УНИВЕРЗИТЕТ У БЕОГРАДУ

ФАКУЛТЕТ ОРГАНИЗАЦИОНИХ НАУКА

Тијана З. Чомић

Унапређење званичне статистике применом *Big Data* концепта

докторска дисертација

Београд, 2019. године

UNIVERSITY OF BELGRADE

FACULTY OF ORGANIZATIONAL SCIENCES

Tijana Z. Čomić

Improvement of official statistics by applying the concept of Big Data

Doctoral Dissertation

Belgrade, 2019

Ментор:

др Зоран Радојичић,

ред. проф. Факултета организационих наука, Универзитет у Београду

Чланови комисије:

др Александар Ђоковић,

доц. Факултета организационих наука, Универзитет у Београду

др Свјетлана Јанковић Шоја,

доц. Пољопривредног факултета, Универзитет у Београду

Датум одбране: _____

Унапређење званичне статистике применом *Big Data* концепта

Сажетак:

Убрзани развој информационо-комуникационих технологија, експанзија Интернета и друштвених мрежа довео је до информационе експлозије – појаве велике количине података који су на располагању практично свима, што указује на неопходност увођења иновација у процесе производње званичних статистичких података. Већина ових података су распрострањени по глобалној мрежи без реда и структуре и пред истраживачима је изазов да их прикупе и обраде на ваљан начин. Централни проблем који се разматра у дисертацији представља унапређење система званичне статистике применом Big Data концепта, кроз примену нових метода и техника истраживања, чиме се олакшава процес анализе великих количина података и побољшавају укупне перформансе система званичне статистике. Кључни научни доприноси дисертације огледају се у унапређењу методолошког поступка процеса статистичког истраживања и формалном опису модела и метода који омогућавају њихову примену. Предложени Big Data модел је флексибилан, проширив, пружа добре перформансе, омогућава интеграцију различитих извора података који укључују и нестандардне формате података. За потребе тестирања примењивости модела на просторима Републике Србије, спроведено је инстраживање које је потврдило полазне претпоставке, чиме су створили услови за развој и имплементацију модела.

Кључне речи: *Званична статистика, Big Data, статистичко истраживање, модел, подаци, информације, индикатори, одрживи развој*

Научна област: Техничке науке.

Ужа научна област: Примењена рачунарска статистика.

УДК број:

Improvement of official statistics by applying the concept of *Big Data*

Abstract:

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. The central problem that is being addressed in the ph.d thesis is exploring possibilities of improving the system of official statistics using Big Data, through the application of new methods and research techniques, which facilitates the process of analyzing large amounts of data and improves the overall performance of the official statistics system. The key scientific contributions of the dissertation are reflected in the improvement of the methodological procedure of the statistical research process and the formal description of the models and methods that enable their application. The proposed Big Data model is flexible, expandable, provides good performance, allows the integration of various data source that include non-standard data formats as well. For the purposes of testing the applicability of the model on the territory of the Republic of Serbia, a research was carried out that confirmed the initial assumptions, which created the conditions for the development and implementation of the model.

***Key words**: Official Statistics, Big Data, Statitstical Research, Model, Data, Information, Indicators, Sustainable development*

Scientific field: Technical science.

Scientific subfield: Applied computational statistics.

UDK number:

Садржај

1 Увод

1.1 Предмет истраживања

Савремене технологије, пре свега интернет, мобилна телефонија и друштвени медији, дале су велики подстицај развоју друштва, креирајући потребе које до скора нису постојале или су биле на знатно нижем нивоу. Техничко-технолошки развој, нарочито у области Информационо – комуникационих технологија (ИКТ) убрзао је процес глобализације и глобалног комуницирања (Podoski & Žilionis, 2008). У претходном периоду развој цивилизације, економије и друштва уопште, био је значајно спорији а информације нису биле доступне свима, поготово не у исто време. Приступ информацијама доводио је до сегментације корисника, доводећи у бољи положај све они који су им имали приступ. Међутим, данас је приступ информацијама омогућен готово свима. Довољно је имати приступ интернету, знање како да се информације пронађу а остало је питање технике и брзине доласка до информација. Развој информатичког друштва иде управо у смеру што бржег добијања информација, уз истовремено смањивање трошкова да се до њих дође.

Од мноштва података који су доступни корисницима, већина је дата у неструктурираној форми (текст, слика, мултимедијлни запис и сл.). Да би се овако структурирани подаци користили морају се претходно обрадити. Обрадом података се бави статистика на традиционални начин, држећи се важеће научне методологије, у складу са којом се прикупљени подаци изражавају на различитим мерним скалама. При томе, највећи број метода статистичке обраде и анализе се обавља над нумеричким обележјима (мереним на ординалним или интервалним скалама), који су високо структурирани. На истим научним поставкама базира се и методолошки приступ на којима се заснива званична статистика (Yan, et al, 2012).

Званична статистика јесте неопходан елемент у информационом систему демократског друштва који снабдева владу, економију и јавност подацима о економској, демографској и социјалној ситуацији и стању животне средине. У том циљу званичне статистичке агенције обезбеђују и, на непристрасној основи, чине

доступном званичну статистику која испуњава захтев практичне корисности, уважавајући право грађана на јавну информацију (UNSTATS, 2015).

Развој званичне статистике се може поделити у три раздобља:

- **Прво**, које се базира на традиционалним методама прикупљања података користећи узорак и теорију великих бројева, методе оцењивања и сл.;
- **Друго**, у коме долази до коришћења административних извора података и обједињавања података којима располажу различити произвођачи података у једну или више база коришћењем јединственог кључа (матичног броја лица, ПИБ-а фирме и слично);
- **Треће**, које је тек у зачетку, а иницира коришћење података са интернета или података који су продукт модерних технологија, уређаја, друштвених медија и сл. (Letouzé & Jütting, 2015).

Убрзани развој ИКТ довео је до информационе експлозије – појаве велике количине података који су на располагању практично свима. Већина ових података су неструктурирани - распрострањени су по глобалној мрежи без реда и структуре и пред истраживачима је изазов да их прикупе и обраде на ваљан начин. Интересовање истраживача за нове изворе података, базиране првенствено на *Big Data* технологијама, почело је паралелно са развојем интернет технологија, мобилне телефоније и друштвених медија. Термин *Big Data* се први пут појављује 1997. у раду научника из *NASA*, где они описују проблеме са којима се суочавају приликом визуализације података које поседују, с обзиром на то да су те базе података прилично велике, што ствара проблем са меморијом рачунара па чак и на удаљеним, спољним дисковима (Cox & Ellsworth, 1997). Те податке, које не могу да сместе на меморију која им је на располагању назвали су *Big Data*. Од тада се у круговима који се баве питањем база података које су толико велике и комплексне да их је немогуће обрадити традиционалним статистичким софтверима усталио термин „Big Data " (Ularu, 2012).

У међувремену, корисници званичних статистичких података, истраживачи, креатори политика и остали, све више користе могућности за употребу мноштва података који не долазе из извора званичне статистике и претварају их у њима

корисне информације (Hammer, 2017). Осим доступности, ове изворе података карактерише и брзина дисеминације. Без обзира на јасно одсуство методолошких поставки, које могу довести (и доводе) у питање квалитет добијених података, корисници наводе да тако добијени подаци могу бити од користи за брзо идентификовање проблема, потреба, пружање услуга, али и за предвиђање и спрачавање криза, а ради добробити становништва (Riley & Smith, 2013).

Из наведеног јасно следи неопходност пуне имплементације *Big Data* у систем званичне статистике, чиме би се обезбедила примена методолошког оквира за употребу нових извора података. Да бисмо дошли у фазу имплементације *Big Data* у систем званичне статистике, не само као новог извора података, неопходно је обезбедити да ти подаци испуњавају следеће критеријуме (ESS, 2015):

- Релевантност,
- Непристрасност,
- Доступност,
- Независност,
- Транспарентност,
- Поверљивост и
- Упоредивост.

У раду се разматра потенцијал примене *Big Data* концепта у домену званичних статистичких истраживања, која карактерише врло чврст законодавни и методолошки оквир. Велика пажња посвећује се тзв. метаподацима, односно „подацима о подацима", који служе бољем разумевању податка, начину на који је он добијен, прецизности мерења, репрезентативности и сл (Eurostat, 2015).

1.2 Циљ истраживања

Основни циљ ове дисертације је да пружи допринос развоју методолошког поступка за примену *Big Data* концепта у систему званичне статистике. Кроз анализу случајева из праксе, тестираних у системима званичне статистике широм света, указано је на неопходност што хитнијег увођења *Big Data* концепта у програме статистичких истраживања. Увођење овог концепта омогућава остваривање његових основних предности пре свега кроз повећање брзине

извештавања, дефинисање нових индикатора, израду флеш оцена као и смањење оптерећења статистичких јединица (испитаника).

Дефинисани *Big Data* модел процеса статистичког истраживања (*Big Data* модел) се базира на традиционалном GSBPM (*Generic Statistical Business Process Model*) моделу (традиционални модел), али је унапређен и прилагођен *Big Data* концепатима.

Након тестирања предложеног *Big Data* концепта у систему званичне статистике анализирани су најважнији предуслови, ограничења и предности коришћења *Big Data* технологија у области примењених статистичких истраживања.

Добијени резултати, приказани у раду, представљају основ за редефинисање методологије статистичког истраживања, уопште, не само у домену званичне статистике.

1.3 Полазне хипотезе рада

Главна хипотеза, која је тестирана у истраживању, гласи:

> Имплементацијом концепта *Big Data,* унапређује се процес статистичких истраживања, отварају се могућности примене нових метода и техника истраживања, олакшава се процес анализе великих количина података и побољшавају укупне перформансе система званичне статистике.

На основу дефинисаног предмета истраживања могу се издвојити посебне хипотезе:

> Х0.1. *Big Data* концепт се може применити за израду индикатора званичне статистике.
>
> Х0.2. Имплементацијом *Big Data* концепта омогућава се развој корпуса нових метода и техника у области званичне статистике.

Даљим прецизирањем наведених посебних хипотеза, формулане су појединачне, које се односе на елементарне чиниоце предмета истраживања:

> Х0.1.1. *Big Data* концепт се може користити у домену званичне статистике за добијање брзих, прелиминарних (*flash*) оцена појединих индикатора у кратком периоду.
>
> Х0.1.2. *Big Data* концепт се може користити за оцењивање параметара на микро нивоу.
>
> Х0.2.1. *Big Data* ресурси се могу користити као алтернативни извори података за званичну статистику.
>
> Х0.2.2. Имплементацијом предложене методологије истраживања базираног на *Big Data* концептима могуће је извршити мерење појединих индикатора одрживог развоја у Републици Србији.

1.4 Научне методе истраживања

Имајући у виду да је концепт коришћења *Big Data* нов и још увек на нивоу истраживања могућности, као и чињеницу да већина званичних статистика и даље користи традиционалне концепте, основне методе истраживања које су коришћене током израде дисертације подразумевају прикупљање, анализу и систематизацију постојеће литературе и метода у области коришћења *Big Data* у сврхе званичне статистике, као и резултата практичне примене постојећих метода статистичке анализе.

За проблем коришћења *Big Data* за мерење индикатора одрживог развоја коришћена је анализа случаја, односно бенчмаркинг (*benchmarking*) анализа већ развијених модела. Такође, ове методе су коришћена и приликом дефинисања неопходне инфраструктуре за примену *Big Data* у званичниој статистици, укључујући и неопходну законску регулативу.

2 Технике и методе прикупљања и обраде података у званичној статистици

2.1 Развој званичне статистике

Статистика је термин која се двојако користи. Прво значење овог термина односи се на бројчане податке који представљају агрегате посматраних обележја појединаца. Одатле проистиче и друга примена овог термина, за науку која се бави овим подацима, методима њиховог прикупљања, анализом и интерпретацијом (Vuković & Vukmirović, 2004). Статистика као наука произашла је из математике као шире области а статус науке стекла је захваљујући развоју специфичних теорија и методологија. Основа статистике као науке лежи у закону великих бројева и теорији вероватноће. Управо Закон великих бројева и теорија вероватноће омогућавају да се на основу релативно малих (или довољно великих) узорака дође до закључака за читаву популацију.

Потреба да се различите „појаве" преброје и квантификује датира још од давних времена, а ове потребе су по правилу економске природе: број пореских обвезник, вредност имовине, број расположивих војника и слично, све су то подаци који су од давнина биле основ „доносиоцима одлука". Дакле, потреба за квантификовањем појава довела је до развоја статистике као науке која је требало да омогући да се подаци прикупе за мање времена и уз ниже трошкове. Одатле и први забележени примери пребројавања чине зачетке званичне статистике.

Постоје докази документовани рукописима на папирусу да је први попис становништва спроведен у Египту још у време фараона и то у периоду од 3340 до 3056 године пре Христа у циљу наплаћивања пореза и одређивања способности становништва за обављање војне службе (Capponi, 2011).

Утемељење за развој статистике, која почива на теорији вероватноће, дали су Паскал и Фермат кроз коресподенцију из 1654. (TSM, 2018). Раном развоју друштвених статистика, пре свега демографије, допринели су *William Petty* и *John Graunt* израдом таблица живота (*life tables*) 1662. којима је се рачуна вероватноћа преживљавања за сваку старосну кохорту (Bacaer, 2011). Пиониром и зачетником

статистике сматра се Томас Бајес који је у есеју из 1763. први покушао да искористи теорију вероватноће као инструмент индуктивног закључивања, односно да се од појединачног дође до општег тј. од узорка до популације (Fisher, 2006).

Закон о званичној статистици

Из дефиниције статистике као науке проистекле су и дефиниције званичне статистике. Према Закону о званичној статистици Републике Србије „Званична статистика обезбеђује, на непристрасним основама, бројчане и репрезентативне податке и информације о масовним економским, демографским и друштвеним појавама и о појавама из области радне и животне средине, и то за све кориснике: привредне субјекте и њихова удружења, државне органе, органе аутономних покрајина и органе јединица локалне самоуправе, културне, образовне и научне институције, као и за најширу јавност“ (RZS, 2017).

На нивоу Европе, основно тело задужено за статистику је Евростат који функционише у оквиру Европске комисије. Европски статистички систем (*European Statistical System - ESS*) подразумева мрежу институција, у коју су поред Евростата укључене и друге, пре свега национални статистички институти (*NSIs*) али и друге институције - произвођаче званичне статистике земаља чланица. Поред тога у ову мрежу укључене су и ЕЕА и ЕФТА земље кроз партнерство са *ECC*, који свој рад координира и са земљама кандидатима али и са другим међународним институцијама као што су *ОЕЦД, УН, ММФ* и Светска банка. Основна улога *ECC* је да координира пружање упоредивих статистика на нивоу Европе, што значи да је рад ове мреже фокусиран на европске политике, кроз хармонизацију методологија те се са ширењем и развојем политика Европске уније и овај фокус проширује (Eurostat, 2017).

Организација уједињених нација (ОУН) од 1946. године развија међународни статистички систем и преко великог броја својих статистичких органа прати токове у светској привреди (UNSTATS, 2017; UNECE, 2018; ILOSTAT, 2018). Поред ОУН постоје и друге организације које статистички прате светску економију: Организација за економску кооперацију и развој (OECD), Европска унија (EU),

Међународни статистички институт (*ISI*), Интер-амерички статистички институт (*IASI*) итд.

Статистичким системом ОУН управљају Економски и друштвени савет УН преко **Статистичке комисије.** Задаци Статистичке комисије су следећи (UNSTATS, 2017):

- пружање подршке у развоју националних статистика,
- побољшање упоредивости националних статистичких показатеља,
- развијање статистичке канцеларије при Секретаријату УН,
- пружање помоћи органима УН у вези сакупљања, интерпретације и ширења статистичких информација,
- унапређивање статистике и статистичких метода уопште.

У оквиру Економског и друштвеног савета УН постоји и пет регионалних комисија. Свака регионална комисија има своје статистичко одељење. Преко конференција које организују регионалне комисије врши се усклађивање статистичких система унутар региона и успоставља се веза са статистичким системима других међународних организација (Eurostat, OECD итд.).

Статистичка служба OECD-a има задатак да обезбеди функционисање јединственог статистичког система задовољавајући потребе за статистичким показатељима земаља – чланица. Статистичка служба функционише при економском и статистичком одељењу. Статистички систем OECD-a је повезан и усклађен са статистичким системом УН-а (OECD, 2018).

Званична статистика у Републици Србији базира се на Закону о званичној статистици, стратегији и годишњим плановима који представљају обавезујући правни оквир за све институције које су у оквиру система званичне статистике. Према овом Закону, Републички завод за статистику је „главни произвођач и дисеминатор званичних статистичких података, као и одговорни стручни носилац, организатор и координатор система званичне статистике у Републици Србији и представља званичну статистику Републике Србије у међународном статистичком систему" (RZS, 2017).

2.2 Извори података

Сваком истраживању, па и статистичком, претходи дефинисање предмета и циљева истраживања као и индикатора и информација које је истраживањем потребно добити. Када су циљеви и излазни индикатори јасно дефинисани, потребно је одредити из којих извора и на који начин ће се потребни подаци прикупити. Извори података могу бити (Kumar, et al., 2013):

- примарни и
- секундарни.

Подаци који се добијају из секундарних извора називају се секундарни подаци и то су већ постојећи подаци, односно подаци који су прикупљени у неке друге сврхе, али који се могу користити и за сврхе конкретног истраживања. За разлику од секундарних података примарни подаци су подаци који се по први пут сакупљају и то за потребе конкретног истраживања (Ханић, 2003). Једна од основних разлика између примарних и секундарних података је у трошковима неопходним за њихово прикупљање. Коришћење секундарних података је неупоредиво економичније него прикупљање примарних те је пре спровођења истраживања неопходно истражити релевантне секундарне податке и оценити да ли се могу користити у сврху конкретног истраживања или је ипак неопходно спровести ново прикупљање података које изискује значајне трошкове.

Званична статистика се, по правилу бави прикупљањем примарних података. Подаци које прикупља званична статистика су од великог значаја на нивоу једне државе и у највећој мери доносиоци одлика користе управо податке званичне статистике. Широк спектар података које прикупља званична статистика подразумева и комплексна истраживања која су од користи најширем кругу корисника. Поред прикупљања примарних података националне статистике имају и овлашћења за обједињавање података прикупљених од стране других институција унутар статистичког система, који су у овом случају, из угла националне статистике, секундарни извори података. Овде ћемо навести као пример Народну банку, пореске институције, органе Министарства унутрашњих послова, различите регистре и сл. Према изворима података које производи званична статистика, развој статистике може се поделити у неколико раздобља:

- прво, које се базира на традиционалним методама прикупљања података користећи узорак и теорију великих бројева, методе оцењивања и сл.;
- друго, у коме долази до коришћења административних извора података и обједињавања података којима располажу различити произвођачи података у једну или више база коришћењем јединственог кључа (матичног броја лица, ПИБ-а фирме и слично).
- као резултат интернет револуције, у последњих неколико година у повоју је и ново, треће раздобље, које је тек у зачетку, а иницира коришћење података са интернета или података који су продукт модерних технологија, уређаја, друштвених медија и сл.

Имајући у виду разлику у **цени** прикупљања примарних у поређењу са ценом коришћења секундарних података јасно је да је тенденција све сврсисходније коришћење секундарних извора података чак и у званичној статистици.

С друге стране, сама природа прикупљања података базираних на репрезентативном узорку доводи у питање **квалитет** прикупљених података услед растућег неодговора изабраних јединица, односно њиховог одбијања да учествују у истраживању. Уколико неодговор доводи до грешке неодговора (*nonresponse error*), онда ће и оцене индикатора бити пристрасне. Пристрасност која настаје услед неодговора немогуће је оценити без додатног истраживања које би имало за циљ да утврди да ли постоје разлике у карактеристикама јединица коју су одбиле да учествују у истраживању и оних које су пристале. У случају да не постоји разлика и да неодговор није систематски, не би дошло до пристрасности. Међутим, уколико разлика постоји, оцене параметара су пристрасне (Holbrook et al., 2008).

Још један од изазова приликом прикупљања примарних података је и питање **правовремености**. У претходном периоду, када је развој цивилизације, економије и друштва уопште био значајно спорији, информације су биле ограничен ресурс. Они који су имали приступ информацијама били су у бољем положају од оних који нису. Међутим, данас је приступ информацијама омогућен готово свима. Довољно је имати приступ интернету, знање како да се информације пронађу а остало је питање технике и брзине доласка до информација. Развој информатичког друштва

иде управо у смеру што бржег добијања информација, уз истовремено смањивање трошкова да се до њих дође.

2.3 Методи прикупљања података

Под прикупљањем података подразумева се обједињавања и мерења информација о одређеној појави применом различитих инструмената. Традиционалне методе прикупљања података подразумевају један од следећа три метода (Aker, et al. 2012):

- посматрање
- испитивање и
- експериментисање

Преглед метода за прикупљање података коришћених у Републичком заводу за статистику приликом анкетних истраживања у домаћинствима

Националне статистике широм света као основну функцију имају прикупљање примарних података. Примарни подаци су оригинални подаци, тј. подаци који се прикупљају по први пут, за потребе конкретног истраживања. Имајући у виду да су примарни подаци најскупљи, с обзиром да захтевају вишефазни процес, почевши од креирања упитника и методологије па до самог процеса прикупљања података и креирања база података, битно је да у позадини стоји јак интерес за њиховим прикупљањем, било материјални или не.

Подаци које националне статистике прикупљају по правилу су подаци који немају директну комерцијалну вредност односно подаци које се, по правилу не користе директно за доношење пословних одлука, као што је на пример случај са маркетинг истраживањима. Њихова карактеристика је да се углавном користе за креирање одређених политика на нивоу државе, иако њихова коректна интерпретација може у великој мери да има утицаја и на реални сектор у смислу да може да да слику економије у земљи, одређеном пословном сектору, демографским карактеристикама одређеног подручја и слично. Имајући у виду ове карактеристике података које прикупљају националне статистике, јасно је зашто је ова функција управо њима поверена у свим државама. Дакле, националне статистике баве се прикупљањем примарних података који су од општег интереса у једној земљи.

У зависности од потреба одређених истраживања, у националним статистикама користе се различити методи прикупљања података. Као најопсежнији метод прикупљања примарних података, пописи су, по правилу, поверени националним статистикама, имајући у виду опсежност овог метода као и високу цену коштања читавог процеса. Попис подразумева, као што и сама реч каже, процес систематског прикупљања података о **свим** члановима посматране популације. Када је реч о, на пример, становништву једне земље јасно је да ова акција мора да буде координисана са самог државног врха, односно институције која је одговорна за статистику државе. Поред пописа, националне статистике користе и друге методе за прикупљање података, као што су административни извори, регистри и слично (Carlson, et al. 2010).

Међутим, имајући у виду цену и комплексност спровођења пописа, најчешћи метод прикупљања податка у националним статистикама су анкете, било да се ради о домаћинствима, појединцима или пословним субјектима. Анкета је метод прикупљања података који се спроводи на узорку, који се бира на начин да у најбољој мери репрезентује карактеристике посматране популације (Михаиловић, 2004). У овом делу рада детаљније ћемо се посветити различитим методама прикупљања података приликом анкетних истраживања домаћинстава.

Основна карактеристика анкетних истраживања је да се она спроводе на узорку који у најбољој мери репрезентује популацију. Величина узорка зависи пре свега од циља истраживања, али и од нивоа до ког желимо да приказујемо податке а да репрезентативност буде испуњена. При томе, као главни ограничавајући фактор у избору величине узорка је пре свега буџет којим се за то истраживање располаже, те се на основу њега одређује и ниво на коме желимо репрезентативност. Већи узорци дају поузданије резултате од малих узорака. Међутим, са повећањем узорка не повећава се пропорцијално поузданост резултата али повећава се цена. С тога, приликом планирања узорка увек треба балансирати између цене и прецизности коју истраживач жели (Chaudhuri & Stenger, 2005).

Методи прикупљања података

Основни метод за прикупљање података код анкетних истраживања је интервју лицем у лице који подразумева учешће обучених анкетара, који су упознати са методологијом истраживања и који има задатак да забележи одговоре испитаника на питања дефинисана у упитнику. При томе, најчешће анкетар са собом има одштампани упитник са питањима и упутствима које следи приликом постављања питања. Након процеса анкетирања у овом случају неопходно је податке прикупљене на папиру пребацити у електронски базу података како би се радила даља анализа и објединили сви резултати. Овај метод који подразумева коришћење папирних упитника приликом спровођење интервјуа (анкетирања) познат је као *PAPI* метод (*Paper and Pen Interviewing*). Са технолошким развојем и увођењем рачунара у готово све пословне процесе, дошло је и до напредовања у области прикупљања података. У новије време предности развоја технологије уводе се у све сфере статистичких истраживања па тако и у сам процес прикупљања података (Вукмировић, 1994). Тако традиционални метод прикупљања података у процесу интервјуисања лицем у лице све више замењује *CAPI* метод (*Computer Assisted Personal Interviewing).* Овај метод подразумева да анкетар са собом има рачунар са већ инсталираним софтвером који је приказује питања на екрану рачунара, и одговоре уноси директно у апликацију чиме се превазилази фаза уноса података.

У процесу анкетирања могуће је користити и друге методе, које нису лицем у лице. *CATI* метод (*Computer Assisted Telephone Interviewing*) подразумева да анкетар позива број телефона испитаника који му је додељен и путем телефона врши анкетирање (McCarthy, 1989; Вукмировић, и остали, 1995). У новије време све чешће је у примени и тзв. *CAWI* метод (*Computer Assisted WEB Interviewing*) који се користи за он-лине истраживања, при чему се испитанику најчешће доставља линк ка истраживању и испитаник самостално приступа процесу попуњавања упитника (Chizawsky, et al., 2011). Овај метод за разлику од претходно поменутих не захтева присуство анкетара.

Основна предност ових метода, које подразумевају коришћење рачунара, у односу на *PAPI* метод је у томе што приликом коришћења рачунара, постоји могућност уграђивања различитих контрола у саму апликацију, тако да анкетар не мора да води

рачуна о тзв. „скоковима" са питања на питање и грешке самих анкетара се своде на минимум.

У Републичком заводу за статистику у Србији (РЗС) још увек најраспрострањенији метод у анкетним истраживањима је *PAPI* метод. Као што је поменуто, *PAPI* метод подразумева учешће анкетара у процесу прикупљања података. РЗС располаже широком мрежом искусних анкетара на целој територији Србије који се по потреби ангажују за различита истраживања. Међутим, *PAPI* метод је веома подложан грешкама анкетара, тако да је у овом случају потребна вишеструка контрола рада самих анкетара како би се грешке и пропусти минимизирали (Čomić, 2016). Такође, поједина истраживања захтевају да у анкетирању учествују анкетари који нису раније имали искуства са сличним истраживањима, чиме се мрежа анкетара шири.

Поред *PAPI* метода РЗС примењује и *CAPI* и *CATI* метод код појединих истраживања. РЗС у оквиру својих просторија располаже са два *call-centra* са укупно 25 телефонских линија које анкетари користе приликом анкетирања. Контрола рада анкетара је код *CATI* методе знатно ефикаснија, с обзиром да контролор може, у сваком тренутку да се укључи у телефонски разговор анкетара и испитаника и провери да ли анкетар на исправан начин поставља питања испитанику, да ли се придржава методологија као и какав приступ има ка испитаницима. *CAWI* метод још увек није у широј примени у РЗС-у, нарочито када је реч о истраживањима код домаћинстава, с обзиром да у Србији, према подацима Истраживања о употреби информационо комуникационих технологија из 2018, тек 72,9% домаћинстава поседује интернет прикључак, док скоро четвртина популације (24,2% лица) никада није користило интернет (РЗС, 2018а).

Предност коришћења *CAPI, CATI* или *CAWI* методе у односу на *PAPI* лежи и у томе да се период обраде података скраћује јер се из процеса искључује фаза ручног уноса података, а и само чишћење базе је знатно краће с обзиром да је велики број контрола већ уграђен у апликацију за анкетирање (Evans & Mathur, 2005).

3 *Big Data* концепт

3.1 *Big Data* и информатичка револуција

Информатичка револуција доводи и до промена парадигми у науци. Захваљујући новим технологијама може доћи до замена улога у којима свет спознајемо на бази података а не теорије, индукцијом а не дедукцијом. Иако многи научници сматрају да наука о подацима (*data science*) нема упориште у филозофији, она изражава концептуални оквир о томе како се свет може сагледати, из друге перспективе. Филозофија је овде битна јер пружа интелектуални оквир о томе шта је то што се жели добити из података, односно која питања треба поставити, како их поставити, како смислено тумачити резултате који се добијају и како се на основу тога надограђује постојеће знање о појави (Kitchin, 2014).

Имајући у виду да је концепт *Big Data* релативно скоријег датума, тешко је причати о „историјату" развоја *Big Data*. Концепт *Big Data* је тек у повоју и у већини случајева још увек је на концептуалном нивоу, док се права имплементација *Big Data* тек очекује када се на свим нивоима испуне претпоставке за њихово коришћење (Bok et al., 2018).

Иако подаци као појам нису ништа ново, као ни чињеница да их је временом све више и више, када је термин „*Big Data* " први пут употребљен и од стране кога, не може се са сигурношћу рећи. Ипак, неки извори наводе да је врло могуће да је то био Џон Маши (*John Mashey*), средином 1990-их година прошлог века, који је у том периоду био водећи стручњак у компанији *Silicon Graphics, Inc*. из САД (Diebold, 2018)). Скоро деценију и по касније, од 2010-2012. године, *Big Data*, као појам и уопште тема, постаје једна од најактуелнијих у ИТ свету, а интересовање стручњака (па и оних који то нису), као и бављење истом, у константном је порасту.

Користећи један од најзначајних извора *Big Data* – интернет и његов најзначајнији бренд (*Google*) можемо закључити да се термин „*Big Data*" уочљиво намеће 2011 године, да би само за годину дана надмашио „*data mining*" као до тада водећи научни поступак за напредну анализу података и наставио са експоненцијалним растом (остављајући „*data mining*" у благом линеарном паду) (Слика 3.1).

Слика 3.1: Претраживање термина на *Google-y* (Google, Новембар 2018)

Напомена: Бројеви представљају интересовање за претраживање у односу на највишу тачку на графикону за дати регион и време. Вредност од 100 је највећа популарност термина. Вредност од 50 значи да је термин пола толико популаран. Резултат 0 значи да није било довољно података за овај термин.

Тобиас Преис (*Tobias Preis*), професор из *Warwick Business School* и његови сарадници користили су *Google Trends* за тестирање теорије да корисници интернета из земаља које имају већи БДП (бруто домаћи производ) по глави становништва имају веће склоности ка тражењу информација везаних за будућност у односу на оне везане за прошлост. Налази сугеришу да постоји веза између онлајн понашања и економских индикатора (Preis, et al., 2012). Аутори ове студије су 2010. године анализирали захтеве за *Google* претраживањима који су се односили на будућу годину 2011. и на оне који су се односили на претходну 2009. годину како би добили "оријентациони индекс за будућност". Затим су овај индекс упоредили са БДП по глави становништва сваке земље и пронашли да постоји велика тенденција да *Google* корисници са већим бруто друштвеним производом траже више информација везаних за будућност. Резултати су већ тада наговештавали да постоји потенцијална повезаност између економског успеха неке земље и

претраживања вршених од стране њеног становништва, што је обухваћено помоћу *Big Data*.

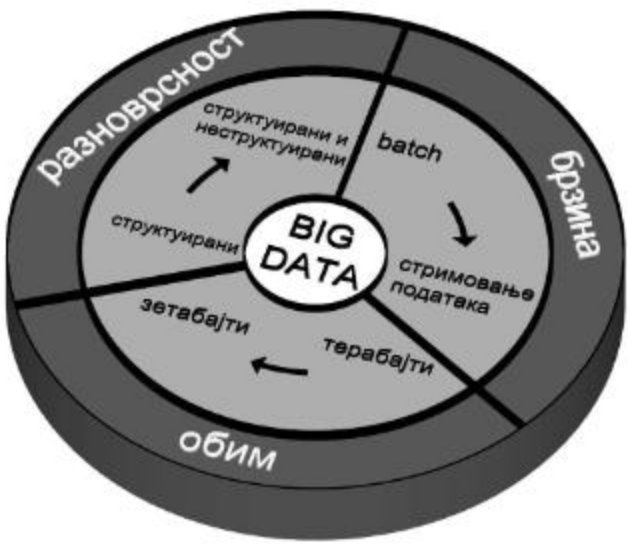
3.2 *Big Data*, основне дефиниције

С обзиром да не постоји једна, опште прихваћена, дефиниција појма *Big Data*, у овом раду наводимо неколико одабраних, ради стварања опште и што комплетније слике о томе шта *Big Data* заправо представља. Већина ових дефиниција су у основи сличне:

- *Big Data* је термин који се користи када се говори о компилацији података и информацијама, које су толико велике и комплексне да је тешко вршити њихову обраду помоћу стандардних, тренутно доступних метода и алата за уређивање података. Под тешкоће спадају прикупљање, ажурирање, складиштење, претраживање, графички приказ, утврђивање расподеле, и нарочито анализа података. Тренд коришћења великих компилација података и међусобно повезаних информација се наставља због појаве нових, додатних информација које се добијају њиховом обрадом, у поређењу са обрадом мањих, одвојених компилација које укупно садрже исту количину података. Кроз обраду већих скупова могуће је уочити пословне трендове, установити квалитет тржишних или других истраживања, спречити болести, борити се против криминала, установити саобраћајне услове у реалном времену, итд. (Cavanillas, et al., 2013).
- Термин *Big Data* се, у најкраћем, односи на информације које се не могу обрадити и анализирати на традиционална начин, коришћењем конвенционалних процеса и алата (Dumbill, 2013).
- *Big Data* се дефинише као скуп података који превазилазе могућности постојећих софтвера за управљањем базама података у смислу прикупљања, смештаја, управљања и анализе (Brown, et al., 2013).
- *Big Data* је термин који се користи да опише експоненцијални раст и доступност података, како структурираних, тако и неструктурираних. *Big Data* је битан у пословању, као и у друштву, у истој мери као интернет, јер више података омогућава прецизније спровођење анализа података у разним областима (SAS, 2018).

- *Big Data* представља велику количину података који се производе из многих извора, алармантном брзином, обимом и разноврсношћу а као резултат различитих дигиталних процеса и друштвених медија. Како би се значајне и вредне информације извукле из *Big Data*, неопходно је имати оптималну моћ обраде, аналитичке могућности и вештине (IBM, 2018).

Оно што је заједничко за већину дефиниција *Big Data* је употреба у неком облику три "V", која представљају почетна слова од енглеских речи: *Volume* (обим), *Variety* (разноврсност) и *Velocity* (брзина). Ову дефиницију увео је још 2001, аналитичар из ИТ индустрије Даг Лани (*Doug Laney*), аналитичар запослен у *Gartner-y*, да би данас била опште прихваћена у ИТ индустрији (Слика 3.2) (Laney, 2001).



Слика 3.2: *Big Data* (Zikopoulos, et al., 2012)

Обим (*Volume*) података. До повећања обима података углавном долази услед убрзаног развоја ИКТ. Евидентан је раст количина података базираних на трансакцијама: у телекомуникацијама (мобилна телефонија, пре свега), банкарском пословању, осигурању, медицинским услугама, и сл. Расте и број неструктурираних података који долазе од стране друштвених медија. Свакодневно

се повећава количина података који се читају са сензора и сличних уређаја: климатски сензори, бројачи саобраћаја, GPS уређаји, скенери на касама у малопродаји и сл. Такође, увођење *Smart city* технологија које се базирају на увођењу интелигентних уређаја у домаћинствима и технологија за мерење (потрошње струје, гаса и сл. – *smart meters*) (Radenković, 2017).

У не тако давној прошлости, велике количине података су представљале проблем како код прикупљања, тако и њиховог складиштења. Развојем нових технолошких решења (пре свега *Cloud* технологија) трошкови складиштења података су значајно смањени (Minelli & Chambers, 2013).

Међутим, појављују се други проблеми који се односе на употребу великих количина података из различитих извора: избор значајних података унутар великих скупова података, утврђивање њиховог квалитета и стварања вредности из њих (пословна аналитика). Доступност и право коришћења велике количине података које су у власништву приватних компанија (нпр. мобилни оператори) су посебно значајна додатна ограничења за ширу употребу *Big Data*.

Генерално, извори *Big Data* се могу класификовати на различите начине. Једна класификација, која се користи у системима званичне статистике приказана је у табели 3.1.

Табела 3.1: *Big Data* извори у званичној статистици (Vale, 2013; Hackl, 2016, Radenković, et al., 2017)

Назив	Кратак опис	Пример
Административни извори	Подаци у електронском облику који се налазе у систему али су неструктурирани или толико велики да се не могу на конвенционални начин обрадити	Скенирани документи
Подаци са скенера	Подаци који се добијају употребом оптичких читача	Подаци са каса у малопродаји добијени преко бар кода

Назив	Кратак опис	Пример
Трансакциони подаци	Подаци који су резултат трансакција између два ентитета, било да је реч о лицима или пословним субјектима	Трансакције са кредитним картицама, он-лине трансакције (укључујући и оне које се обављају путем мобилних телефона)
Подаци са спољних сензора	Подаци који су резултат употребе сензора на јавним местима	Сензори на путевима, сензори за климатске услове
Сателитски снимци	Подаци добијени путем сателита, укључујући хидрометеоролошке податке	Сателитски снимци пољопривредних површина
Подаци са уређаја за праћење	Подаци настали као последица регистровања путање, појединца или објекта,	Мобилни телефони – праћење локације, *GPS* уређаји – праћење моторних возила
Бихевиорални подаци	Подаци настали као последица праћења понашања онлине корисника, дисагрегирани по социо-демографско-економским варијаблама	Он-лине претраге (о производима, услугама или неким другим подацима), прегледи он-лине страница
Ставови и мишљења	Подаци који су резултат онлине праћења и истраживања јавног мњења	Коментари на социјалним медијима (facebook, tweeter, youtube)
Интернет интелигентних уређаја (IoT)	Подаци настали као резултат повезивања већег броја корисника, уређаја, сервиса и апликација на интернет.	Подаци о утрошку електричне енергије у интелигентним и паметним стамбеним објектима
Остали подаци	Сви остали извори података који нису категоризовани а могу се користити у сврхе званичне статиситке	Фотографије које су направљене коришћењем дронова

На нивоу ЕУ утврђена је листа доступних *Big Data* извора по доменима и статистичким институцијама земаља који су их идентификовали. Ова листа извора објављена је на основу истраживања које је спроведено у оквиру *ESSnet Big Data* пројекта у коме су учествовале све земље чланице (ESSnet, 2016).

Између осталог, циљеви пројекта су били:

- Идентификација *Big Data* извора (укључујући њихову трајност и доступност у различитим земљама),
- Процена могућности коришћења одабраних извора за анализу података у областима становништва, пољопривреде и туризма.

Идентификација резултата/нових производа из пилот студија који могу бити корисни у овим областима.

У табели 3.2 дат је резултат инвентуре потенцијалних *Big Data* извора у Холандији, прирећен у сарадњи са Холандским статистичким институтом (CBS).

Табела 3.2: Потенцијални *Big Data* извори у Холандији (ESSnet, 2016).

Извор података	Могући индикатори
AIS подаци	Кретање бродова
Евиденција позива	Страни туристи
	Дневна популација
	Посетиоци на фестивалима
Google трендови	Статистика здравља
Сензори на путевима	Густина саобраћаја
	Економски индикатори
Подаци са скенера	Информације о ценама на бази података са скенера
Smart city подаци	Кретање људи током фестивала
Друштвени медији	Сентимент анализа на друштвеним медијима
	Употреба друштвених медија
Twitter	Теме на *Twitter*-у
	Друштвена кохезија
	Основне емоције
	Осећај безбедности
	Холандски туристи

	Страни туристи
	Детектовање сезонских промена
Веб странице	Информације о ценама на бази веб страница
	Отворена радна места
	Цене становања
	Метаподаци о компанијама
	Детектовање и процена густине настањености животиња

Осим актуелног стања, истраживање је садржало и планове националних статистичких институција везано за коришћење потенцијалних *Big Data* извора, након 2018 године (Слика 3.3). На основу ових података може се закључити да ће се у будућности ЕСС, када је реч о *Big Data* изворима највише ослањати на интернет (*web*) простор, мобилне сервисе (за праћење локације), административне податке које производе друге институције, интернет претраживања и друштвене мреже.

Слика 3.3: Коришћење потенцијалних *Big Data* извора, пре и после 2018. године унутар ECC (ESSnet, 2016)(Wang, et al., 2018)

Разноврсност (*Variety*)

Подаци се у данашње време јављају у разним облицима. Развој информационих технологија, пре свих довео је до појаве различита врста података који се могу прикупити. Једна од основних подела је на структуриране, полуструктуриране и неструктуриране податке (Goes, 2014).

- **Структурирани подаци** су постојећи подаци, смештени у базама података, по свим правилима складиштења података.
- **Полу-структурираних подаци** се користе за описивање структурираних података који се не уклапа у формалну структуру модела података. Ови подаци не садрже ознаке које раздвајају семантичке елементе, а који поседују способност спровођења хијерархије унутар података.
- **Неструктурирани подаци** су у основи информације које немају унапред дефинисани модел података (мета податке) и/или се не уклапају добро у базу података. Неструктурирани подаци су обично текстуални или мултимедијални подаци, али могу бити и нумерички – као што су датуми, бројеви, и сл.

Конкретно, под неструктурираним подацима се подразумевају следећи формати:

- Текст,
- Аудио
- Видео,
- Слике,
- Геопросторни подаци,
- Интернет подаци.
    - *click streams*
    - *log*-ови

Уочено је да подаци све више постају "неструктурирани", тачније, раст неструктурираних података прати експоненцијални тренд. И количина структурираних података такође расте, али у складу са линеарним трендом (Moreno, 2011).

На нивоу ЕУ, већина података која се користи у истраживањима на бази *Big Data* концепта је дато у неструктурираној форми, што се може закључити на основу анализе примењених *Big Data* извор унутар ЕСС (Табела 3.3). На упитник је одговорило 19 земаља ЕУ.

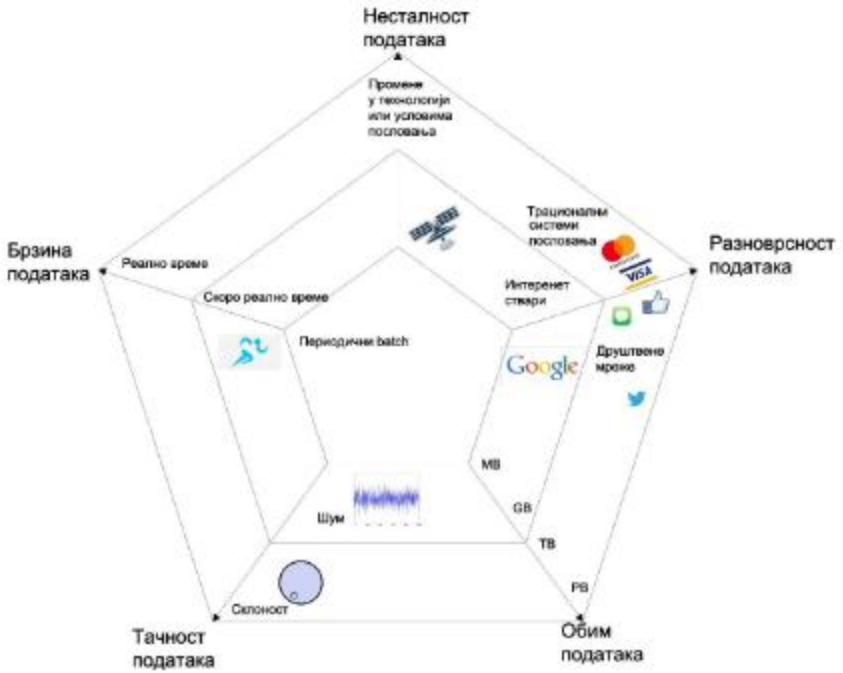
Табела 3.3: *Big Data* извори који се користе у званичним статистичким институцијама земаља ЕУ (ESSnet, 2016)

Извор	**Број институција**
Web странице	13
Праћење мобилног сензора - Локација мобилног телефона	8
Подаци јавних агенција	8
Интернет претраживања	8
Бизнис подаци - Комерцијалне трансакције	8
Подаци са сензора - Фиксирани сензори - Сензори саобраћаја/веб камере	7
Друштвене мреже: *Facebook, Twitter, Tunlr* итд.	6
Бизнис подаци - Кредитне картице	6

Брзина (*Velocity*).

Време које је неопходно да се добије крајњи резултат истраживања у значајној мери дефинише *Big Data* концепт (McAfee & Brynjolfsson, 2012). Под брзином се у *Big Data* концепту претпоставља реално или приближно реално време за достављање (праћење) резултата за разлику од традиционалног истраживања где се на коначан резултат чека данима, неретко месецима, чак и годинама (нпр. званична статистика – БДП обрачун).

У дефиницију *Big Data* све више се додају још 2 „V" која се односе на *Veracity* (тачност) и *Value* (вредност) или *Volatility* (неконзистентност - у смислу подложности променама) (Слика 3.4). У имплементацији *Big Data* концепта у систем званичне статистике посебно је значајна *Veracity* - димензија тачности, (Samuel, et al., 2015).



Слика 3.4: 5V које дефинишу *Big Data* (Hammer, et al., 2017)

У наставку истраживања, под појмом *Big Data* концепта подразумеваћемо скуп технологија и техника које омогућава обухват, обраду и анализу података који су исувише разнолики, пребрзо се мењају и којих има превише да би им „традиционална" ИКТ инфраструктура и методологија омогућила ефикасно коришћење у процесу извештавања, односно доношењу одлука.

3.3 Друштвени медији

Један од најчешће коришћених извора у *Big Data* су друштвени (социјални) медији, нарочито друштвене (социјалне) мреже. У овом тренутку се они највише користе и у пилот пројектима у области званичне статистике који су предмет наредних поглавља.

Друштвени медији (*Social Media*) се често појмовно мешају са друштвеним мрежама (*Social Network*). Међутим, иако су друштвене мреже попут *Facebook-a*, *Twitter-a*, *LinkedIn* и *Instagram-a* најзаслужније за популарност друштвених медија, оне су само део њих. Основна категоризација друштвених медија приказана је у табели 3.4.

Табела 3.4: Основна категоризација друштвених медија (Safko, 2012)

Назив категорије			
Друштвене мреже	Размена фотографија	Аудио записи	Микро-блоговање
Лајвкастинг (*Livecasting*)	Виртуелни светови	Играње	*RSS* и Агрегатори
Претраживање	Мобилна телефонија	Интерперсонални медији	…

4 Примена концепта *Big Data* у званичној статистици

Имајући у виду све добробити информатичке револуције, званична статистика ће, иако по многима ригидан систем, што пре морати да се прилагоди променама и искористи благодети нових технологија како би ишла у корак са потребама доносиоца одлука и како би уз смањење трошкова, повећање квалитета и правовремено могла да одговори потребама доносиоца одлука али и других корисника података. „Ригидност" званичне статистике резултат је строгих методолошких поступака прописаних у форми „препорука" од стране Статистичке комисије УН, односно „европским стандардима" прописаним од Евростата. Као резултат, званична статистика улази у трећу фазу свог развоја у којој ће се поред пописа, истраживања на узорку и коришћења административних извора података користити и подаци који су доступни, као секундарни подаци, на интернету (UNSTATS, 2014).

Међутим, ови нови извори података су, због свог волумена и брзине развоја и даље велика непознаница за традиционалну статистику те се у овој фази мора прво извршити темељна анализа и процена квалитета података доступних захваљујући новим технологијама. Такође, питање које се поставља је како *Big Data* могу допринети прецизнијем и правовременом мерењу економских, друштвених и еколошких појава (Hackl, 2016).

Иако административни извори (који су често главни извор података у званичиној статистици) спадају у „изворе обимних података" и испуњавају (или могу испуњавати) критеријум обима и разноврсности, оно због чега се административни извори не могу сматрати *Big Data* -ом је брзина којом званичне статистике добијају ове податке. Наиме, они се добијају на годишњем, кварталном или месечном нивоу те не испуњавају критеријум брзине који дефинише *Big Data*. Када би се ови подаци добијали на дневном или чак и на недељном нивоу они би се могли сврстати у *Big Data*. Упаривање различитих административних извора по одређеном кључу (матични број, број социјалног осигурања и сл.) даје огромну количину података на нивоу појединаца. Предност овога је у могућности приказивања података на

нижим нивоима (насеље, општина) него што је случај са традиционалним прикупљањем података који даје информације најчешће до нивоа региона.

4.1 Нове технике прикупљања података

Пад стопе одговора приликом анкетних истраживања, било да се ради о истраживањима која се врше на домаћинствима или истраживањима у оквиру пословног сектора, све је израженији. У том смислу коришћење *Big Data* доносиоцима одлука може донети информације у реалном времену нарочито у областима као што су статистика цена, запослености, индустријске производње, демографије и сл. (OECD, 2013). *Big Data* имају потенцијал пружања релевантнијих и благовременије податке него што је то случај са традиционалним методама као што су анкете и административни извори (Bok, et al., 2018). Међутим, већина извора *Big Data* је у власништву приватног сектора те је неопходно законодавство које би омогућило коришћење ових података у званичној статистици. Непостојање законодавства је један од главних разлог зашто *Big Data* подаци нису још увек у широј примени у званичној статистици.

Без обзира на велике могућности *Big Data,* не треба размишљати о овим подацима као о замени за традиционалне методе прикупљања података, већ више као могућег извора који би употпунио статистички систем. И поред покушаја да *Big Data* пруже што исцрпније податке, пре свега кад се ради о подацима који долазе са социјалних медија, они су, по својој природи увек парцијална, са различитим празнинама, пристрасностима и неизвесностима. *Big Data* су произведени од стране система који су дизајнирани и тестирани у одређеним научним оквирима и окружени мноштвом различитих контекста и интереса (Kitchin, 2016).

Треба имати на уму да *Big Data* никада не говоре сами за себе и неопходно их је систематизовати и на прави начин извући одговарајуће информације из њих за тачно дефинисан део популације.

Дакле, да би се *Big Data* користили као додатни извор података у званичној статистици, неопходно је да се испоштују све релевантне фазе *GSBPM* и да се испуне сви принципи кода праксе.

4.2 *Big Data* у УНЕЦЕ

УНЕЦЕ Група високог нивоа за модернизацију званичне статистике (*The UNECE High-Level Group for the Modernisation of Official Statistics*) *HLG-MOS* формирана је од стране Конференције европских статистичара 2010. године у циљу надгледања и координације међународних послова који су вези са модернизацијом статистике (UNECE, 2018). Модернизација статистике се одвија у смеру подршке праћењу индикатора одрживог развоја и концепта „*Data Revolution* за одрживи развој" (UNECE, 2018а).

Пројекти које је спроводила *HLG-MOS* потпуно су комплементарни са сличним иницијативама о којима ће бити речи у наредним деловима рад (ЕСС и УНСД).

У наставку су приказани резултати прва два пројекта спроведена од стране *HLG-MOS* на тему *Big Data :*

- Пројекат 2014
- Пројекат 2015

Као резултат оба ова пројекта један је главни закључак: „нема пуно случајева у којима *Big Data* као такви, могу бити адекватан извор података за званичну статистику, али они имају улогу да, уколико се користе заједно са другим изворима података дају бољи увид у текућа и нова питања." (UNECE, 2018а). То помера фокус рада *HLG-MOS* са *Big Data* на интеграцију са већим бројем извора података што је и предмет пројекта за 2016.

Иако се прелази на истраживање могућности коришћења вишеструких извора података, у овом делу рада приказаћемо резултате претходних пројеката и зашто су управо ти резултати довели до померања фокуса.

Као припремна фаза пројекта урађена је анализа текућег стања, класификације и само увођење овог термина у статистички речник, као и инвентар текућих и планираних пројеката из области *Big Data.*

У једном од првих радова насталих у припремној фази наводе се примери коришћења или планове за коришћење и тестирање *Big Data* у званичној статистици (UNECE, 2018). Примери који се у раду наводе су следећи:

- Статистика саобраћаја и статистика транспорта – Холандија
- Статистика друштвених медија – Холандија
- Статистика цена – Пројекат Евростата
- Статистика туризма – Студија изводљивости Евростата
- Коришћење информационо комуникационих технологија (ИКТ) – Студија изводљивости Евростата

Пример: Статистика саобраћаја и статистика транспорта

Пример који се овде наводи за случај Холандије у оквиру статистике саобраћаја је, касније резултирао у објављивању првих икада званичних статистичких података који су добијени на бази *Big Data* у Холандији (CBS, 2015.). У Холандији постоји преко 60.000 сензора (*road sensors*) од којих је 20.000 на аутопутевима који служе за бројање возила различитих величина, сваког минута. Приликом обраде ових података показало се да сам квалитет података значајно варира, како из минута у минут тако и из дана у дан. Како би се овај проблем превазишао развијени су одређени филтери који су подешени на стохастичко понашање приласка возила у оквир сензора. Кориговањем података и комбиновањем дневног „профила" које даје сензор на једном путу, покривеност и квалитет података су унапређени. На основу овога могуће је извести индексе стања на холандским путевима на регионалном нивоу (Daas, et al., 2014a).

Овај пример представља пример једног од типова извора *Big Data* података. Међутим, од идеје до пуне примене ових података прошло је неколико година. Изазови са којима су се аутори сусрели приликом развоја методологије и начине како ове изазове превазићи како би добијени подаци могли бити и публиковани приказани су у раду *High frequency Road Sensor Data for Official Statistic* (Puts, et al., 2015). У овом раду се објашњава да је филтрирање један од добрих начина за едитовање *Big Data* као и да је једноставно смањење димензија модела применом

метода главних компоненти (*Principal component analysis* - PCA) добар приступ за рад са овом врстом података.

Пример: Статистика друштвених медија

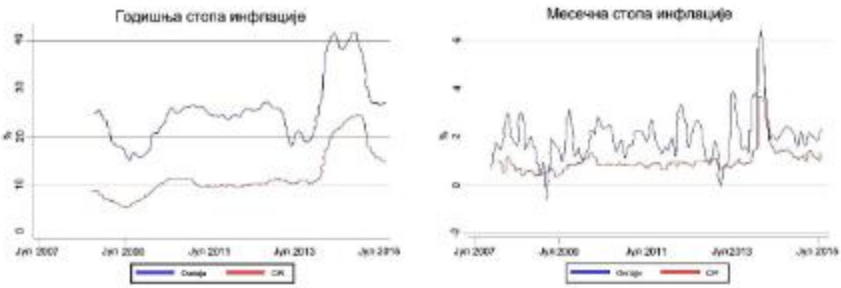
Овај пример је, као и претходни, на подацима из Холандије. У Холандији се дневно објави преко три милиона порука на јавним друштвеним медијима (Daas, et al., 2014a). На друштвеним медијима појединци размењују информације, учествују у дискусијама и комуницирају са својим пријатељима и породицом. Како би истражили које поље званичне статистике би могло бити покривено из овог извора података, све објаве су сагледаване из два угла: садржај и области. Најчешће коришћен медиј у Холандији је *Twitter* те су и коришћени подаци са овог медија. Анализа је показала да је око 50% „бесмислено брбљање", док је остатак порука био у вези активности у слободно време (10%), посао (7%), медији (5%) и политика (3%). Касније студије су показале да је расположење на холандским друштвеним медијима високо корелисано са поверењем потрошача. Посматрано расположење је било стабилно када се посматра на месечном или недељном нивоу али не и када се посматра на дневном нивоу.

Када је реч о друштвеним медијима као изворима податак за *Big Data* највећи проблем је покривеност популације. Узмимо за пример *Tweeter*. Ма колико он био глобално распрострањен и место где појединци изражавају своје ставове о мноштву тема, корисници *Tweeter* су само један, специфичан део друштва. Са друге стране, велики број корисника *Tweeter-a* нису активни већ пасивни корисници, односно не објављују своје „*Tweet-ove*" (ставове) већ само прате друге. Према студији *An Exhaustive Study of Twitter Users Across the World* која је спроведена на 36 милиона корисника, 25% корисника *Tweeter-a*, никада није *Tweet-овало*. Поред тога, 74% корисника је старости од 15 до 25 година старости, а само 6% спада у категорију старих 46 и више година. Отуда, преко *Tweeter-a* доступни су нам само подаци о специфичном делу друштва, не улазећи овде у специфични профил активних корисника (Beevolve, 2017).

Пример: Статистика цена

Пример који се овде наводи је коришћење *web scraping* технике за аутоматско прикупљање података о ценама са интернета. *Billion Prices Project* је пројекат Института за технологију у Масачусетсу (*MIT*) кроз који се корист цене сакупљене од стотина он лине продавница на дневном нивоу како би се спровела економска истраживања (Cavallo & Rigobon, 2016). Овај пројекат је настао као наставак *InflacionVerdadera.com* насталог у Аргентини 2007, који је имао за циљ израчунавање алтернативног индекса цена у Аргентини, користећи он лине податке о ценама два велика супермаркета у Буенос Аиресу. У свом раду, Кабаљо објашњава како је од 2003 инфлација у Аргентини константно расла као резултат експанзивне монетарне политике која је имала за циљ стимулацију потрошње и избегавање апресијације валуте (Cavallo, 2012). Тако је инфлација постало битно политичко питање током 2006. године. Комбинација субвенција и контролисаних цена није била у могућности да задржи ниво цена па је у Јануару влада донела драстичну одлуку: да реорганизује Национални статистички институт и отпусти лица која су до тада била задужена за обрачун индекса цена чиме је Институт изгубио на кредибилитету код медија и у академским круговима. Након ових драстичних мера, подаци о инфлацији су били знатно нижи док су истраживања која се баве очекиваном инфлацијом показивале чак три пута више вредности. Као одговор на то развијена је алтернативна методологија за мерење цена која користи онлајн податке, односно *Big Data*. Овај алтернативни индекс цена прати потпуно другачији тренд него што су званични подаци и вредности индекса су константно биле два до три пута веће, али са друге стране прате изненађујуће сличан образац кретања (Слика 4.1).

(а) Индекс цена

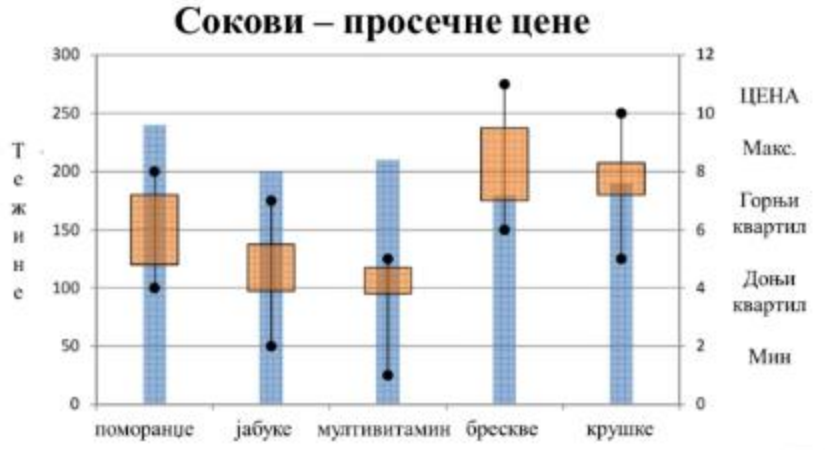


(б) Годишња стопа

(в) Месечна стопа

Слика 4.1: Поређење Индекса цена мерено онлајн и традиционалном методологијом на примеру Аргентине (Cavallo & Rigobon, 2016)

Евростат је уложио напор да дефинише методологију за вишенаменску статистику потрошачких цена (*Multipurpose consumer price statistics*) повећањем нивоа димензија прикупљених и објављених информација. Циљ је да уведе нови статистички производ „индикативни ниво цена" базиран на просечним (медијалним) нивоима цена, који би служио као допуна постојећим индексима цена. Квалитет тако добијених индикативних нивоа цена би био исказиван кроз мере дисперзије – квартилну девијацију (Слика 4.2) (Barcellan, 2013).



Слика 4.2: Пример индикативни нивоа цена (Barcellan, 2013)

За прикупљање података овако дефинисане мере цена, предвиђено је коришћење следећих извора:

I. Подаци са скенера, уз могућност прикупљања огромне количине детаљних информација о ценама производа, карактеристикама производа, као и обиму продаје производа. Такве информације треба да буду мапиране на ширем нивоу производа на структурирани модел података

II. Употреба електронских уређаја за традиционално прикупљање података за потребе израчунавања хармонизованог индекса цена (*Harmonised Index of Consumer Prices - HICP*) ХИЦП-а у комбинацији са истовременим прикупљањем додатних информација о карактеристикама производа на структурални (хармонизирани) начин

III. Употреба прикупљених цена са интернета, као и сродних информације у складу са дефинисаним моделом података.

IV. Комбинација детаљних микропoдатака (прикупљених у складу са три претходна приступа) с методологијом PPP-а (*Purchasing power parity*) која осигурава већу географску и временску покривеност.

Предложена методологија је тестирана кроз пилот студију. Добијени резултати - цене за 156 производа, и закључци истраживања су публиковани у децембру 2012. године (Eurostat, 2012).

Такође, планиран је и развој напредног *open source* софтвера за *web scraping* који је препознат као широко примењива метода у прикупљању *Big Data* за потребе статистика цена у статистичким институтима (Boettcher, 2015).

Пример: Статистика туризма

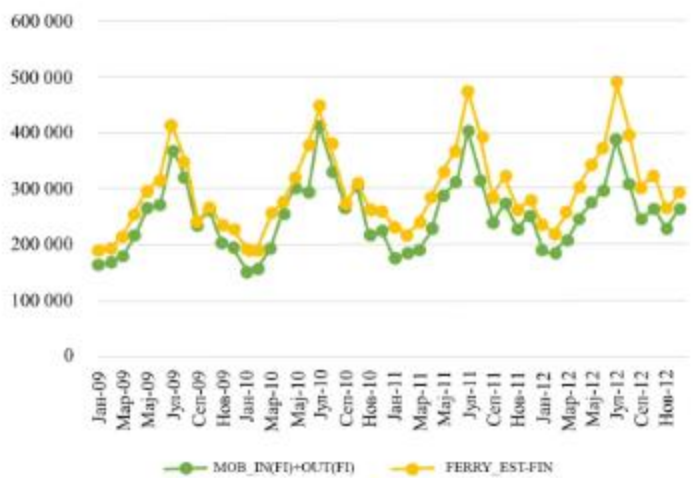
За потребе истраживања потенцијала података о позицији мобилних телефона у статистици туризма као и оцена њихових предности и недостатака израђена је студија изводљивости коју је наручио Евростат. Главни закључци ове студије били су (Ahas et al., 2014):

- Подаци о позицији мобилних телефона су веома лимитирани пре свега због законодавних ограничења. Неопходни су централизовани оквири за Националне статистике и друге заинтересоване стране како би се легално приступило овим подацима
- Неопходни су лонгитудинални подаци како би се прикупили што прецизнији подаци о кретању претплатника.
- Податке мобилне телефоније је боље користити као додатни него као јединствен извор података за обрачун индикатора из области туризма.
- Коришћењем овог извора као додатног, побољшава се правовременост, омогућен је приступ подацима који су раније били недоступни (чиме је омогућен обрачун нових индикатора), побољшавају се могућности за калибрацију података, боља „резолуција" (подаци на нижем нивоу) и већа прецизност у простору и времену.
- И друге статистичке области могу имати корист од овог извора података.

Методологија која је основ ове студије изводљивости била је методологија за статистику туризма развијена од стране конзорцијума који је учествовао у пројекту

- и карактеришу је комплетност (географска и временска покривеност је много богатија него код традиционалних статистичких метода), прецизност, конзистентност, ваљаност (дефиниције су у сагласности са ЕУ регулативом о статистици туризма), правовременост, смањена оптерећеност (и извештајних јединица и буџета) и приватност података (Ahas, et al., 2010). Све ово су захтеви који сваки статистички податак мора да испуни па је одавде јасно да подаци мобилне телефоније могу бити адекватан кандидат за допунски или алтернативни извор података у званичној статистици.

Илустрацију добијених података можемо сагледати на слици 4.3 која приказује упоредни преглед улазних месечних путовања од Финске ка Естонији и излазних од Естоније ка Финској добијених на основу података о позиционирању са мобилних телефона и актуелног броја путника који су се превозили бродом (феријем), добијеним од стране Финске транспортне агенције.



Слика 4.3: Упоредни преглед улазних месечних путовања од Финске ка Естонији и излазних од Естоније ка Финској добијених на основу података мобилне телефоније (MOF_IN(FI)+OUT(FI) и Финске транспортне агенције (FERRY_EST-FIN) (Ahas et al., 2014)

На основу месечних временских серија за приказани период може се закључити да се на основу података о позиционирању са мобилних телефона може добити веома

конзистентна процену укупног броја путника са долазних и одлазних путовања. На основу додатне анализе закључено је да мобилном позиционирање потцењују стварне количине путника, због необухвата других националности и транзитних путника на броду.

Пример: Коришћење информационо комуникационих технологија (ИКТ)

Идеја ове студије изводљивости је да се тестирају могућности коришћења података са интернета за израду статистика из области коришћења информационо комуникационих технологија. Студија је требала да покаже који се подаци могу прикупити са интернета, односно са веб сајтова. Процес прикупљања података одвијао се у два смера у зависности од јединица посматрања:

- Домаћинства: испитаници су замољени да инсталирају програм за мониторинг а подаци су слати Националним заводима за статистику.
- Пословни субјекти: „жетва" података са веб сајтова (у другој фази ови подаци су допуњени подацима са програма за мониторинг или фајлова о логовању на сервер)

УНЕЦЕ-ов пројекат из 2014. године – Улога *Big Data* у модернизацији производње статистичких података имао је за циљ да идентификује и дотакне главне изазове употребе извора *Big Data* у званичној статистици и има је три главна циља (UNECE, 2014):

- Да идентификује, испита и да смернице за статистичке организације о главном стратешким и методолошким питањима које концепт *Big Data носи са собом*.
- Да покаже изводљивост ефикасне производње како нових производа тако и базичних статистичких података применом *Big Data* као и могућности да се ови приступи на исти начин примене у различитим националним контекстима.
- Да олакша трансфер знања, експертизе, алата и метода за производњу података користећи *Big Data*.

Овај пројекат је од великог значаја јер је, можда по први пут, указано на неопходност глобалног приступа изазову који са собом носи употреба *Big Data* у званичној статистици. Ово подразумева чињеницу да је немогуће овим проблемом бавити се изоловано већ је неопходно удруживања како појединаца са специфичним знањима тако и институција. Али, са друге стране и размена искуства је овде стављена у први план, коко не би дошло до расипања знања или дуплирања рада. Алати развијени овим пројектом стављени су на располагање широј заједници пре свега кроз увођење *Sandbox-a,* заједничко окружење за израчунавање које служи међународној сарадњи по питању *Big Data.* Ово окружење које је доступно преко веб-а за чување и анализу великих сетова података креирано је и коришћено као платформа за колаборацију институција које су биле партнери на пројекту (Jug, et al., 2016). Такође, *Sandbox* окружење је омогућило и искоришћавање потенцијала *Big Data,* кроз технолошке иновације, нове методологије, партнерства и вештине, што би за појединачне статистичке организације било тешко да мобилишу на тако високом нивоу.

Као један део пројекта, спроведено је и неколико експеримената примене *Big Data* у званичној статистици (UNECE, 2014):

- Подаци са паметних бројила,
- Индекс потрошачких цена,
- Подаци о отвореним радни местима,
- Употреба мобилних телефона,
- Истраживање ставова коришћењем податка са социјалних медија,
- Статистика саобраћаја и
- Коришћење ИКТ.

О ником од ових експеримената већ је било речи, што само потврђује чињеницу да је кооперација и повезивање експертизе кључна за успешне резултате у примени *Big Data.* Овде ћемо навести неке од основних резултата експеримената о којима није било раније речи.

Пример: подаци са паметних бројила

> Паметна бројила су електронска бројила која омогућавају аутоматско прикупљање података о потрошњи електричне енергије у домаћинствима и малим пословним субјектима. Ова паметна бројила омогућавају увид у потрошњу електричне енергије у сваком тренутку. Поред података о потрошњи струје коришћени су и подаци о температури из метеоролошких станица као и подаци о цени струје у датом времену који су дефинисани од стране дистрибутера електричне енергије.

Кроз пројекат тестирано је шест димензија квалитета података. Резултати добијени из овог извора података имају велику предност у односу на актуелне изворе података најмање из два разлога. Као прво, уколико је омогућен пун приступ овим подацима, подаци нису подложни типичним грешкама које се праве у класичним анкетним истраживањима, као што су грешке испитаника, грешке анкетара, пристрасност оцена услед неодговора и због грешака необухвата, грешке у обради и импутацији података. Грешке које су могуће због механичких кварова бројача су смањене на минимум имајући у виду да је у интересу корисника да бројачи буду што прецизнији а, са друге стране, уколико и дође до грешке на неком од мерача, захваљујући великом броју података ова грешка је, углавном занемарљива и не утиче на резултат. Овим је омогућена прецизна оцена, што је један од предуслова квалитета података званичне статистике. Подаци добијени на овај начин су веома релевантни, јер могу бити приказани чак и на много нижем нивоу него што је случај са анкетним подацима, а додатно могу бити упарени и са подацима о времену и цени струје. Имајући у виду да су подаци доступни из сата у сат, они на тај начин могу бити и обрађивани чиме се обезбеђује правовременост података (практично у реалном времену). За ове податке је могуће израдити метаподатке чиме се омогућава интерпретативност података. Такође, ови подаци лако могу бити повезани са другим изворима података, нарочито уколико се располаже геопросторним подацима домаћинства (кохерентност). Коначно, приступачност овим подацима зависи од типа уговора између статистичке организације и власника података. Свакако приступ агрегираним подацима би требало омогућити широј јавности (UNECE, 2014).

Примена бројача би требало да буде шире разматрана имајући у виду предности ИоТ и чињенице да у данашње време, многи кућни апарати комуницирају што међусобно, што са неким централним сервером. *People metre-и* су одавно у примени у маркетиншке сврхе али би се могла размотри њихова примена и за потребе друштвених статистика у оквиру званичне статистике.

Пример: Подаци о отвореним радни местима

> Подаци о отвореним радним местима најчешће су доступни на веб страницама за запошљавање, ко што су странице националних завода за запошљавања или приватних веб портала који се баве овим питањима (у случају Србије највећи портал за новоотворене позиције је, поред сајта НСЗ, Infostud, http://www.infostud.com/). Идеја експеримента је била да се користећи *web scraping* технику, *scrap-yjy* подаци са највећих сајтова за запошљавање.

Код овог метода прикупљања података поставља се питање грешака мерења. Такође приступ подацима није лак. Са друге стране, висока правовременост података је предност јер је могуће поставити процес који прикупља и чисти податке и аутоматски рачуна индикаторе на недељном нивоу. За имплементацију овог концепта неопходно је остварити партнерства са власницима портала јер је тиме омогућено прикупљање детаљнијих и тачнијих података (UNECE, 2014a).

У фокусу УНЕЦЕ-овог 2015 пројекта били су експерименти, *Sandbox,* и тренинзи. У оквиру пројекта спроведено је четири експеримента (UNECE, 2018a):

- Прегледи на онлајн енциклопедији *Wikipedia*
- Трговина на бази података са *UN ComTrade* базе података
- Подаци са *Twitter-a*
- Подаци са сајтова пословних субјеката

Пример: Прегледи на онлајн енциклопедији *Wikipedia*

> Овај *Big Data* пројекат се фокусирао на прегледе страница посвећених местима који припадају UNESCO светској заоставштини и анализирано је укупни 1.068

места а праћени су прегледи од 2012-2015 унутар *Wikipedia-e* као најпосећеније онлајн енциклопедије. Према званичној ЕУ статистици, 44% појединаца у доби од 16 до 74 године који живе у ЕУ консултовали су викије (*wikis*), нарочито *Wikipedia-e* у циљу постизања знања у 2013 (Beręsewicz, et al., 2018). Циљ пројекта је био да се открију трендови у прегледима одређених страница како би се проверила евентуална корелација са активностима из области туризма. Добијени су подаци који су релевантни за статистику културе и регионалне статистике. На тај начин покривени су сегменти који раније нису били покривени статистиком културе (UNECE, 2018a).

Ниво прецизности зависи од тога за који се конкретан индикатор користи. Ако је индикатор „број посетилаца на страници" онда је прецизност велика, али ако је индикатор „укупан број посетилаца, онда то није случај (Beręsewicz, et al., 2018). Резултати су доступни након неколико сати чиме се обезбеђује правовременост извештавања. Подаци о посетама страница су јавно доступни на сајту *Wikipedia-e* и добро документовани чиме је омогућена једноставна интерпретација резултата (Интернет 2). Подаци су упоредиви у простору и времену иако се не могу сматрати комплетним јер нису у могућности да одговоре на све потребе корисника.

4.3 *Big Data* у ESS

Имајући у виду комплексност концепта *Big Data* и специфична знања и регулативе неопходне за његову имплементацију у систем званичне статистике, на нивоу Европске уније одлучено је да се уложе заједнички напори свих националних статистика како би се заједничким радом тестирале могућности коришћења овог концепта. Иницијална каписла за почетак рада на ову тему је Меморандум из Шевенингена који је усвојен у септембру 2013. године од стране Савета ЕСС (Eurostat, 2013). Ово је почетни документ на који се ослања сав будући рад ЕСС на тему *Big Data*. Овим меморандумом наглашена је растућа потреба за правовремени и економичним а истовремено високо квалитетним подацима, као и потреба за новим решењима за смањење стопе неодговора. Стога, националне статистике треба да укључе, у мери у којој је то могуће, све изворе података, укључујући и *Big Data* у концептуални развој података које задовољавају потребе модерног друштва. Након усвајања Меморандума у оквиру Евростата оформљена је Радна група (*Task*

Force) за *Big Data*. Од усвајања Меморандума креће развој *Big Data* у званичној статистици ЕУ. Први велики догађај који је организован од стране ЕСС на ову тему, одржан је у Марту 2013 у Риму. Састанак је имао за циљ имплементацију Меморандума кроз остваривање следећих циљева (Cervera, 2014):

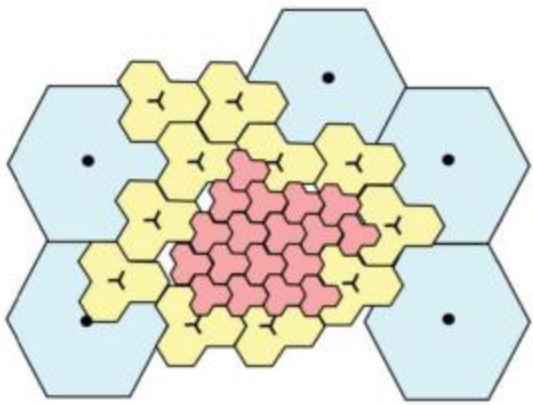
- Повећање свести о изазовима *Big Data* у оквиру ЕСС и проналажење начина за њихово решавање.
- Идентификовање и размена најбољих пракси у вези коришћења *Big Data* у званичној статистици
- Идентификовање синергија између ЕСС, приватног сектора и академије, где заједнички напори могу бити од интереса за националне статистике или за читав ЕСС
- Идентификовање законских, техничких и методолошких захтева који морају бити испуњени да би се *Big Data* користили у званичној статистици.
- Пружање експертске помоћи при изради акционог плана од стране ЕСС и његове имплементације кроз Годишњи програм званичне статистике Евростата.

Током састанка представљени су први пилот пројекти на којима је рађено:

- „Лица и места" (о мобилности лица коришћењем података мобилне телефоније) (Ricciato, et al., 2016).
- „Оцена тржишта рада" (о прогнозама стопа незапослености коришћењем *Google Trends*) (Perduca, 2014).
- „Употреба ИКТ у привредним субјектима" (коришћењем техника *webscraping*-а и *text mining*-ом).

Рикато и др. (2016) су у студији која се бави мерењем густине насељености на бази употребе мобилних телефона описали технологију на бази које се прикупљају подаци о коришћењу мобилне телефоније. Наиме, инфраструктура оператера мобилних мрежа (ОММ) састоји се из великог броја „ћелија" различитих величина, које покривају простор који може бити величине од 10 метара до неколико километара. Ове ћелије емитују сигнал који примају мобилни телефони тако да након било ког „догађаја" – примање или упућивање позива, примање или слање

SMS-ова – мобилни телефон открива у оквиру које ћелије се налази и ова информација се заувек чува у такозваној *Call Detail Record (CDR)* бази података како би „догађај" био наплаћен. Поред овога мрежа бележи и прелазак из једне у другу ћелију.



Слика 4.4: Пример покривености вишеслојним ћелијама са повећањем величине ћелија (и смањењем густине ћелија) од унутрашњости ка спољашњим деловима града (Ricciato, et al., 2016)

Циљ ове студије је био да се изгради методолошки оквир за прикупљање и обраду података мрежних оператера који може бити примењен и за више ОММ. Приликом обраде ОММ података треба имати увиду да пресликавање *лице: мобилни телефон* није увек 1 на 1 (не слика се увек једно лице на један мобилни телефон) већ су могуће следеће ситуације:

1. Пресликавање 1 на 1: идеални случај, где једно лице носи један мобилни телефон
2. Пресликавање 1 на више: где једно лице може носити више мобилних телефона (или других уређаја која примају сигнал мобилне телефоније)
3. Пресликавање 1 на 0: где нека лица не носе мобилне телефоне

4. Пресликавање 0 на 1: где уређај који прима сигнал не одговара ни једном лицу већ сигнал примају машине за међусобну комуникацију (*machine-to-machine - M2M*)

У 1. и 4. случају долази до превеликог обухвата. Учесталост за пресликавања 1 на више и 1 на 0 зависи од демографских карактеристика. Аутори дају предлоге како се ово може, макар делимично превазићи. У овом раду показани су, визуално, резултати два сценарија. Оба сценарија су резултат симулација тако да аутори напомињу да се „резултати не могу узети као дефинитивни доказ перформанси модела у реалном свету али су свакако информативни и дају иницијалне индикације о томе шта се може очекивати као резултат у практичној примени".

Други пример употребе *Big Data* у званичној статистици је рад Perduca, (2014), који је користио *Google trends* ради побољшавања предикције стопе незапослености, полазећи од једноставног ауторегресионог модела који укључује једну доцњу стопе незапослености. Подаци су коришћени су на месечном нивоу. Почетни ауторегресиони модел је облика (Perduca, 2014):

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

Где је y_t незапосленост у месецу t, a и b коефицијенти, а y_{t-1} незапосленост у месецу *t-1*. e_t је грешка модела.

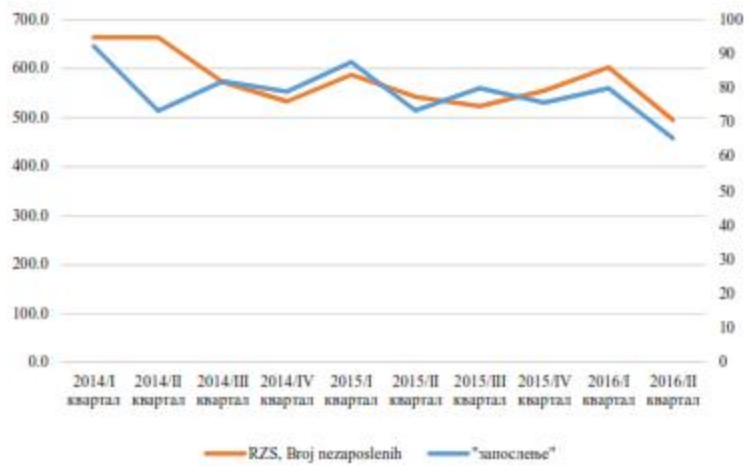
Након оцене оваквог модела, у модел су укључене додатне варијабле које представљају индекс претраживања три различита термина на *Google trends*, а доводе се у везу са стопом незапослености. То су термини:

- „pole emploi" – француска агенција за запошљавање
- „etre au chomage" – бити незапослен
- "indemnité" – џепарац

Као закључак оцењивања модела је да укључивање додатних предиктора побољшава квалитет модела. Међутим, јасно је да је наведени модел поједностављен и да је неопходно обавити комплекснија оцењивања и комплексније моделе, који би укључили и сезонску компоненту у обзир.

Међутим, овај модел се базира на предикцији која укључује званичне податке из претходног месеца, а информације са *Google trends* се користе у циљу побољшања модела. Било би занимљиво оценити регресиони модел у коме је зависна променљива број незапослених а независне променљиве само подаци о претрагама различитих термина са *Google trends*-а. Овде највећи изазов представља одабир термина за који би се посматрао тренд претраге. Овде свакако треба имати у виду и сезонску компоненту.

Навешћемо овде најједноставнији пример претраге термина „запослење" на територији Републике Србије у периоду од 2014 с обзиром да званични подаци о броју незапослених постоје за тај период. Како су званични подаци за Србију дати на кварталном нивоу, и за претрагу на *Google*-u узели смо просту аритметичку средину претраживања овог термина. Резултат је дат на слици 4.5.



Слика 4.5: Упоредни приказ претраге термина „запослење" и званичних података о броју незапослених у Републици Србији

Са слике 4.5 се види да постоји веза између ове две варијабле. Коефицијент корелације за ове две варијабле износи 0.54, што је солидно високо, имајући у виду да је посматрана једноставна линеарна веза.

За наведену анализу коришћен је апроксимативни механизам самоселекције (*self-selection mechanism*) који се спроводи у три фазе и базира се на три нивоа покривености: укупна интернет популација, популација интернет корисника (у последња 3 месеца) и популација корисника *Google-a* (Слика 4.6) (Beręsewicz, et al., 2018).

Слика 4.6: Апроксимативни механизам самоселекције у коришћењу Google trends-a (Beręsewicz, et al., 2018)

Истраживање могућности *Google trend* апликације настала је заједно са настанком ове апликације 2006. године. *Google* је 2009 на свом истраживачком блогу објавио један од првих покушаја примене индикатора базираних на *Google trend* апликацији за предвиђање социо-економских индикатора. И овде је, слично као у раду о коме је било речи претходно коришћен ауто-регресивни модел, укључујући доцњу од 1 и од 12 месеци како би се искључио утицај сезоне. Циљ овог модела није предвиђање будућност већ садашњости. Модел је коришћен за оцену продаје моторних возила, незапосленост, планирање дестинације за путовање и поверење потрошача (Choi & Varian, 2009).

Закључак рада је био да укључивање података са *Google trend* побољшава перформансе модела. Имајући у виду веома високу правовременост података са *Google trend-a* коришћење ових података могао би у великој мери да помогне оцењивању индикатора у много брже него што је то био случај раније.

На крају овог прегледа, навешћемо да су Светска банка и Статистичка дивизија УН формирали су глобалну радну групу за примену *Big Data* у званичној статистици

(*Global Working Group - GWG*) која је израдила је каталог *Big Data* пројеката који су релевантни за званичну статистику, индикаторе одрживог развоја и друге статистике потребне за доношење одлука о јавним политикама, као и за управљање *и праћење програма/пројеката јавног сектора (GWG, 2018).*

"У суштини, сви модели су погрешни, али су неки корисни."

Box & Draper (1987)

5 Развој методологије статистичког истраживања на бази *Big Data* концепта

5.1 Основне напомене

Појава *Big Data* навела је научну заједницу да преиспита до тада важећу (традиционалну) методологију научног истраживања и тако покрене револуцију у научном размишљању и методама. Познато је да су најранија научна истраживања у људској историји била заснована на експериментима. Касније се појавила теоријска наука која је карактерисала проучавање различитих закона и теорема. Међутим, пошто је теоријска анализа сувише сложена и није увек примењива за решавање практичних проблема, научници су почели са применом метода заснованих на рачунарској симулацији, што је довело до развоја рачунарске науке. *Big Data* концепт је увео нову парадигму: у раду са подацима великог обима, истраживачи могу да откривају одређене законитости (информације, знање и интелигенцију) само на основу увида и „рударења" података (*data mining*). При томе, они чак ни не морају директно да приступе предметима који се проучавају (Hey, 2008; Jin, at al. 2015).

Ова парадигма је у стручној и научној јавности, почетком овог миленијума, названа четвртом парадигмом научних истраживања, која је заправо раздваја науку која се бави подацима и рачунарску науку– Табела 5.1.

Табела 5.1: Еволуција науке, према парадигмама

Парадигма	Трајање	Назив парадигме	Кратак опис	Пример
1.	Хиљаду година	Експериментална наука	Опис природних феномена	Посматрање природних појава
2.	Последњих неколико стотина година	Теоријска наука	Проучавање различитих закона и теорема	Њутнови закони, Максвеллова једначина...
3.	Последњих неколико деценија	Рачунарска наука	Симулација комплексних феномена	Симулациони модели
4.	Данас	Наука о интензивној употреби података	Коришћење велике количине података	*Big Data*, наука о подацима

Добитник Туринговe награде (*Turing Award*), која се сматра неком врстом Нобелове награде у области рачунарства, истиче да је четврта парадигма можда једини системски начин за решавање неких од највећих глобалних изазова с којима се данас суочавамо, и да четврта парадигма није само промена у начину научног истраживања, већ и промена у начину на који људи мисле (Hey et al., 2009).

"Данашњи градови и владе и даље функционишу у складу са принципима који су развијени пре два века, током индустријске револуције. Да бисмо решили проблеме 21. века као што су експлодирајући раст становништва и климатске промене, потребан је нови начин размишљање, које може да нам пружи *Big Data*. Дигиталне мрвице за хлеб, које остављамо иза себе у свакодневном животу - које откривају о нама више него о што смо били спремни да откријемо - пружају моћно средство за решавање многих друштвених проблема" (Pentland, 2013).

Као илустрацију, навешћемо проблем деменције, за који је на глобалном форуму наведено да представља један од највећих глобалних изазова човечанства данас, уз напомену да је деменцијом погођено око 44 милиона појединаца, очекује се да ће се до 2030. године број оболелих удвостручити и утростручити до 2050. године (Deetjen, et al., 2015).

Тренутно не постоји лек за деменцију или поуздан начин за успоравање напретка, а министри здравља земаља Г8 поставили су циљ проналажења лека или терапије за модификовање болести до 2025. године. У том циљу на глобалном самиту посвећеном деменцији (*G8 Global Dementia Summit*) у децембру 2013. године, као сет приоритета наведено је: боље коришћење доступних података, дељење ресурса и сарадња истраживача. У том правцу, министри здравља Г8 дали су мандат OECD -у да извештава о томе како се *Big Data* могу користити у циљу ефикаснијег истраживање деменције (Fox & Petersen, 2013).

Нова парадигма доводи и до одређених опречних ставова. Тако се за *Big Data* често наводи да представља праксу без теорије (*„The no theory thesis"*), док неки аутори чак иду до тога да наводе очекивање и/или страховање да *Big Data* представља крај теорије у науци (Leonelli, 2014; Mazzocchi,2015).

Андерсон, уредник *Wired magazine*, је још 2008. године изнео став да ће у ери петабајтних информација и суперкомпјутера, традиционално статистичко закључивање базирано на тестирању хипотеза, као научна метода постати застарела. Он сматра да нема више потребе за теоријом и хипотезама, као и дискусијом да ли експериментални резултати оповргавају или подржавају постављене хипотезе. У новој ери, оно што је битно јесу софистицирани алгоритми и статистички алати, способни да се из масовне количине података пронађу информацију која се може претворити у знање (Anderson; 2008).

На тај начин он уводи дискусију на терен поређења: истраживање засновано на подацима (*Big Data* приступ) насупрот истраживању заснованом на тестирању хипотеза (традиционални приступ) (*data-driven research versus hypothesis-driven research)*.

Присталице *Big Data* приступа држе се индуктивних алгоритама који се користе у већини експертних система и машинском учењу, за разлику од дедуктивних метода који се већином користе у традиционалним истраживањима. Индуктивно расуђивање генерално не даје финални статус. Резултати закључивања вероватно ће променити претходне закључке. Код *Big Data* приступа могуће је наставити расуђивање до бесконачности. Најбољи индуктивни алгоритми могу да еволуирају:

они "уче", редефинишу начин обраде података према најпримеренијој употреби која се може направити. Истиче се да стално учење, које се никада не завршава, производи несавршено али корисно знање, те да сличност са људским мозгом сигурно није случајност (Malle, 2013).

Као основна предност *Big Data* приступа наводи се да процес изградње модела заснованог на великој количини података у мањој мери зависи од теоријских претпоставки и ограничења.

Приликом дефинисања методолошког поступка статистичког истраживања на бази *Big Data* концепта узели смо у обзир Кантов предговору делу *Metaphysical Foundations of Natural Science* (1786), где је истакао колико је наука рестриктивна, указујући на захтев да се научна спознаја систематски уреди, у складу са рационалним принципима (Watkins & Marius, 2014).

Такође, извесно је да процес изградње модела вођен масовном количином произведених података и мање зависи од теоријских претпоставки и постављених хипотеза. Ипак, то не значи да ће *Big Data* у потпуности заменити когнитивне и методолошке поступке, који се примењују у унапређују кроз векове филозофске и научне мисли. Дакле, може се закључити да у овом тренутку није реално говорити о "крају теорије", већ о истраживању нових могућности.

Имајући у виду предмет истраживања дисертације, а узевши у обзир основне принципе система званичне статистике и Кодекс праксе европске статистике који укључују научне методе и методологију у статистичку праксу, имплементација концепта *Big Data* у процес статистичких истраживања мора се базирати на научним принципима и методама (UNSTATS, 2015; Eurostat, 2017a).

Тако је постављена главна хипотеза ове дисертације, која гласи:

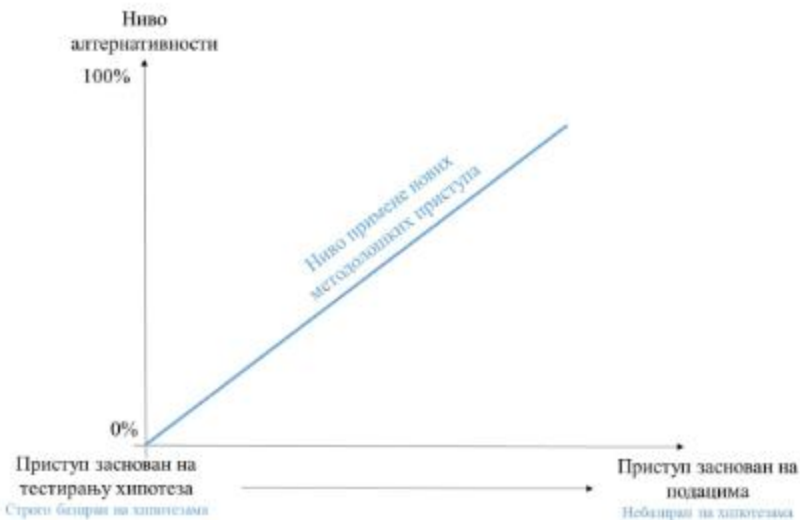
„Имплементацијом концепта *Big Data*, унапређује се процес статистичких истраживања, отварају се могућности примене нових метода и техника истраживања, олакшава се процес анализе великих количина података и побољшавају укупне перформансе система званичне статистике".

Једна од полазних хипотеза истраживања је:

Х0.2.1. *Big Data* ресурси се могу користити као алтернативни извори података за званичну статистику.

Сам појам алтернативности у овом случају претпоставља да *Big Data* не могу у потпуности заменити традиционална статистичка истраживања у свим областима, нарочито у домену система званичне статистике (Lazer, et al., 2014).

Ниво алтернативности може ићи од 0 до 100% (Слика 5.1).



Слика 5.1: Ниво алтернативности извора података и методолошки приступ

- 0% претпоставља да *Big Data* не може да представља суплемент статистичких података, што указује да ће се користити традиционална статистичка истраживања (*hypothesis-driven research*) и одговарајуће статистичке методе (дескриптивне анализа, статистичко оцењивања и закључивање) у случајевима када их је могуће применити (Tam & Clarke, 2015a).
- 100% претпоставља да *Big Data* у потпуности замењује традиционални статистички извор података и примењиваћемо алтернативне поступке

индуктивног резоновања у форми технолошког емпиризма где ће машинско учење и вештачка интелигенција водити до аутоматизованог откривања знања - истраживање засновано на (вођено) подацима (*data drivern approach*). Према томе, овај приступ добија атрибут "небазиран на хипотезама" (*hypothesis-neutral*), као начин стварања знања који је у стању да замени традиционална статистичка истраживања заснована на хипотезама. Анализа огромних количина података ће дати нове и често неочекиване корелације, обрасце и правила, које ће се у итеративном поступку смењивати новим, унапређујући претходно стечена знања.

- Између ова два екстремна приступа, евидентан је и мешовити приступ, који подразумева примену традиционалних статистичких поступака и модела над великом количином података (*Big Data*). У овим случајевима фокус је на прилагођавању стандардних статистичких модела великим подацима, чија величина прелази капацитет једног рачунара односно одговарајућих софтверских решења, било да је реч о ограничењима везаним за меморију рачунара, ограничењима статистичких алата или предугој обрада података (Wang, et al., 2018; Drovandi, et al, 2017).

5.2 Основне претпоставке за реализацију модела

Као сет изазова који се постављају пред званичне статистичке институције могу се класификовати на следећи начин (Vaccari, 2014):

- Законодавни, нпр. о приступу и коришћењу података
- Приватност података, нпр. управљање јавним поверењем и прихватање поновне употребе података и њихово повезивање са другим изворима
- Финансије, нпр. потенцијални трошак коришћења података наспрам предности за њихово коришћење
- Управљање, нпр. политике и директиве о управљању и заштити података
- Методологија, нпр. квалитет и одрживост статистичких метода
- Технологија, нпр. питања у вези са информационим технологијама

У складу са наведеним изазовима, дефинисали смо предуслове за развој методологије која омогућава имплементацију *Big Data* модела процеса статистичког истраживања, који се огледају у следећем (Čomić, et al., 2017):

1. Законодавни и стратешки оквир
2. Приступ подацима
3. Партнерства
4. Кадрови – *data scientists*
5. Технолошка инфраструктура

5.2.1 Законодавни и стратешки оквир за примену *Big Data* у званичној статистици

За пуну примену *Big Data* концепта у систему званичне статистике, неопходно је обезбедити законодавни и стратешки оквир (Cheng, 2014). Овај оквир се доноси на државном нивоу, за шта постоје примери из најразвијенијих земаља (Hajirahimova & Aliyeva, 2015).

Политика Сједињених Америчких Држава о *Big Data* - Иницијатива за истраживање и развој концепта *Big Data*. *Big Data* се, од стране научне заједнице, пословних удружења и државних органа у САД и у једном броју западних земаља, већ сматра за стратешки ресурс, баш као и нафта, и велики значај се даје проблемима у овој области. У марту 2012. године, Обамина администрација је званично покренула Иницијативу за истраживање и развој концепта *Big Data* којом је опредељено чак 200 милиона долара намењених за развој алата и техника за приступ, организовање и складиштење података, садржаних у огромним количинама дигиталних података. Иницијатива има за циљ проучавање нове инфраструктуре и методологија за *Big Data* истраживања како би се знатно олакшала примена алата и техника за стицање знања и информација из *Big Data* уз побољшање могућност њиховог коришћења за научна открића (Kalil, 2012).

Стратегија о *Big Data* аустралијске јавне службе.

У августу 2013., аустралијска савезна Влада објавила је Стратегију о *Big Data*. Стратегијом се промовише реформа услуга јавног сектора кроз коришћење *Big*

Data, развијање бољих јавних политика и заштиту приватности грађана како би Аустралија била међу светски најнапреднијима земљама у пољу *Big Data.* Ова визија има за циљ да подржи напредне и нове услуга као и пословне могућности партнерства, побољшани развој политика, као и заштита података о личности и усклађивање улагања Владе у ИКТ (AGIMO, 2013).

Сервис података Велике Британије. Овај сервис се фокусира на то да учини податке доступним у циљу стварања позитивних ефеката у сфери истраживања, економије и политике. Наиме, они су фокусирани на поновну употребу постојећих сетова података у циљу информисања креатора политика, генерисања нових становишта гледања и покретања дебата. Политика отворених података (*Open data*) у Великој Британији је прошла фазу покретања и пажња је проширена на квалитет података. Сервис података Велике Британије је финансиран од стране Савета за економска и социјална истраживања *(Economic and Social Research Council - ESRC)* Велике Британије (Open Data Strategy, 2014).

Политика отворених података – Република Ирска

Политика отворених података (*Open Data Strategy*) 2017 – 2022 обезбеђује оквир за податке којима се може приступити како би се осигурало да јавне услуге буду испоручене на ефикаснији, транспарентан и одговоран начин. Два основна циља Стратегије су дисеминација јавних података у власништву владе у отвореном формату у циљу слободне употребе и ангажовање шире јавности (универзитета, института и осталих) како би се промовисала употреба ових података у корист свих сектора економије (Open Data Strategy, 2017).

Политика *Big Data* у Француској. У оквиру међувладиног семинара о дигиталној економији која је одржана 28. фебруара 2013. године, француски премијер представио мапу пута владе за ову индустрију. Француска влада је, кроз *Програм инвестиција за будућност,* доделила 11,5 милиона евра за 7 пројеката из области обраде *Big Data* (Vellaipandiyan, 2013).

Јапанска Стратегија интегрисаног ИКТ за 2020. и Декларација у циљу постајања светски најнапредније ИТ-нације. Ова документа усвојена су 2012. и

2013. године, респективно (Doyukai, 2011). Такође, Савет за регулаторну реформу је 2013. године поставио смернице за јапанске компаније за коришћење *Big Data* без кршења закона о приватности. Истовремено, влада је издвојила је 13,2 милијарде јена за истраживање и развој у области *Big Data* (Japan, 2013).

Политика *Big Data* у НР Кини. Јуна 2014. године, кинеска народна политичка консултативна конференција одржала је форум у Пекингу о томе како користити *Big Data* технологију за побољшање способности управљања. У новембру 2013. године, Национални биро за статистику Кине потписао је низ споразума са 11 највећих кинеских предузећа, укључујући *Baidu, Alibaba, China Unicom,* и сл. са циљем максимизације ефеката примене *Big Data* (Damin & Jinjing, 2014).

Политика *Big Data* у Републици Кореји. Република Кореја је најдаље отишла када је реч о стратешки циљевима. Савет Председника за националну ИКТ стратегију је 2011. године покренуо радну групу за *Big Data* у оквиру *Big Data* иницијативе. Ова иницијатива има за циљ успостављање (Hajirahimova & Aliyeva, 2015):

- свевладину *Big Data* мрежу и аналитички систем;
- размену података између владе и приватног сектора;
- дијагностички систем за јавне *Big Data*;
- технике и технологије управљања и обраде *Big Data.*

Иницијатива је укључила и неколико јавних и приватних партнерстава са националним институтом Кореје у оквиру Националне агенције за информатичко друштво - *Big Data Strategy Center* и академским институтом у оквиру Националног Универзитета у Кореји - Институтом за *Big Data.* У новембру 2012. године, Национална комисија за науку и технологију науке и технологије развила је Мастер план за *Big Data.* У Јужној Кореји је 2013. године основан *Big Data* институт, са циљем да се оствари напредак у глобалној трци насталој у домену образовања за *Big Data* и то кроз развој мултидисциплинарних истраживања у овој области (SNU, 2013).

Као део *Smart nation government* утврђен је Пилот пројекат за Примену *Big Data* у званичној статистици (Ahn & Hwang, 2013). У јулу 2013. године, корејска влада је

најавила трећу верзију Мастер плана, обезбеђујући тиме нову парадигму за функционисање владе у циљу да се постигне власт орјентисана на грађане, транспарентна и способна власт која комуницира са грађанима. Национална влада такође планира да користи *Big Data* као део предстојећег пописа становништва, што ће, према проценама донети уштеду од 140 милиона долара (Desai & B. Nuño, 2015),

Статистика Холандије се све више фокусира на употребу *Big Data* у производњи званичне статистике. Званична статистичка агенција Холандије улаже много у *Big Data* са циљем да постану глобални лидер у области статистике *Big Data*. Један корак који има за циљ да додатно уобличи те амбиције је званично представљање новог Центра за *Big Data*, 2016. године (CBS, 2016). Том приликом потписан је споразум о билатералној сарадњи на *Big Data* са статистиком Кореје.

У **Републици Србији** не постоји стратешки оквир за развој и примену *Big Data*. Од стратешких докумената који тангирају ову област може се издвојити Стратегија развоја електронске управе у Републици Србији за период од 2015. до 2018. и Акциони план за спровођење иницијативе Партнерство за отворену управу у Републици Србији за период 2018-2020 - Акциони план (Интернет 3). Ипак, у овим документима *Big Data* се не помињу, већ само отворени подаци (*open data*). У акционом плану дефинисане су обавезе које представљају надградњу на већ утврђене активности у оквиру спровођења иницијативе отварања података у Србији кроз пројекат „Отворени подаци – отворене могућности" који спроводи Канцеларија за информационе технологије и електронску управу (у даљем тексту: ИТЕ), у сарадњи са Програмом Уједињених нација за развој (UNDP) који подржава развој ове иницијативе у Србији (Акциони план, 2018).

5.2.2 Приступ подацима и партнерства

Очигледно је да употреба података из *Big Data* извора у функцији производње службене статистике захтева суочавање са бројним изазовима. Пре свега, поставља се питање расположивости података, који су у власништву приватних компанија (Varian, 2018). Подаци у власништву јавног сектора су релативно лако доступни, захваљујући позитивној правној регулативи, која уређује стање у размени административних података (најчешће кроз закон о статистици). Да би се

обезбедила употреба података из *Big Data* извора који се налазе у приватном сектору, неопходно је направити партнерство са њиховим власницима, било да се подаци купују или добијају по некој другој основи. За то је неопходно да постоји стратешки оквир на нивоу државе, као што је приказано у претходном одељку, а који ће створити правни основ за формирање приватно – академског - јавног партнерства.

Један од могућих приступа је да се аналитика пренесе на власника података. Статистика Новог Зеланда је изразила интересовање да то уради са скенираним подацима, с обзиром да власник података има неопходну рачунарску и софтверску инфраструктуру. *Cost/benefit* анализа је указала да је то бржи и јефтинији начин за обраду и анализу.

Нема сумње да је власништво над подацима који су увезаним аналитичким алгоритмима добило велики значај за будућност многих, ако не и свих, компанија које послују у приватном сектору. Доказ за то су највредније компаније на свету (*Apple, Alphabet Inc. (Google), Facebook, Microsoft, Netflix, …*) које свој успех на тржишту у многоме могу да захвале огромним капацитетом за прикупљање, организовање, контролу и комерцијализацију података, уз све контроверзе које их при томе прате, пре свега везано за њихов монополски положај и злоупотребу прикупљених података (Coy, 2017; Koulopoulos, 2018).

Да би се јавно – приватна партнерстава успешно реализовала, неопходан је заједнички напор методолога (из јавног сектора - статистике) и ИТ стручњака из компаније која располаже подацима и одговарајућом технологијом (из приватног сектора) да би се развиле технике за смањење обима и сложености података уз очување квалитета добијених резултата (Tam & Clarke, 2015).

Квалитет података

Имајући у виду кључну улогу високо квалитетних података у доношењу одлука на државном нивоу али и глобално, као и неопходност интегритета званичних статистика, УН је донео Фундаменталне принципе на којима почива званична статистика (UNSTATS, 2015). Такође, квалитет података у оквиру ЕСС почива на

Коду праксе (*Code of Practice – CoP*) који је у потпуности у складу са Фундаменталним принципима УН-а (Eurostat, 2017a). Управљање квалитетом података је у фокусу званичних статистика и оно прожима све делове процеса статистичког истраживања.

Кад је реч о европској званичној статистици, велика пажња посвећује се тзв. мета подацима, односно „подацима о подацима", који служе бољем разумевању податка, начину на који је он добијен, прецизности мерења, репрезентативности и сл. (Metadata, 2018). Дакле, сваки податак који се сматра податком званичне статистике мора да испуни следеће критеријуме (Eurostat, 2015):

- Релевантност,
- Непристрасност,
- Доступност,
- Независност,
- Транспарентност,
- Поверљивост,
- Упоредивост.

Република Србија је у процесу приступа Европској унији прихватила хармонизацију статистике са ЕСС те принципе на којима овај систем почива. ЕСС се базира на основних 15 принципа, описаним у Кодексу праксе европске статистике, који се тичу институционалног оквира, процеса статистичке производње и резултата статистичке анализе (Eurostat, 2017a):

- Принцип 1: Професионална независност
- Принцип 2: Овлашћење за прикупљање података
- Принцип 3: Адекватност ресурса
- Принцип 4: Посвећеност квалитету
- Принцип 5: Поверљивост статистичких података
- Принцип 6: Непристрасност и објективност
- Принцип 7: Ваљана методологија
- Принцип 8: Одговарајуће статистичке процедуре
- Принцип 9: Умерено оптерећење давалаца података

- Принцип 10: Рационалност трошкова
- Принцип 11: Релевантност
- Принцип 12: Тачност и поузданост
- Принцип 13: Правовременост и поштовање најављених рокова објављивања
- Принцип 14: Кохерентност и упоредивост
- Принцип 15: Доступност и разумљивост

Поштовање ових принципа мери се специфичним индикаторима, који служе за процену усклађености националних статистичких система са *CoP*-ом. Како би се обезбедили квалитетни подаци, на нивоу ЕСС примењује се јединствен модел статистичког пословног процеса под називом *Generic Statistical Business Process Model (GSBPM)* (GSBPM, 2013).

У Републичком заводу за статистику (Завод) је развијен Општи модел статистичког производног процеса који се примењује као стандардни методолошки поступак код нових истраживања (RZS, 2012). Овај модел се ослања на међународне стандарде, препоруке и најбољу праксу коју предлажу кључни актери за постављање методолошких стандарда у области званичне статистике: UNECE, Eurostat и OECD.

На крају, напоменимо, да би се ма који податак који се добија коришћењем *Big Data* извора користио у званичној статистици (као ексклузивни, алтернативни или допунски извор), мора, макар у некој мери, да задовољава наведене принципе (1-15). Осим тога, тако добијени подаци се могу сматрати званичним статистичким подацима тек ако су на располагању и одговарајући мета подаци који их прате и дефинишу.

Питање приватности

Имајући у виду саму природу *Big Data,* јасно је да је приватност оно што је најосетљивије питање које се поставља приликом њиховог коришћења. Не сме се заборавити чињеница да, суштински, нису компаније те које производе *Big Data* податке. Произвођачи су сви (појединци, предузећа) свесни тога или не, који својим свакодневним активностима, у свом дому или ван њега, употребом модерних технологија, готово сваког тренутка производе податке. Готово свака трансакција

се бележи, посета интернету, разговор мобилним телефоном, плаћање картицом, дневна рута кретања, вожња аутопутем, и сл. оставља негде податке који се складиште у (великим) базама података. Такође, у већини случајева, опет, свесно или не, грађани и компаније дају пристанак да се ти њихови подаци користе од стране компаније која им пружају одређену услугу.

Дакле, доступност готово свим подацима, па и личним, је неминовност. Питање које је суштинско јесте на који ће начин компаније које располажу личним подацима чувати интегритет и приватност корисника. Губитак поверења од стране корисника и злоупотреба индивидуалних података доводи до губљења клијената те и до пада профита или чак краха компаније. Због тога су компаније, које су и најчешћи произвођачи *Big Data* у великој мери заинтересовани да приватност њихових корисника остане сачувана. Са друге стране, национална законодавства прописују и законе о заштити личности чиме се, и законодавно, обезбеђује приватност података. Међутим, имајући у виду мисију званичне статистике, целокупна друштвена заједница може имати користи од коришћења *Big Data* у званичне сврхе, уз приоритетну заштиту индивидуалних података. Сваки национални Закон о званичној статистици у себи садржи клаузулу о заштити података и статистичари су у обавези да поступају у складу са овом клаузулом. Коришћењем агрегираних података у великој мери се превазилази овај проблем јер се на тај начин приказују подаци за одређени скуп лица.

На нивоу ЕУ ову област уређује Уредба (ЕУ) 2016/679 Европског парламента и Савета из априла 2016. (Уредба) о заштити физичких лица у односу на обраду података о личности и о слободном кретању таквих података (*General Data Protection Regulation*), која се од 25.05.2018. примењује и у Републици Србији (GDPR, 2016).

Приликом израде метаподатака који су у вези са индикаторима израчунатим коришћењем *Big Data,* посебну пажњу треба посветити документовању начина на који је обезбеђена приватност коришћених података, како не би дошло до губитка поверења, како у статистичку институцију која производи те индикаторе, тако и у компанију која обезбеђује податке, уколико је реч о подацима који се добијају од стране трећих лица (Metadata, 2018).

Питање коришћења индивидуалних података и њихове заштите, произведених од стране различитих произвођача званичне статистике, није новијег датума, нити се појављује од настанка *Big Data* концепта. О томе се размишљало и приликом коришћења административних података, који су данас у великој мери коришћени (и све више се користе) од стране званичне статистике, нарочито када је реч о подацима пореске администрације (Nordbotten, 2008). Досадашња искуства и модели који су у примени у системима званичне статистике (неки од њих су приказани у овој дисертацији) показали су се као веома операбилни за размену података уз висок степен заштите.

Размена података у Републици Србији

У Републици Србији размена структурираних података између различитих органа путем дефинисаних канала врши се на основу прецизно дефинисаних протокола пријављивања на систем. Национални оквир интероперабилности је усвојен на седници Владе Републике Србије 2014. године. Након тога је објављена Листа стандарда интероперабилности која ће омогућити лакшу комуникацију међу системима (Стратегија, 2015).

2017. године завршен је национални Портал отворених података, као централно место на којем су обједињени подаци државне управе и локалне самоуправе. За сада је 27 институција отворило 104 скупа података са укупно 353 датотека, што је довело до великог помака Србије у овој области (мерено индексом отворених података Србија је на првом месту у региону и 41. на листи чланица УН). Ови подаци су стављени на располагање грађанима, приватном и невладином сектору (Интернет 3).

Када је реч о *Big Data* изворима који су у власништву приватног сектора (као што је већина извора наведених у табели 3.1), њихов коришћење је веома отежано, услед одсуства приватно – академског - јавног партнерства у области *Big Data*, за који у Републици Србији не постоји јасно дефинисана стратегија, као и пратећа правна регулатива.

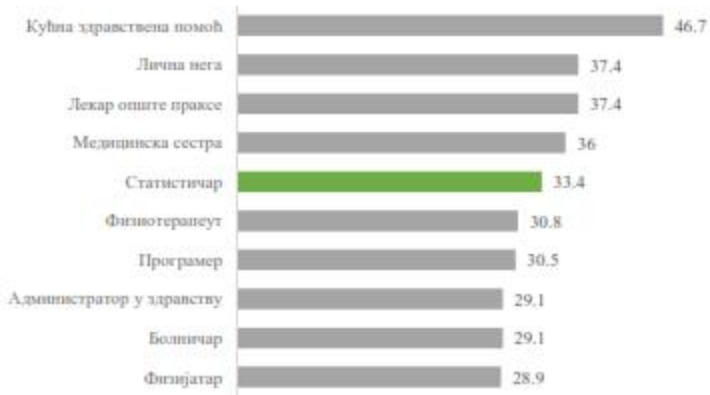
5.2.3 Кадрови

Према *TDWI* (*The Data Warehousing Institute*) истраживању, спроведеном 2011. године, на питање које су потенцијалне баријере у имплементацији *Big Data* технологије, као разлог број један наведен је недостатак вештина и особља потребног за *Big Data* аналитику. Наведено је често споменуто у бројним чланцима као уистину разлог број један за неуспех или уопште не започињање имплементације *Big Data* технологије (Russom, 2011).

С обзиром да се ради о релативно новој технологији тј. технологијама, велики је и недостатак кадра који је оспособљен за рад са *Big Data* те су из тог разлога компаније натеране на импровизацију, на преквалификацију запослених, додатно образовање што доводи до додатних трошкова у циљу евентуалне добити у будућности. Иако се ради о *open source* пројектима и даље постоје трошкови који могу бити доста велики. Говори се и о трошковима особља, њихове едукације, те опортунитетном трошку, а не само о трошку *hardwear-a*.

Један од водећих економиста компаније *Google* Hal Varian, још 2011. године је изјавио да сматра да ће у наредним годинама посао статистичара, односно стручњака за податке (*data scientist*) бити један од најтраженијих и најплаћенијих у будућности (Интернет 4).

Ова прогноза се показала тачном, с обзиром на податке са сајта *Indeed*, који указују да новоотворена радна места стручњака за податке (статистичара) у САД-у забележила раст од 75%, од јануара 2015. до јануара 2018. године (Интернет 5). Такође, Биро за статистику рада (*The Bureau of Labor Statistics*) очекује да ће статистичар бити најбрже растуће занимање, а која се не односе на здравље до 2026. године (Слика 5.2) (BLS, 2018).



Слика 5.2: Десет најбрже растућих професија од 2016 до 2026. године у САД (BLS, 2018)

Слична је ситуација и са статистичарима који су запослени у сектору званичне статистике. Пред њима је изазов у овладавању нових знања и вештина које ће их квалификовати за звање стручњака за податке. Нови профил стручњака за податке, комбинује у једно: статистичара и ИКТ специјалисту (Hackl, 2016). Од стручњака за податке се очекује да се разумеју у програмирање, развој апликација, управљање мрежама (*cloud* технологијама), управљање и анализу *Big Data*, између осталих вештина (OECD, 2016).

Република Србија

Иницијативе Дигитална Србија, коју су основале неке од водећих компаније у Србији (Слика 5.3) у свом Дигиталном манифесту наводе резултате које желе да постигну у области образовања која тангира нове технологије (IDS, 2017).

Слика 5.3: Оснивачи Иницијативе Дигитална Србија (IDS, 2017)

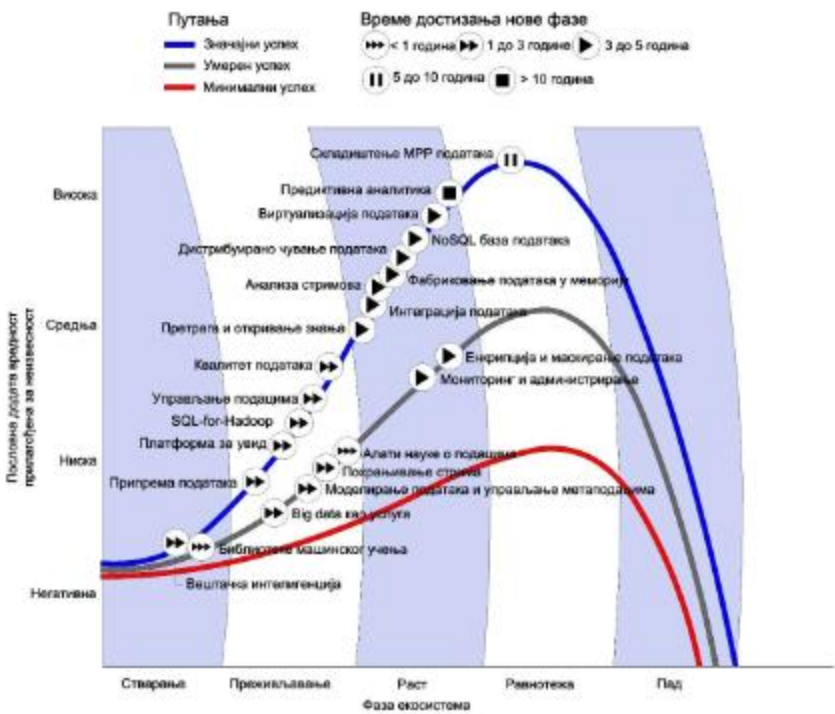
Према овој иницијативи, стратешки циљ је да до 2025. године број инжењера, професионалаца са техничких факултета и дигиталних експерата у Србији треба да буде десетоструко већи од данашњег.

Да би се то остварило, неопходно је образовни систем и привреда, заједно са ИТ организацијама, направе један обједињени екосистем. Реч је о некој врсти партнерства у којој је задатак образовног система да преноси релевантне технолошке вештине и знања од раног доба и да препозна талентоване студенте и омогући им усавршавање у ИТ области. Универзитети треба да прерасту у центре за иновације и развој, повезане са ИТ индустријом. У закључку се наводи да целокупно друштво треба да препозна информационо-комуникационе технологије као релевантну и атрактивну област за учење и рад (IDS, 2017).

5.2.4 Технолошка инфраструктура

Технологија *Big Data* је настала из обраде екстремно великог обима података са интернета и постепено се примењивала на све већи број пословних домена у

последњој деценији. Развој *open source* технологије отвореног кода и извора убрзано је сазрео до тачке у којој *Big Data* технологије, интегрисане са традиционалним технологијама обраде података - обезбеђују моћни скуп технолошких алата. Приказ развоја за 22 технологије током читавог животног циклуса података са слике 5.4 указује на место и потенцијал *Big Data* технологије која се налази у успону и приближава фази раста (Forrester, 2016). Такође, са слике се може закључити да се „победничке технологије" базирају на извештавању у реалном времену, предиктивним алатима и интегрисаним решењима, диктираним од стране великих клијената, што све карактерише *Big Data*.



Слика 5.4: Развој технологија током читавог животног циклуса података (Forrester, 2016).

Већина етаблираних технологија сада укључује технологију *Big Data* као део свог портфеља производа (IBM, Microsoft, Oracle, SAP, SAS,...). Развијене су различите

Big Data платформе које омогућавају извршавање различитих задатака за потребе различитих корисника на једној јединственој платформи коришћењем софтвера за аналитику.

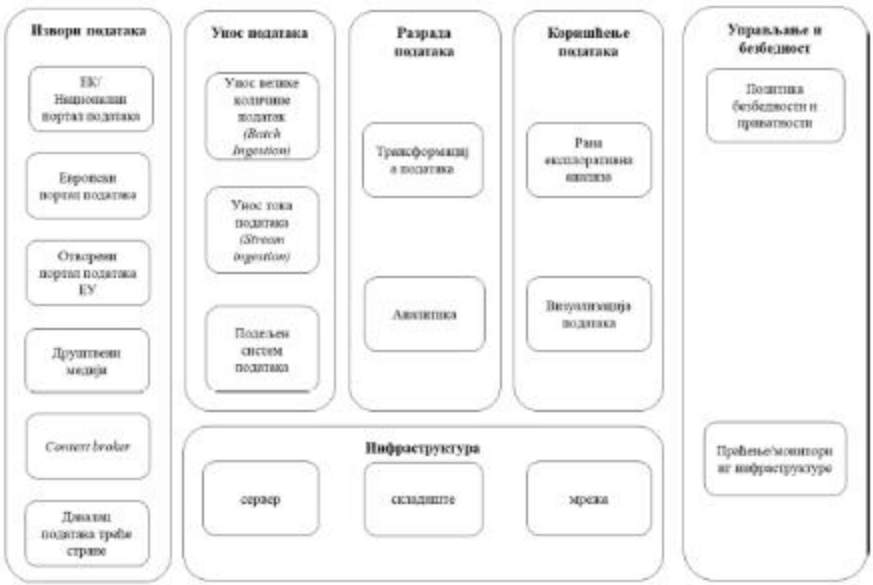
Од *Big Data* платформе и *Big Data* софтвера за аналитику очекује се да има одговарајуће функционалности које се тичу управљања подацима, *Hadoop* система, машинско учење, интеграцију података, безбедности и сл. (Vukmirović, 2017). Према овим функционалностима, најбоље рангиране *Big Data* платформе и *Big Data* софтвери за аналитику приказани су на слици 5.5 (Интернет 6).

Amazon Web Service, Google BigQuery, Arcadia Data, Microsoft Azure, Informatica PowerCenter Big Data Edition, GoodData, Actian Analytics Platform, Google Bigdata, Wavefront, IBM Big Dat, Datameer, Attivo Active Intelligence Engine, DataTorrent, Cloudera Enterprise Bigdata, Opera Solutions Signal Hubs, FICO Big Data Analyzer, Palantir Bigdata, Oracle Bigdata Analytics, Qubole, Sincsort, MapR Converged Data Platform, Hortonworks Data Platform, Amdocs Insight, Splunk Bigdata Analytics, Celebrus Technologies, VMware, HPCC Systems Bigdata, Pentaho Big Data Analytics, Pachyderm, BlueTalon, Flytxt, MongoDB, BigObject, Rabikloud, SAP Bigdata Analytics, Next Pathway, CSC Big Data Platform, Kognito Analytical Platform, 1010data, GE Industrial Internet, DataStax Bigdata, Mu Sigma Bigdata, MicroStrategy Bigdata, Opera Solutions Bigdata

Слика 5.5: Најбоље рангиране *Big Data* платформе и *Big Data* софтвери за аналитику (Интернет 6).

Паралелно са готовим решењима која се нуде на тржишту, инфраструктура и пратећи алати за *Big Data* се развијају и као сопствена решења. Када је реч о званичној статистици, обрада велике количине података захтева нове типове представљања података (семантичке податке, базу података графова), нове технике закључивања (аналитичке технике засноване на вештачкој интелигенцији (ВИ) увезане са робусном статистичком анализом), визуелизацију, аналитичке језике (као што су R и SAS), и коришћење хардвера високих перформанси (Tam & Clarke, 2015).

На нивоу ЕУ направљено је јединствена платформа: *CEF Big Data Test Infrastructure* (BDTI), као комплетно решење које егзистира у виртуелном окружењу и омогућава европским организацијама (укључујући и статистичке заводе) да експериментишу с технологијама *Big Data* (Слика 5.6). Осим саме инфраструктуре, BDTI регистрованим корисницима ставља на располагање одређене скупове података и аналитичка решења (BDTI, 2018).



Слика 5.6: Архитектонске компоненте (BDTI, 2018)

BDTI је резултат једне од активности у оквиру ISA програма – *Big Data for Public Administrations,* који подржава развој дигиталних решења са циљем да омогући јавним администрацијама, предузећима и грађанима у Европи да имају користи од интероперабилних, прекограничних и међусекторских јавних услуга (BDTI, 2018).

6 Основне поставке *Big Data* модела

6.1 Фазе животног циклуса *Big Data* пројекта

Једна од основних замерки које се односе на *Big Data* је да углавном представљају артефакте људских интеракција, односно да нису производ системског прикупљања података (Ospina, 2018). Са методолошког становишта, овако прикупљене информације нису добијене на основу података који су прикупљани у статистичке сврхе. Аналогија се може тражити са употребом административних извора – регистара, која је већ констатована у претходном поглављу. Из тог разлога, полазну основу предложеног решења представљају модели процеса статистичког истраживања *Generic Statistical Business Process Model* (GSBPM), који се примењује у статистичким заводима широм света (GSBPM, 2013). У наставку ове дисертације као референтни GSBPM, користи се „Модел процеса статистичког истраживања", који се користи у Републичком заводу за статистику Србије - РЗС (RZS, 2012). За потребе истраживања наведени модел назива се традиционалним моделом.

Овај традиционални модел је модификован и прилагођен за употребу *Big Data* технологија. Добијено решење представља *Big Data* модел процеса статистичког истраживања (*Big Data* модел). У истраживању се имплементира пројектни приступ истраживању и у наставку дисертације приказане су фазе животног циклуса пројекта базираног на *Big Data* моделу – *Big Data* пројекат.

Предложено решење је тестирано кроз студију случаја која се односи на утврђивање потенцијала *Big Data* концепта у процесу дефинисања и праћења индикатора одрживог развоја. При томе, *Big Data* се третира као савремени концепт у званичној статистици, базиран на алтернативним изворима података, који омогућава производњу нових статистичких производа, у краћем временском периоду, уз редукцију трошкова и смањење оптерећења корисника.

Фазе традиционалног модела процеса статистичког истраживања су (RZS, 2012):

1. Утврђивање потреба,
2. Пројектовање,

3. Реализација и тестирање производног система,
4. Прикупљање,
5. Обрада,
6. Анализа,
7. Дисеминација,
8. Архивирање и
9. Евалуација.

У циљу хармонизације статистичких процеса у домену званичне статистике, циљ је био да се што је могуће више, прате наведене фазе традиционалног модела, с тим што се евалуација нашла испред фаза Дисеминације и Архивирања, с обзиром на значај фазе евалуације у процесу статистичког истраживања. На тај начин добијене су фазе животног циклуса *Big Data* пројекта базираног на *Big Data* моделу (Табела 6.1).

Табела 6.1: Фазе животног циклуса *Big Data* пројекта

Фаза	**Назив фазе**
1	Утврђивање потреба
2.	Пројектовање
3.	Реализација и тестирање производног система
4.	Прикупљање података
5.	Обрада података
6.	Анализа резултата
7.	Евалуација добијених резултата
8.	Дисеминација
9.	Архивирање

У наставку поглавља дат је преглед наведених фаза, без детаљног описа свих појединачних подфаза животног циклуса *Big Data* пројекта. Акценат је стављен на оне подфазе *Big Data* модела које се значајно разликују од традиционалног модела истраживања.

6.2 Фаза 1: Утврђивање потреба

Прва фаза традиционалног модела процеса статистичког истраживања односи се на Утврђивање потреба.

Подфазе ове фазе су:

1.1. Утврђивање потреба за информацијама

1.2. Консултације и потврђивање потреба

1.3. Дефинисање потребних резултата

1.4. Идентификовање концепата (статистичких стандарда)

1.5. Провера расположивости података

1.6. Израда предлога пројекта

Већ у овој, уводној фази, долази до концептуалних разлика између традиционалног и *Big Data* приступа.

У подфази 1.1. Утврђивање потреба за информацијама, редовна периодика објављивања резултата истраживања у статистикама је месечна, квартална, годишња и вишегодишња. Вероватно највећа промена/прогрес за кориснике резултата званичне статистике настаје у дисеминацији резултата, које се огледа у потенцијалу скраћивању времена за производњу и публиковање резултата до нивоа реалног или готово реалног времена (V - *Velocity*). То је од нарочите важности за добијање брзих, прелиминарних (*flash*) оцена појединих индикатора у кратком временском периоду, што представља једну од појединачних хипотеза, које се односе на елементарне чиниоце предмета истраживања, а коју смо доказали у претходном тексту, на примеру статистике запослености (Hammer, et al., 2017).

Подфаза 1.2. Консултације и потврђивање потреба: нове могућности стварају нове потребе. Захваљујући великом броју нових извора података (V – Volume) који се могу увести у систем званичне статистике шири се потенцијални асортиман производа који могу да се ставе на располагање корисницима (V - *Variety*).

И остале подфазе из почетне фазе традиционалног модела процеса статистичког истраживања, као што су: дефинисање потребних резултата, идентификовање концепата (статистичких стандарда), провера расположивости података и израда предлога пројекта истраживања, захтевају одређене корекције у складу са предложеним моделом базираним на *Big Data* концептима.

Ово се посебно односи на проверу расположивости података и израду предлога пројекта истраживања.

Провера расположивости података

Ова подфаза је важна јер детерминише даљи поступак истраживања (Слика 6.1):

- Уколико постоје традиционални извори података који могу задовољити потребе корисника (по дефинисаним критеријумима за квалитет података и извора званичне статистике), наставак истраживања се изводи у складу са традиционалним моделом процеса статистичког истраживања.
- Уколико не постоје традиционални извори података а постоји потреба корисника (везано за брзину извештавања (V - *Velocity*) или увођење новог индикатора), наставак истраживања се имплементира у складу са *Big Data* моделом, поштујући пројектни приступ, а у складу са *Big Data* концептима и одговарајућим изворима података (V - *Variety*) који су дефинисани у трећем поглављу дисертације.

Слика 6.1: Фаза 1. Провера расположивости података у *Big Data* моделу

У истом истраживању могуће је комбиновати традиционалне и *Big Data* изворе, што је веома чест случај у домену званичне статистике. С обзиром да је традиционални модел детаљно описан у литератури (RZS, 2012), у наставку текста разматран је дефинисани *Big Data* модел.

6.3 Фаза 2: Пројектовање

Процена ризика извођења *Big Data* пројекта

Пре него што се пређе на детаљу разраду *Big Data* пројекта, са подфазама традиционалног пројектовања:

- пројектовање резултата,
- дефинисање променљивих,
- методи и инструменти прикупљања података,
- методологија узорка,
- методологија статистичке обраде и
- пројектовање производних система и токова рада,

неопходно је урадити процену ризика извођења *Big Data* пројекта у домену коришћења података о личности и нарушавања приватности. У ту сврху кренули смо од иницијалне контролне листе (*checklist*), развијене од стране *Global Pulse* иницијативе и прилагодили је потребама званичне статистике (UNDP, UN Global Pulse, 2016). Ова листа представља део свеобухватнијег пројекта процене, разумевања и минимизације ризика, како би се максимизирали позитивни ефекти *Big Data* пројекта. Овај алат не представља правни акт, и није заснован на националном законодавству, тако да смо га прилагодили захтевима Уредбе (ЕУ) 2016/679 Европског парламента и Савета из априла 2016. (Уредба) (GDPR, 2016).

У табели 6.2, у пољу „*Објашњење*" дате су основне дефиниције из Уредбе које представљају смернице везане за попуњавање контролне листе.

Разлог што се процена ризика ради после израде концепта истраживања је у томе што се процена ризика односи на сваку фазу животног циклуса *Big Data* пројекта. Препорука је да се одговарајућа питања која се постављају у контролној листи разматрају од стране експертског тима који чине: руководилац пројекта, доменски експерти - стручњаци за поједине теме везано за предмет истраживања, методолози, статистичари, информатичари, стручњаци за податке (*data scientists*) и стручњаци за заштиту података. Овај експертски тим делује проактивно, тако што осим подршке коју пружају при одговарању на питања из контролне листе, стоје на

располагању руководиоцу пројекта у евентуалном ублажавању потенцијалних ризика.

Напомена:

Постављена питања су са алтернативним одговором, следећих модалитета:

a. Да,
b. Не,
c. Не знам и
d. Није применљиво

За сваки одговор:

- „Није применљиво" - потребно је дати објашњење у коментарима.
- „Не знам" - треба аутоматски сматрати фактором ризика који захтева консултацију са одговарајућим стручњаком (доменским експертом), како би се добио финални одговор („Да", „Не" или „Није применљиво").

Коначна одлука се доноси на основу контролне листе, и доноси се тек онда када не постоји ни један одговор „Не знам".

Табела 6.2: Контрола листа за процену ризика извођења *Big Data* пројекта

Део 1:	**Типови података**	
1.1.	Да ли ће се у истраживању користити подаци о личности - које се односе на физичко лице чији је идентитет одређен или може да се одреди? a. Да b. Не c. Не знам d. Није применљиво	Коментар:
Објашњење:	*Подаци о личности - сви они подаци који се односе на неко одређено или одредиво физичко лице, на основу којих оно може бити идентификовано чиме се угрожава њихова приватност: нпр. Име и презиме, адреса становања, ЈМБГ, корисничко име, број телефона, адреса е-поште, статичка IP адреса, идентификатор уређаја, итд.*	
1.2.	Да ли ће се користити псеудонимизовани подаци који не идентификују појединца	Коментар:

	директно, али који би се могли користити за издвајање јединствене особе применом постојећих и лако доступних средстава и технологија? a. Да b. Не c. Не знам d. Није применљиво	
Објашњење:	*Подаци о личности који су псеудонимизовани, а који би могли да се припишу неком физичком лицу помоћу додатних информација морају да се сматрају информацијама о физичком лицу чији идентитет може да се одреди.* *„Псеудонимизација" је обрада података о личности на такав начин да подаци о личности више не могу да се повежу с конкретним лицем на које се подаци односе без коришћења додатних информација, под условом да се такве додатне информације чувају одвојено и да се на њих примењују техничке и организационе мере да би се обезбедило да подаци о личности не могу да се повежу с физичким лицем чији је идентитет одређен или се може одредити*	
1.3.	Да ли ће се користити осетљиви (сензитивни) подаци? a. Да b. Не c. Не знам d. Није применљиво	Коментар:
Објашњење:	*Осетљиви (сензитивни) подаци - подаци о личности који откривају:* - *расно или етничко порекло,* - *политичка мишљења,* - *верска или филозофска уверења,* - *чланство у синдикату.* *Обрада:* - *генетичких података,* - *података о здравственом стању,* - *података о сексуалном животу,* - *података о кривичним осудама и кривичним делима или повезане мере безбедности;* *Анализа:* - *Ако се процењују лични аспекти, а посебно анализа или предвиђање аспеката у вези са учинком на послу, економским стањем, здрављем, личним склоностима или*	

	интересима, поузданошћу или понашањем, локацијом или кретањем, како би се направили или користили лични профили; - *Ако се обрађују подаци о личности угрожених физичких лица, а посебно деце; или ако обрада обухвата велику количину података о личности и утиче на велики број лица на које се подаци односе.*	
Контролна тачка	Ако је одговор **„Да"** на најмање једно од наведених питања (од 1.1. до 1.3), повећава се ризик извођења *Big Data* пројекта	
Део 2	**Прикупљање података**	
2.1.	Да ли су прикупљени подаци из система званичне статистике??	Акција:
	a. Да	*Део 3.*
	b. Не	*Одељак 2.2.*
	c. Не знам	*Консултације*
2.2.	Како су подаци прикупљени?	Акција:
	a. Директно од испитаника (анкета, интервју и сл.)	*Одељак 2.2.*
	b. На други начин (интернет, друштвени медији, провајдери мобилне телефоније....)	*Одељак 2.3.*
	c. Не знам	*Консултације*
2.3.	Постоји ли правни основ за прикупљање и обраду прикупљених података? a. Да b. Не c. Не знам d. Није применљиво	Коментар:
2.4.	Да ли добављач података (*data provider*) поседује правни основ за прикупљање и обраду прикупљених података? a. Да b. Не c. Не знам d. Није применљиво	Коментар:
Контролна тачка	Ако је одговор **„Не"** на најмање једно од наведених питања (2.1, 2.3 и 2.4), повећава се ризик извођења *Big Data* пројекта	
Део 3.	**Коришћење података**	
3.1.	*Спецификација циљева:* Да ли је дефинисана сврха за коју ће се користити подаци?	Акција:
	a. Да	*Одељак 3.2*

	b. Не	*Врати се на фазу 1.*
	c. Не знам	*Консултације*
3.2.	*Компатибилност циљева*: Да ли је сврха за коју ће се користити подаци компатибилна са сврхом за коју су они прикупљени?	Акција:
	a. Да	*Одељак 3.3*
	b. Не	*Врати се на подфазу 1.1*
	c. Не знам	*Консултације*
3.3.	*Минимизација потребних података:* Да ли су сви подаци који ће прикупити истраживањем заиста неопходни?	Акција:
	a. Да	*Одељак 3.4*
	b. Не	*Врати се на фазу 1.*
	c. Не знам	*Консултације*
Објашњење:	*Приступ подацима, њихова обрада и употреба треба да буде на минималном потребном нивоу, односно да буду неопходни, адекватни и релевантни у односу на циљеве истраживања. Податке треба чувати само онолико колико је то потребно, у складу за правном регулативом.*	
3.4.	*Регулатива и законска усклађеност:* Да ли је употреба података у складу са (а) важећим законима и (б) условима под којима су добијени подаци?	Акција:
	a. Да	*Одељак 3.5*
	b. Не	*Ризик*
	c. Не знам	*Консултације*
	d. Није применљиво	*Коментар*
Објашњење:	*Неопходно је проверити да ли су добијене све регулаторне и друге потребне дозволе за наставак пројекта. (На пример: коришћење података везано за кориснике мобилне телефоније могу бити ограничени у складу са законима о телекомуникацијама. Требало би проверити постојеће споразуме, лиценце, услове коришћења на платформама друштвених медија или условима сагласности коришћења*	
3.5.	*Квалитет података:* Да ли су сви подаци који ће прикупити истраживањем у складу са основним концептима квалитета?	Акција:
	a. Да	*Одељак 3.6*

	b. Не	*Ризик*
	c. Не знам	*Консултације*
3.6.	*Сигурност података:* Да ли постоје процедуре и прописи којима се обезбеђује поверљивост података и интерна заштита података	Акција:
	a. Да	*Део 4.*
	b. Не	*Ризик*
	c. Не знам	*Консултације*
	d. Није применљиво	*Коментар*
Део 4. Процес комуникације		
4.1.	*Транспарентност:* Да ли постоји план комуникације везан за коришћење података (јавно или интерно - са другим одговарајућим заинтересованим странама)?	Акција:
	a. Да	*Одељак 4.1*
	b. Не	*Ризик*
	c. Не знам	*Консултације*
	d. Није применљиво	*Коментар*
4.2.	*Ниво транспарентности:* Да ли постоје неки ризици и евентуалне штете повезане са објављивањем прикупљених података или резултирајућим извештајима?	Акција:
	a. Да	*Ризик*
	b. Не	*Део 5.*
	c. Не знам	*Консултације*
	d. Није применљиво	*Коментар*
Део 5. Треће стране – Партнери у истраживању (*outsourcing* компаније)		
5.1.	Да ли партнери у истраживању (ако их има - за потребе чувања података (*cloud*), *hosting-a*, пословне аналитике, и сл.) имају капацитет да поштују највише стандарде и основне принципе за заштиту података како је наведено у овој контролној листи?	Акција:
	a. Да	*Део 6.*
	b. Не	*Ризик*
	c. Не знам	*Консултације*
	d. Није применљиво	*Коментар*
Део 6. Ризици и потенцијалне штете		
6.1.	*Ризици*: Да ли употреба прикупљених података представља ризик од штете неком појединцу или групама појединаца, односно да ли они могу	Акција:

	бити директно идентификовани, видљиви или познати?	
	a. Да	*Ризик*
	b. Не	*Одељак 6.2.*
	c. Не знам	*Консултације*
	d. Није применљиво	*Коментар*
6.2.	*Потенцијалне штете:* Да ли употреба прикупљених података угрожава неког појединца или групу појединаца, односно доводи од штетних последица по њих?	Акција:
	a. Да	*Ризик*
	b. Не	*Део 7.*
	c. Не знам	*Консултације*
	d. Није применљиво	*Коментар*
Објашњење:	*Ризике треба проценити одвојено од штета. Постојање ризика не значи да ће нужно настати штета. Приликом одговора на ово питање, важно је усредсредити се на могуће ризике. Треба имати на уму да типична анализа података резултира генерисањем новог скупа података. Такав исход треба узети у обзир и као ризик и мора се посебно проценити за ризике, штете и користи пре даље употребе / објављивања резултата истраживања. Такође, пристрасност у статистичком смислу представља скривени ризик који се може произвести као резултат коришћења података.*	

Део 7. Коначна одлука и образложење одлуке

Коначна процена - На основу одговора у одељцима 1-7, проценити да ли су ризици и штетне последице несразмерно високи у поређењу са очекиваним позитивним резултатима *Big Data* пројекта.

- Питања 1.1 - 1.3; 4.2; 6.1-6.2 постоји одговор „Да" - ризик је присутан.
- Питања 2.1 - 2.4; 3.4-3.6; 4.1; 5.1 постоји одговор „Не" - ризик је присутан.

- Одговор „Не знам" на било који од питања, сматраће се као „фактор ризика". Процена ризика се врши искључиво уколико се на сва питања одговори модалитетима: „Да", „Не" или „Није применљиво".
- Сваки појединачан одговор „Није применљиво", захтева објашњење у пољу Коментари

- Уколико је констатован потенцијални ризик, неопходно је да се:
  - Процени вероватноћа наступања **потенцијалног ризика,** као и вероватноћа, величина и озбиљност **потенцијалних штета**
  - Предложе корективне акције за смањивање потенцијалних штетних ефеката.

- Уколико је констатовано да су неки од ризика или штете нејасни или високи, неопходно је да се изврши свеобухватнија процена ризика уз укључивање стручњака за сигурност података и приватност.
- Ако је констатовано да је вероватноћа ризика и штете веома ниска (или не постоји) у односу на вероватноћу позитивног утицаја, треба наставити са *Big Data* пројектом. Увек имати на уму да треба предузети што више мера за ублажавање ризика (чак и ако је констатован ризик на ниском нивоу) (Слика 6.2).

Слика 6.2: Процена ризика извођења *Big Data* пројекта

Према томе, спровођење *Big Data* пројекта подразумева извођење следећих подфаза фазе 2. Пројектовање:

2.1. Процена ризика извођења *Big Data* пројекта
2.2. Пројектовање резултата
2.3. Дефинисање променљивих
2.4. Избор метода и инструмената за прикупљање података
2.5. Методологија узорка
2.6. Методологија статистичке обраде
2.7. Пројектовање производних система и токова рада

Док се подфазе 2.2. и 2.3. есенцијално не мењају код имплементације *Big Data* пројекта, подфаза 2.4. Избор метода и инструмената за прикупљање података је

круцијално повезана са основним концептом *Big Data* приступа који карактерише мноштво потенцијалних извора расположивих података (*V-Variety*), описаним и категоризованим у претходном тексту. Такође, од ове подфазе директно зависе подфазе 2.6. Методологија статистичке обраде, односно 2.7. Пројектовање производних система и токова рада.

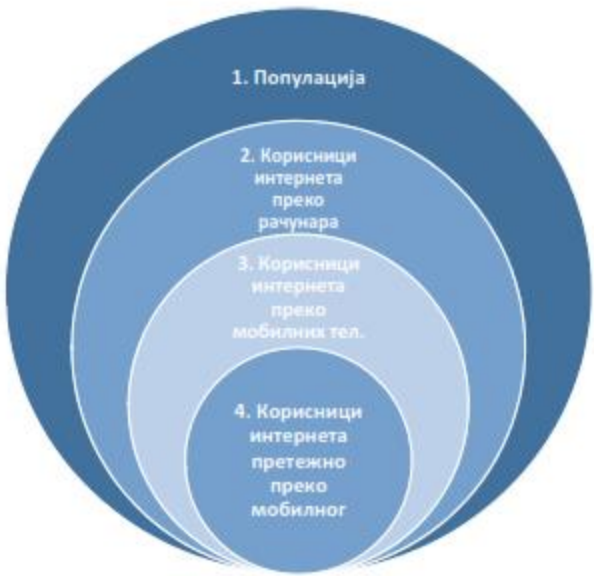
Избор метода и инструмената прикупљања података је везан за предмет и циљеве истраживања, наведеним у оквиру фазе 1. Утврђивање потреба али и за потенцијал *Big Data* извора података. Након њихове идентификације и евалуације ризика, врши се коначан избор метода и инструмената за њихово прикупљање. При томе се мора водити рачуна о начину на који ће се вршити (статистичка) обрада тако прикупљених података.

Када је реч о подфази 2.5. Методологија узорка мора се имати у виду да сетови података добијени из *Big Data* извора нису нужно (случајни) узорци циљне популације, и углавном то није случај. Са друге стране, анализа података, па и статистичко закључивање код традиционалног модела истраживања базира се на непристрасним оценама параметара популације добијеним на основу случајних узорак (Kish, 1965; Rubin, 1976; Sarndal et al., 1977, Chaudhuri & Stenger, 2005). Евентуално се користе статистички модели базирани на неслучајним узорцима који не поседују пожељне карактеристике репрезентативних узорака (Puza & O'Neill, 2006, Lavrakas, 2008, Alvi, 2016).

Типичан пример *Big Data* извора представља употреба друштвених медија (као што је *Twitter*) који представљају значајан извор података за традиционална истраживања, посебно јавног мњења. С обзиром да постоји мало информација о корисницима појединих друштвених медија (не постоји оквир узорка), немогуће је одредити да ли су кориснички профили репрезентативни за општу/циљну популацију. Чак је вероватно да ће неке подгрупе становништва бити недовољно заступљене у било којем потенцијалном узорку корисника друштвених медија, захваљујући стопи диференцијалног усвајања нових технологија - *Digital divide* (Tam & Clarke, 2015).

Генерално, доношење закључака о општој популацији само на основу групе испитаника одабране из популације без јасно дефинисаног оквира узорка, без одговарајућег моделирања и прилагођавања, подлежу пристрасности (Smith, 1983).

На слици 6.3 приказана је класификација популације интернет корисника. Она је по својој природи хијерархијска и подељена је у 4 ниво, почев од најмање субпопулације корисника интернета претежно преко мобилне телефоније, који су уједно и припадници популације корисника мобилног интернета, преко корисника рачунара, до опште популације (Antoun, 2015).



Слика 6.3: Хијерархијска класификација популације интернет корисника (Antoun, 2015)

Од посебне је важности да се у извештајима који се доносе на бази *Big Data* модела јасно нагласи да ли се добијени резултати могу сматрати репрезентативнима у статистичком смислу (да ли се базирају на случајним узорцима). Само тада се могу израчунати узорачке грешке и урадити извештај о квалитету истраживања (*quality report*).

Уколико узорак није репрезентативан у статистичком смислу, тада ћемо напустити концепт случајног узорка и говорити о сетовима података и неслучајним узорцима.

У складу са *Big Data* концептима, очекивано је да се структурирани подаци добијени из *Big Data* извора могу разматрати као потенцијални оквир случајних узорака, за разлику од полуструктурираних и неструктурираних података из којих се могу бирати само неслучајни узорци.

С обзиром на уобичајену праксу, систем званичне статистике преферира концепт случајног узорка, тако да се алгоритам модела наставља питањем да ли је могуће изабрати случајни узорак из *Big Data* извора. Уколико је одговор „да", примењује се позната методологија избора узорка. Уколико то није могуће ова подфаза (2.4) се прескаче (Слика 6.4).

Слика 6.4: Фаза 2. Провера расположивости случајног узорка у *Big Data* моделу

Пројектовање методологије (статистичке) обраде и анализе података односи се на развој нових и/или прилагођавање постојећих статистичких, односно аналитичких метода (методе *Big Data* аналитике), која се примењују у Фази 5. Обрада и 6. Анализа података. Методе *Big Data* аналитике се користе у случајевима када није могуће применити стандардизоване статистичке процедуре, за обраду

полуструктурираних и неструктурираних података (Minellin & Chambers, 2013). Аналитичке методе првенствено обухватају следеће технике (Gandomi, A. H, Murtaza (2015):

- Аналитика текстуалних података (*text mining*),
- Анализа аудио података,
- Анализа видео података,
- Анализа друштвених медија и
- Предиктивна аналитика.

Детаљнији опис ових техника следи у наредним фазама *Big Data* модела.

6.4 Фаза 3: Реализација и тестирање производног система

Ова фаза подељена је на шест подфаза и у њој се врши имплементација пројектованих решења из претходне фазе, тако што се прате предложене подфазе традиционалног модела:

3.1. Израда инструмената за прикупљање података
3.2. Израда или проширење компоненти процеса
3.3. Интегрисање производног система са осталим системима
3.4. Тестирање производног система
3.5. Тестирање статистичког процеса
3.6. Финализација производних система

Израда инструмената за прикупљање података зависи од избора извора за њихово прикупљање. Упитници и остали инструментариј везан за структуриране податке су добро дефинисани у традиционалним моделима истраживања, тако да се у наставку модела концентришемо на полуструктуриране и неструктуриране податке.

Приликом имплементације модела инфраструктуре треба имати у виду да се одређене софтверске апликације и сервиси имплементирају на *Cloud* инфраструктури. То значи и да се провајдери – компаније који изнајмљују простор на *Cloud-y*, морају тестирати кроз контролну листу за анализу ризика.

6.5 Фаза 4: Прикупљање података

Код традиционалних истраживања, прикупљање података се врши у четири подфазе:

4.1. Избор узорка,

4.2. Организовање прикупљања,

4.3. Спровођење прикупљања и

4.4. Унос података.

Као што је већ објашњено у фази 2, када је реч о *Big Data* истраживањима, није увек могуће располагати случајним узорком. У том случају, истраживање се спроводи и организује у зависности од циљева истраживања и *Big Data* извора који ће се користити.

У фази прикупљања података, неки од примера имплементације *Big Data* извора у званичној статици односе на следеће активности (Tam & Clarke, 2015):

- Дефинисање оквира узорка,
- Креирање/ажурирања пословног регистра,
- Потпуна супституција података из истраживања,
- Делимична супституција примарних података – редуковање броја питања у упитнику – смањење оптерећења испитаника,
- Делимична супституција примарних јединица узорка (за поједине субпопулације) – редуковање величине узорка.

Овако прикупљени подаци доприносе унапређењу наредних фаза истраживања. При томе, потребно је нагласити, да *Big Data* нису ништа мање склони пристрасности, ако нису и више, у односу на традиционалне статистичке податке, из разлога што су базирани на технологијама које се брзо мењају и што у многоме зависе од алата за њихову обраду и анализу (Ruths & Pfeffer, 2014).

Последња подфаза, унос података, се углавном не спроводи, обзиром да се подаци већ налазе у неком електронском формату.

6.6 Фаза 5: Обрада података

Токови рада на *Big Data* истраживању се пројектују кроз два главна процеса: Управљање подацима и аналитика (Слика 6.5).

Слика 6.5: Пројектовани токови *Big Data* процеса (Labrinidis & Jagadish, 2012)

Унутар *Big Data* процеса под обрадом *Big Data* сматраћемо:

- екстракцију информација,
- чишћење података,
- интеграцију података,
- агрегацију података и
- представљање података (визуелизација).

Према томе, обрада података представља интегрални процес који почиње са филтрирањем и чишћењем велике количине података који су на располагању истраживачима, с обзиром да већи део ових података није значајан за резултате истраживања. Основни задатак овог процеса је да се дефинише оптималан филтер, такав да се прикупе само корисне информације, при чему је евидентан ризик да у обраду уђу нерелевантни, односно да се не укључе релевантни подаци. При томе се акценат ставља на генерисања метаподатака како би се што боље описали подаци који се прикупљају. Преферирају се аутоматизовани системи за прикупљање метаподатака коју смањују трошкове рада експерата у дефинисању метаподатака (Lyko, et al. 2016).

Да би тако добијени подаци могли успешно да се анализирају на традиционалан начин, статистичким методама, неопходно да се изразе у структурираном облику погодном за анализу. Уколико то није могуће, примениће се *Big Data* аналитика.

Слика 6.6: Фаза 5. Провера структурираности података у *Big Data* моделу

С обзиром да *Big Data* аналитика захтева аутоматизацију процеса обраде и анализе, евидентна је разлика у структури података у односу на традиционалне (статистичке) методе обраде и анализе података. Код *Big Data* аналитике захтева се структура и семантика података која омогућава обраду на начин да их компјутер разуме (*computer understandable*) и „роботски" решава (*„robotically" resolvable*) (Labrinidis & Jagadish, 2012).

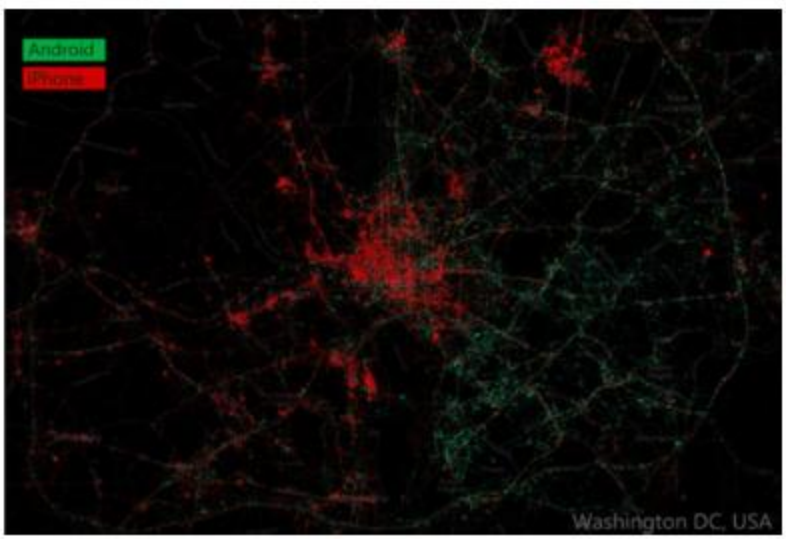
Овај услов захтева улагање пуно напора и средстава у интеграцију и агрегацију података, како би се смањила пристрасност и грешке при закључивању. Алати за интеграцију података састоје се од модула за рационализацију, усаглашавање, семантичку интерпретацију и реструктурирање података између различитих архитектонских приступа (Beyer, et al, 2018).

У званичној статистици, као и већини статистичких истраживања, у домену обраде података, *Big Data* се може користити пре свега у процесима:

- импутације недостајућих података,
- контроле прикупљених података на традиционални начин – валидност и конзистентност,
- повезивања са постојећим подацима – креирање комплексних база података, омогућавајући, између осталог, лонгитудиналну анализу,
- генерисања нових аналитичких алата – унапређење система званичне статистике: увођење нових производа (индикатора), скраћено време дисеминације, флеш оцене, и сл.)

Графичко представљање података – визуелизација (*Data Representation*) је важан концепт у *Big Data* истраживању. Наиме, добро дизајнирана информативна графика може да омогући истраживачу да врши поређења и доноси закључке, без додатног улажења у каузалност посматраних појава (Reinhard, 2010). То се постиже искоришћавањем способности људског ока: тренутним и без додатног напора покретом ока, високом пропустљивошћу и капацитетом за паралелну обраду, препознавањем различитих шаблона и корелација, макро/микро дуалношћу која омогућава преглед целе страницу или фокусирање на најситнији детаљ (Bret, 2006; Majooni, et al., 2018).

Овај приступ је илустрован на слици 6.7 на којој је приказан пример визуелизације коришћења различитих мобилних платформи у Вашингтону, добијен на основу гео-таговања (*geo-tagged*) *Twitter* порука (Washington, DC) (OECD/ITF, 2015).



Слика 6.7: Диференцијација корисника мобилних платформи у Вашингтону (OECD/ITF, 2015)

Из овог примера можемо закључити да се само на основу графичког приказа јасно може сагледати територијална диференцијација корисника андроид и ајфон мобилних платформи, тако да је даља статистичка обрада и анализа података (тестирање хипотезе о једнакости популација, нпр.) у овом случају излишна.

Наравно, разлози који су довели до ове поделе (разлике) - каузалност, остају отворена за даља истраживања.

У екстремним случајевима анализа података се чак и завршава на графичком представљању података, на начин да се процес доношења закључка базира на визуализацији података. Прво правило Едварда Туфтеа о статистичком графичком дизајну је „Покажи податке: Сви графички прикази, без обзира да ли су статистички или не, морају приказати истраживачима довољно информација да им пруже одговоре на њихова питања" (Tufte, 2001). Овај концепт има свој назив – инфографика (*infographic*) и представља концепт визуелног откривања знања базиран на графичким приказима, сликама и анимацијама, који уз минималну количину текста, на брз и једноставан начин даје увид у посматрану појаву (тему) (Lankow, et al., 2012).

Слика 6.8: Инфографика (Smiciklas, 2012)

Практично, инфографика комбинује податке са дизајном како би се омогућило визуелно учење (Слика 6.8). На тржишту постоји велики број алата за визуелизацију података, од којих се у *Big Data* аналитици најчешће користе: *Tableau, QlikSense, Microsoft Power BI, Fusion Charts, Highcharts, Datawrapper, Plotly, Sisense* итд. (Интернет 7).

6.7 Фаза 6: Анализа података

Настављајући се на пројектоване токове *Big Data* процеса долазимо до фазе анализе података (аналитике).

Слика 6.9: Фаза аналитике

Фаза 6: Анализа се одвија итеративно и паралелно са Фазом 5: Обрада. Анализом података може се доћи до информација које захтевају да се обави додатна обрада података. Активности у Фазама 5 и 6: Обрада и Анализа, могу почети и пре завршетка Фазе 4: Прикупљање.

У традиционалном истраживању Фаза 6: Анализа се састоји од пет подфаза које су обично узастопне, али се могу одвијати и паралелно, и могу се понављати:

6.1 Израда првих резултата
6.2 Верификација резултата
6.3 Интерпретација резултата
6.4 Примена правила о поверљивости
6.5 Финализација резултата

Имајући у виду да је предложени *Big Data* модел пројектован за коришћење у систему званичне статистике, намера је да он буде, у што је могуће већој мери, компатибилан са традиционалним моделом. Ово посебно важи за процес анализе података, с обзиром да се у традиционалним обрадама користи статистичка методологија. Зато се у првом кораку анализе података сугерише редукција *Big Data* на "*small data*", односно прелазака са *Big Data* на традиционални концепт анализе података (Derbeko, et al., 2016). Уколико је то није могуће, модел претпоставља коришћење *Big Data* аналитике (Слика 6.10).

Слика 6.10: Провера опције свођења *Big Data* на "*small data*" у *Big Data* моделу

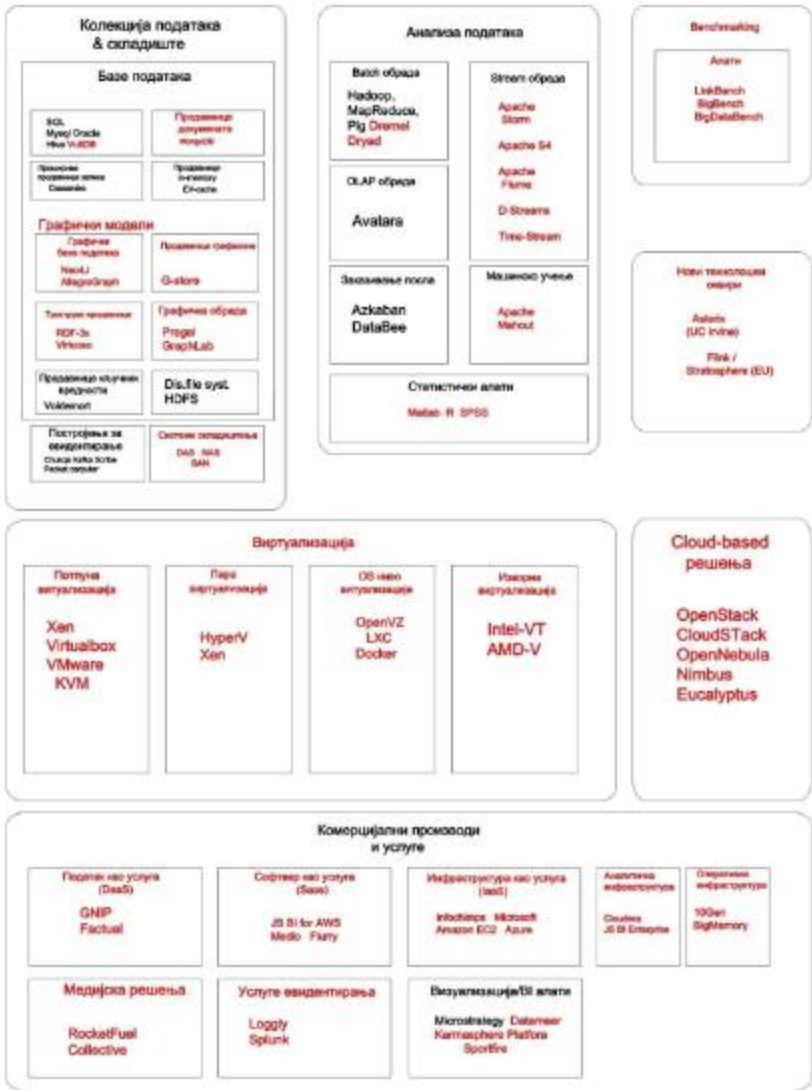
Big Data аналитика

Big Data аналитика је директној зависности од *Big Data* извора и прикупљених података који се користе у наставку истраживања. Многе традиционалне аналитичке методе које добро функционишу у процесу обраде и анализе умерених количина података не могу се применити за велике количине података, што захтева примену нових аналитичких метода, из најмање три врсте разлога (Fan, et al., 2014):

- Прво, традиционалне методе статистичке анализе се заснивају на статистичком закључивању (сигнификантности) базираном на оценама параметара добијеним на малим узорцима посматраног основног скупа (популације). Добијени закључци се потом генерализују на целу популацију. Са друге стране, *Big Data* узорци су масивни и представљају велики део, ако не и целокупну популацију. То доводи до тога да статистичко закључивање није увек релевантно за *Big Data*.
- Друго, многе конвенционалне аналитичке методе за мале узорке не могу се применити на *Big Data*, пре свега због њиховог неструктурираног формата и количине.
- Треће, *Big Data* карактеришу особине које се не разматрају у традиционалним (малим) узорцима, као што је њихова хетерогеност и низ атрибута које доводе до пристрасности (шума) у подацима (лажне корелације, случајне ендогеност и сл.).

С обзиром да је највећи број података из *Big* извора неструктурираног или послуструктурираног облика, неопходно је моделовање података и њихових релација у циљу редукције димензије података и њиховог структурирања. У наставку су описане методе *Big Data* аналитике који се користе за структурирање неструктурираних података, што укључује трансформацију података у организоване сетове, са јасно дефинисаним варијаблама и идентификованим односима међу њима (Hastie et al., 2013).

Алати, технике и технологије које се могу применити у *Big Data* истраживању могу се класификовати на различите начине. Ради илустрације је, на слици 6.11 приказана једна класификација *Big Data* технологија, производа и сервиса добијена на бази анализе случајева из праксе од стране аутора студије (Pääkkönen & Pakkala, 2015). Технологије, производи и сервиси који се не односе директно на *Big Data* решења исписани су црвеним словима.



Слика 6.11: Класификација *Big Data* технологија, производа и сервиса (Pääkkönen & Pakkala, 2015)

За потребе предложеног *Big Data* модела, класификација метода *Big Data* аналитике је извршена на следећи начин (Gandomi & Haider, 2015):

1. Текст аналитика,
2. Аудио аналитика,
3. Видео аналитика и
4. Аналитика друштвених медија.

У наставку поглавља описане су ове методе, без улажења у технологије, комерцијалне производе и сервисе за аналитичку подршку. Наведене *Big Data* методе су фундаментално различите од метода традиционалне статистичке анализе које се спроводе на релативно малим узорцима. Свака од ових метода користи различите технике, од којих су наведене само најважније, са становишта предмета и циљева истраживања докторске дисертације.

1. Текст аналитика

Скоро сваки извор података, а посебно Интернет, садржи мноштво људски (или машински) генерисаног текста који захтева правилно преузимање и обраду. Да би се искористио пуни потенцијал текста у базама података, потребне су специфичне технике за обраду природног језика (Blazquez & Domenech, 2017).

Метода текстуалне аналитике (*text mining*) се односи на процес којим се екстрахују (издвајају) информације из неструктурираних текстуалних података из *Big Data* извора: друштвене мреже, електронска пошта, блогови, онлајн форуми, одговори на анкету, корпоративни документи, вести и логови кол центара (*call center logs*) и трансформишу се у структуриране информације које подржавају доношење одлука заснованих на доказима (*evidence-based decision-making*) (Hurwitz & Nugent, 2013).

Текстуална аналитика укључује следеће технике: Обраду природног језика (*Natural Language Processing*), технике машинског учења (*Machine Learning*), статистичке методе, вештачку интелигенцију, технике класификације, језичко учење (*Linguistic Learning*), семантичка анализа и предиктивно моделирање (Kaur & Deepti, 2016).

2. Аудио аналитика

Аудио (говорна) аналитика је метода која анализира и издваја информације из неструктурираних аудио података. Технолошка решења заснована су на рачунарској лингвистици и машинском учењу, уз јаку регулаторну експертизу (Ernst & Young, 2016).

Аудио аналитика се базира на два технолошка приступа/технике:

- приступ заснован на транскрипту (такође познат као препознавање континуираног говора са великим вокабуларом (*large-vocabulary continuous speech recognition - LVCSR*) и
- фонетски приступ (Gandomi & Haider, 2015).

- **Приступ заснован на транскрипту** се одвија у две фазе: индексирање и претраживање.
    - У првој фази врши се покушај да се утврди и испише говорни садржај звука. Ово се врши помоћу алгоритама аутоматског препознавања говора који одговарају звуковима у речи. Речи се идентификују на основу унапред дефинисаног речника. Ако систем не успе да пронађе тачну реч у речнику, она враћа најсличнију реч. Излаз система је индексна датотека која садржи информације о низу речи изговорених у говору.
    - У другој фази користе се стандардне текстуалне методе за проналажење појма за претраживање у индексној датотеци.
- **Фонетски системи** раде са звуковима или фонемима. Фонеми су перцептивно различите јединице звука на одређеном језику које разликују једну реч од друге. Фонетски засновани системи се такође састоје од две фазе: фонетско индексирање и претраживање.
    - У првој фази, систем преводи улазни говор у низ фонема (за разлику од приступа заснованог на транскрипту где се говор претвара у низ речи).
    - У другој фази, систем претражује излаз прве фазе за фонетско представљање термина за претрагу.

3. Видео аналитика

Видео аналитика укључује различите технике за контролу, анализу и екстракцију информације из неструктурираних видео података (*video streams*).

У почетку фокусиран на покретне објекте, аутоматизована видео аналитика је углавном игнорисала све друге контекстуалне информације. Данашњи врхунски аутоматизовани системи за видео анализу интегришу и уче од свих доступних сензора и од свих претходних информација о сцени, циљевима и очекиваним понашањима како би додатно побољшали ефикасност. Ово омогућава аутоматизована решења за видео анализу за нове домене (Hakeem, et al., 2012).

На пример, паметни системи могу да прикупљају демографске информације о појединцима, као што су старост, пол и етничка припадност. Слично томе, малопродавци могу тачно да утврде број и структуру клијената, измере време боравка у продавници, да открију њихове обрасце кретања, мере време задржавања у различитим областима малопродајних објеката и да прате стање у редовима испред каса, и све то у реалном времену. Додатна вредност се може се добити коришћењем алгоритама за корелисање ових информација са демографским подацима корисника како би се донеле одлуке везане за пласман производа, њихову цену, оптимизацију асортимана и залиха, промоције, визуелног идентитета итд.

Аутоматско видео индексирање и проналажење објеката представља још један домен у видео аналитичким апликацијама. Широка распрострањеност уређаја за онлајн и офлајн снимање и праћење, само додатно наглашава потребу за индексирањем мултимедијалног садржаја у циљу једноставног претраживања и коришћења. Индексирање видео записа може се извршити на основу различитих нивоа информација доступних у видео садржају, укључујући метаподатке, звучни запис, транскрипте и визуални садржај видеа. У приступу заснованом на метаподацима, системи за управљање релационим базама података се користе за претраживање и проналажење видео материјала. Технике аудио аналитике и аналитике текста могу се применити за индексирање видеа на темељу повезаних звучних записа и транскрипата.

4. Аналитика друштвених медија

Аналитика друштвених медија односи се на анализу структурираних и неструктурираних података из канала друштвених медија који се генеришу из два основна извора:

→ Подаци и информације генерисане од стране корисника: сентименти (осећања, ставови) фотографије, видео записи и сл.) и

→ Подаци и информације које настају као резултат интеракције између мрежних ентитета које чине људи и организације.

Аналогно наведеној понуди, методе које се користе у аналитици друштвених медија могу се класификовати у две основне категорије (Gandomi & Haider, 2015):

- **Аналитика заснована на садржају** се фокусира на податке које корисници објављују на платформама друштвених медија, као што су различити постови, коментари, ставови, слике и видео записи. Такав садржај је углавном неструктуриран, обиман, динамичан и често садржи шум (Fan, et al., 2014). Текстуална, аудио и видео аналитика, може се применити за добијање одређених увида из таквих података.
- **Аналитика заснована на структури** (назива се и аналитиком друштвених мрежа - *social network analytics*) бави се синтезом структурних атрибута друштвених мрежа и издвајањем знања (интелигенције) из односа између ентитета који учествују у интеракцији. Структура друштвене мреже је моделирана кроз скуп „чворова" и „веза", који представљају учеснике и односе, респективно. Модел се може визуализовати као граф састављен од „чворова", који представљају индивидуалне актере унутар мреже и „веза" које су њихови међусобни односи. Анализа друштвених мрежа се може вршити преко два основна типа мрежних графова (Heidemann, et al., 2012):
    - У **друштвеним графовима**, веза између пара чворова означава једино да постоји веза (нпр. пријатељство) између одговарајућих ентитета. Такви графикони могу се бити истражити неким од *data mining* техника, како би се идентификовале заједнице или одредила чворишта (тј. корисници са релативно великим бројем директних и индиректних друштвених веза).

- У **графовима активности**, везе представљају стварне интеракције између било којег пара чворова. Интеракције укључују размену информација (нпр. ставове изражене кроз коментаре) и због тога су информативнији од социјалних графова.

Технике аналитике друштвених медија могу се категоризовати на следећи начин (Batrinca & Treleaven (2015):

- Обрада природних језика (*Natural language processing—NLP*),
- Аналитика вести (*News analytics*),
- Истраживање јавног мњења (*Opinion mining*),
- Екстракција (*Scraping*) друштвених медија,
- Сентимент анализа (*Sentiment analysis*) и
- Текстуална аналитика (*Text analytics*).

Предиктивна аналитика

Предиктивна аналитика обухвата различите моделе које предвиђају будуће исходе на основу историјских и тренутно расположивих података, откривањем образаца и релација у подацима, за разлику од експланаторних модела који се користе за тестирање каузалности (Shmueli & Koppius, 2011).

У зависности од типова расположивих података и задатих циљева истраживања, у предиктивној аналитици се јављају две различите парадигме: учење под надзором (*Supervised Learning*) и учење без надзора (*Unsupervised Learning*) (Hastie et al., 2013). У зависности од ових парадигми, евидентне су различите технике предиктивне аналитике, које се базирају на примени мултиваријационих статистичких метода, као што је регресија (нпр. вишедимензионални логит и пробит модели), кластер анализа, машинско учење (нпр. неуронске мреже) (Varian, 2014).

У литератури је евидентан тренд да се под машинским учењем подразумевају следећи алгоритми (Ray, 2017):

1. Линеарна регресија (*Linear Regression*),
2. Логистичка регресија (*Logistic Regression*),

3. Стабла одлучивања (*Decision Tree*),
4. SVM (*Support Vector Machine*),
5. Наивни Бајесов алгоритам (*Naïve Bayes),*
6. kNN (*k- Nearest Neighbors*),
7. Нехијерархијско класификовање (*K-Means),*
8. *Random Forest,*
9. Алгоритми редукције димензионалности (*Dimensionality Reduction Algorithms*) и
10. Алгоритми за побољшање градијента (*Gradient Boosting algorithms*).

На крају излагања везаног за Фазу 6. Анализа, можемо закључити да се практично сва питања и дилеме везана за методолошки приступ у *Big Data* истраживању (моделу) преламају у овој фази истраживачког процеса (*Big Data* аналитика). Намеће се закључак да аналитика представља највећи изазов у имплементацији конкретног истраживања у домену званичне статистике, и да се за свако појединачно истраживање мора тражити одговарајуће методолошко решење које ће у највећој мери одговорити предмету и циљевима истраживања на бази предложеног концепта *Big Data* модела.

6.8 Фаза 7. Евалуација

Код *Big Data* модела последње три фазе су промениле места, у односу на традиционални модел,. тако да фаза евалуације претходи фазама дисеминације и архивирања. У Фази 7: „Евалуација" врши се оцењивање комплетног процеса *Big Data* истраживања и утврђује потреба за евентуалних накнадним итерацијама, односно враћање на претходне фазе истраживања у циљу постизања одређених побољшања.

Код традиционалног модела истраживања, евалуација се најједноставније спроводи кроз израду извештаја о квалитету статистичког процеса (*quality reports*). У систему званичне статистике ови извештаји су стандардизовани и чине саставни део сваког истраживања (Eurostat, 2015).

Код *Big Data* истраживања стандардизовани извештаји о квалитету не могу да се спроведу, уколико се не користи традиционална статистичка методологија (базирана на теорији узорка или статистичким регистрима) што представља основну методолошку препреку у њиховом ширем коришћењу у домену званичне статистике.

Ипак, оно што је у овом тренутку извесно је да су *Big Data*, у зависности од извора, изложени различитим врстама статистичке пристрасности, као што је неодговарајућа покривеност, недовољна репрезентативност, пристрасност у избору узорка, грешке у мерењу. За разлику од грешака насталих у процесу избора узорка, ниво ових типова грешака неће бити смањен повећањем величине скупа података. Као основни извор пристрасности наводи се разлика између циљне (најчешће опште) и популације из које се генеришу *Big Data* (Интернет популација, популација корисника мобилне телефоније, и сл.) (Tam & Clarke, 2015a).

Напоменули смо да је највећи број података из *Big Data* извора исказано у недовољно структурираној форми и да су метаподаци који их описују најчешће непотпуни или неподударни, уколико уопште постоје. Према томе, дугорочна поузданост *Big Data* извора представља реално ограничење за текућу статистичку производњу. За ефикасно креирање и евалуацију дугорочних политика неопходни су конзистентни статистички производи који се прате у временским серијама, дужи временски период, често и кроз много година. У овом тренутку, *Big Data* извори су динамички и нестални - могу се променити у карактеру или нестати током времена. Ова неизвесност у трајању токова података и њихових извора угрожава поузданост статистичке производње и објављивања значајних временских серија.

Једно од решења представља комбинацију *Big Data* са подацима добијеним на основу традиционалних статистичких истраживања (анкете и регистри) у циљу компензација у разлици покривености између популација. Пажљиве анализе помоћи ће у разумевању потреба и проналажењу индивидуалних решења за сваку примену *Big Data* у домену званичне статистике (Hackl, 2016).

У неким случајевима, може бити корисно објавити статистике које се односе на делове популације која је покривена *Big Data* изворима. У прилог овоме иду и

наступајући трендови који указују на повећање интернет и мобилне пенетрације, што доводи до смањење разлике у покривености популација. Такође, треба имати у виду и да су традиционална истраживања све више изложена пристрасности, због све већег одбијања испитаника да учествују у истраживању.

На крају, можемо закључити да провера репрезентативности сетова података добијених из *Big Data* извора остаје отворено питање. Намеће се обавеза да се изврши својеврсна процена последица коришћења ових података, у вези са релевантношћу статистичких производа, њиховом тачношћу, упоредивошћу и другим димензијама квалитета. Такође, потребно је уложити додатни напор за дефинисање метаподатака, који се морају ускладити са новим изворима. Реално је очекивати додатне анализе које ће омогућити процену квалитета *Big Data* и статистичких производа који се добијају на основу ових података, у форми стандардних процедура извештаја о квалитету.

6.9 Фаза 8. Дисеминација

Фаза дисеминације у традиционалном истраживању се састоји од пет подфаза које су обично узастопне, али се могу одвијати и паралелно, и могу се понављати:

7.1. Ажурирање дисеминационих база података
7.2. Производња производа за дисеминацију
7.3. Управљање објављивањем производа за дисеминацију
7.4. Промоција производа за дисеминацију
7.5. Управљање захтевима корисника

У *Big Data* моделу, ова фаза захтева више пажње и опреза, што значи и јачу комуникацију са корисницима, управо због недовољно развијеног и често неразјашњеног методолошког поступка.

Приликом дисеминације резултата, неопходно је увести стандардизоване процедуре које ће омогућити корисницима да буду информисани о ограничењима и потенцијалној пристрасности до које је можда дошло приликом производње одређених информација – резултата. Ове процедуре ће се користити уместо (још увек недостајућих) извештаја о квалитету. Акценат треба да буде на потенцијалним

ограничењима и одступањима од стандардне статистичке методологије на коју су корисници навикли. Такође, неопходно је истаћи резултате провере процене ризика извођења *Big Data* пројекта у домену коришћења података о личности и нарушавања приватности.

Остале активности у овој фази, односе се на успостављање ауторских права над подацима и резултатима, писање публикација, навођење извора података, дистрибуцију података и контролу приступа подацима (Blazquez & Domenech, 2017). У процесу дисеминације, у *Big Data* моделу, се предлаже навођење коришћених извора у складу са највишим стандардима - *Reproducibility Enhancement Principles* (Stodden, 2017).

6.10 Фаза 9. Архивирање

Ова, последња фаза, односи се на архивирање и регистрацију свих прикупљених, обрађених и анализираних података, ради омогућавања њиховог дугорочног чувања података и евентуалне поновне употребе. Радње које треба извршити односе се на чување података у одређеним рачунарским системима, њихово преношење на друге платформе или медије, редовно прављење резервних копија података, стварање повезаних метаподатака, чување документације генерисане током целог процеса, контрола сигурности података и приватности и брисање података ако то захтевају законски прописи

Као што је већ наведено у фази 2. Пројектовање, код Процена ризика извођења *Big Data* пројекта архивирање прикупљених података мора бити у складу са највишим стандардима заштите података, у овом тренутку са уредбом (ЕУ) 2016/679 Европског парламента и Савета из априла 2016 (Уредба) о заштити физичких лица у односу на обраду података о личности и о слободном кретању таквих података (*General Data Protection Regulation*) GDPR.

7 Примена модела – студија случаја: индикатори одрживог развоја

У складу са постављеним циљевима истраживања докторске дисертације, у овом поглављу извршено је тестирање предложеног модела методолошког поступка, како би се потврдио потенцијал *Big Data* приступа у производњи статистичких резултата у домену званичне статистике. У ту сврху направљена је студија случаја која се бави редефинисањем (и имплементацијом) индикатора квалитета живота на националном нивоу у Републици Србији, а који омогућавају дефинисање (и имплементацију) УН индикатора 1.4.1: Удео становништва које живи у домаћинствима која имају приступ основним услугама (*Proportion of population living in households with access to basic services*). У студији нису описане све фазе предложеног модела, већ само оне које су релевантне за спроведено истраживање.

7.1 Фаза 1: Утврђивање потреба

У складу са предложеним *Big Data* моделом, акценат је стављен на имплементацију, што је могуће више, фаза традиционалног модела процеса статистичког истраживања.

Већ у уводној фази, утврђивању потреба, долази до концептуалних разлика између традиционалног и *Big Data* приступа у имплементацији индикатора квалитета живота на националном нивоу у Републици Србији, а које су истакнуте у појединачним подфазама.

7.1.1 Утврђивање потреба за информацијама

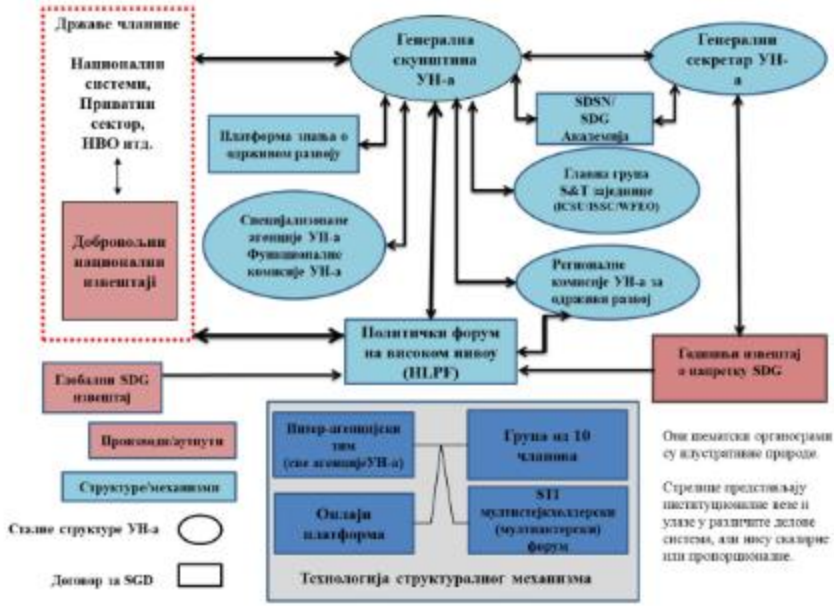
Полазну тачка у нацрту истраживања представља развојна Агенда Уједињених нација (УН) 2030 (Агенда 2030) у којој су дефинисани циљеви одрживог развоја. Агенда 2030 је званично ступила на снагу 1. јануара 2016. након усвајања резолуције на УН самиту у септембру 2015 (SDG, 2016). Према Агенди 2030, у наредних 15 година се од држава потписница (међу којима је и Република Србија) очекује да мобилишу све ресурсе како би искоренили сиромаштво, изборили се против неједнакости и нашли одговоре на климатске промене. Циљеви одрживог

развоја (17 глобалних циљева) проистекли су из Миленијумских циљева развоја УН.

У циљу праћења циљева Агенде 2030, било је неопходно дефинисати индикаторе за њихово праћење (Индикатори одрживог развоја). Приликом дефинисања сета индикатора одрживог развоја (индикатора) за мониторинг, водило се рачуна о следећим ограничењима (Portal, 2017):

- Јасно дефинисати неопходне индикаторе, и при томе користити оптималан број индикатора, како би процес праћења био једноставнији.
- Појединачни индикатор мора да буде јасно дефинисан и не превише компликован - тачно да осликава појаву која се прати и да буде довољно информативан за кориснике.
- Објективност извештавања мора бити задовољена (у складу са најбољом статистичком праксом).
- Неопходно је потпуно документовање методолошког процеса израде индикатора (метаподаци, извештаји о квалитету, извори података, методе коришћене за прикупљање и обраду података итд., морају бити садржани у извештајима о резултатима).

Инфраструктура УН за имплементација циљева одрживог развоја приказана је на слици 7.1 (IAP, 2017). Генерална скупштина УН-а се информише о напретку имплементације кроз Политички форум на високом нивоу (*High-Level Political Forum*) - Форум, који се састаје сваког јула у седишту УН-а у Њујорку. Форум представља централну платформа за праћење и ревизију циљева одрживог развоја, уз пуно учешће свих држава чланица УН, специјализованих агенција и других заинтересованих страна. Форум се информише преко Извештаја о циљевима одрживог развоја Генералног секретара, проценом глобалног и регионалног напретка на основу последњих доступних података из глобалног оквира индикатора циљева одрживог развоја који су припремили УН уз подршку међународних и регионалних организација.



Слика 7.1: Инфраструктура УН за имплементација циљева одрживог развоја (IAP, 2017)

Међуагенцијска експертска група за развој и имплементацију глобалних индикатора (IAEG-SDGs) дефинисала је оквир за праћење остваривања ових индикатора који за сада садржи 230 индикатора о којима је постигнута сагласност. Индикатори се разликују према прецизности, типу, могућностима мерења и сврстани су у три категорије (нивоа), у зависности од степена развоја методологије и расположивости података (IAEG-SDGs, 2018):

1. Индикатори првог нивоа су концептуално јасни, методологија и стандарди су дефинисани, и земље редовно производе податке за ове индикаторе.
2. Индикатори другог нивоа су концептуално јасни, методологија и стандарди су дефинисани, али државе не производе податке за ове индикаторе у редовној динамици.
3. Индикатори трећег нивоа нису развијени, односно нису дефинисане одговарајуће методологије и стандарди.

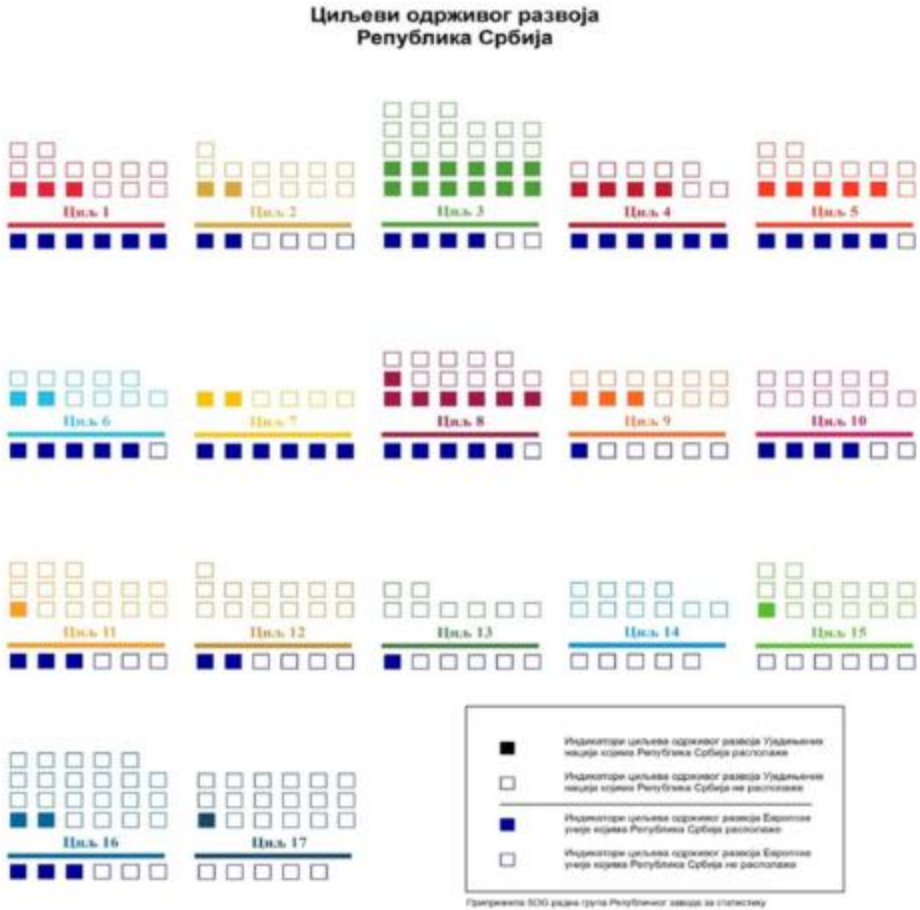
На дан 11. маја 2018: 93 индикатора припада првом нивоу, 72 припада другом, а 62 трећем нивоу. Такође, постоји 5 индикатора, који су класификовани у више нивоа (из разлога што су њихове компоненте припадају различитим нивоима – вишеструки индикатори) (Ndugwa, 2018).

Агенда 2030 предвиђа да глобални показатељи буду допуњени индикаторима на регионалном и националном нивоу. У новембру 2016. године, Европска комисија је дефинисала свој стратешки приступ у спровођењу Агенде 2030, укључујући циљеве одрживог развоја. Европска комисија је посвећена праћењу напретка ка циљевима одрживог развоја у контексту ЕУ. Сет ЕУ индикатора одрживог развоја (ЕУ индикатори) структуриран је на основу 17 дефинисаних циљева одрживог развоја и покрива социјалну, економску, еколошку и институционалну димензију одрживости коју представља Агенда 2030. 100 индикатора су равномерно распоређени унутар 17 циљева, тако да напредак у сваком циљу мери помоћу пет или шест припадајућих индикатора (без разматрања вишеструких индикатора), који одражавају његове широке циљеве и амбиције. Као званични статистички произвођач на нивоу ЕУ, Евростат има евиденцију о производњи статистике за праћење одрживог развоја на нивоу ЕУ (SD in EU, 2017).

Иако је сет ЕУ индикатора у великој мери усклађен са УН листом глобалних индикатора, они ипак не покривају све аспекте циљева одрживог развоја, односно не репродукују глобалну листу УН-а. У складу са тим, ЕУ индикатори су дефинисани на начин да омогућавају праћење одрживог развоја у контексту дугорочних политика ЕУ.

Циљ 1 се у УН Агенди 2030 наводи као „Крај сиромаштва свуда и у свим облицима", при чему се сиромашним сматрају они који живе са мање од 1.90 америчких долара дневно. Ова УН дефиниција сиромашних не одговара простору ЕУ, која у листи циљева одрживог развоја такође дефинише Циљ 1. са истим насловом али укључује праћење аспеката везаних за вишедимензионо сиромаштво и основне потребе становништва (Eurostat, 2018b). Вишедимензионо сиромаштво се односи на сиромаштво у приходима, материјалну ускраћеност (депривацију) и низак интензитет рада (Čomić, 2018). Основне потребе односе се на стамбену ситуацију и приступ здравственој заштити.

Генерални преглед циљева одрживог развоја за Републику Србији приказан је на слици 7.2 (DevInfo, 2018). Овај преглед представља резултат мапирања Индикатора циљева одрживог развоја према УН и ЕУ концепту, у Србији.



Слика 7.2: Упоредни приказ расположивости индикатора циљева одрживог развоја УН и ЕУ којима Република Србија располаже (DevInfo, 2018)

У сврху студије случаја, пошли смо од циља 1 и припадајућег УН Индикатора 1.4.1 (SDG, 2016) који је сврстан у трећи ниво.

→ *Циљ 1:* Крај сиромаштва свуда и у свим облицима.

- *Подциљ 1.4:* До краја 2030. осигурати да сви мушкарци и жене, а посебно сиромашни и рањиви, имају једнака права на економске ресурсе, као и приступ основним услугама, власништву и управљању земљиштем, односно другим облицима својине, наслеђству, природним богатствима, одговарајућим новим технологијама и финансијским услугама, укључујући микрофинансирање

Сагласно са овим подциљем дефинисана су два индикатора:

- *1.4.1*: Удео становништва које живи у домаћинствима која има приступ основним услугама (*Proportion of population living in households with access to basic services*)
- *1.4.2:* Удео одраслог становништва са сигурним правима власништва над земљиштем, (а) са правно признатом документацијом, и (б) који перципирају своја права на земљиште као сигурно, према полу и врсти власништва. (*Proportion of total adult population with secure tenure rights to land, with legally recognized documentation and who perceive their rights to land as secure, by sex and by type of tenure*)

И док за индикатор 1.4.2 постоје одређена методолошка решења Индикатор 1.4.1 је за сада методолошки непокривен (Ndugwa, 2018; Njiru, 2018). Циљ истраживања обухваћеног студијом случаја је да се дефинише, пре свега, најједноставнији начин за прикупљање података за овај индикатор (на националном нивоу), који ће се базирати на комбинацији *Big Data* и традиционалних извора.

7.1.2 Консултације и потврђивање потреба и дефинисање потребних резултата

Предложени приступ се базира на претпоставци да је за Републику Србију корисно да, осим ЕУ индикатора, производи и УН индикаторе одрживог развоја, чиме се стварају предуслови да се задовоље националне и међународне потребе за информацијама (на УН и ЕУ нивоу), као што је илустровано на слици 7.2. У том

правцу, на примеру индикатора 1.4.1, илустрована је потенцијална имплементација *Big Data* приступа.

Према УН препорукама, влада свака земље (тако и Србија) преузима одговорност за прикупљање података и валидацију индикатора, и то на следећи начин (OECD-UNDP, 2018):

- Одговорност за прикупљање **административних података** је на ресорним министарствима или одговарајућим регистрима, уз методолошку подршку коју ће обезбеђивати међународне организације и регионална тела, како би се олакшала размена искустава и конзистентност у свим земљама.
- Одговорност за **анкете и попише у домаћинствима** је на националним статистичким заводима.

Према предвиђеној динамици, подаци ће се прикупљати сваких 3-5 година у зависности од националних статистичких календара различитих земаља. Примена *Big Data* концепта, теоретски омогућава прикупљање података у реалном или готово реалном времену. Остаје још да се одреди одговорност за прикупљање података из *Big Data* извора. (Portal, 2017).

Први корак у методолошкој поставци код израчунавања индикатора 1.4.1 је дефинисање термина „основне услуге". Према УН приступу, основне услуге за потребе дефинисања овог индикатора су организоване у три категорије (UNSTATS, 2018):

1. **Основне инфраструктурне услуге**: вода и канализација, сакупљање и управљање чврстим отпадом, мобилност и транспорт и енергија,
2. **Услуге које утичу на социјални живот**: образовање, здравство, хитне службе, становање, брига о деци, услуге за старије и друге рањиве групе са посебним потребама,
3. **Услуге који утичи на квалитет живота**: јавна сигурност, урбанистичко планирање, култура и забава, спорт и јавни простори

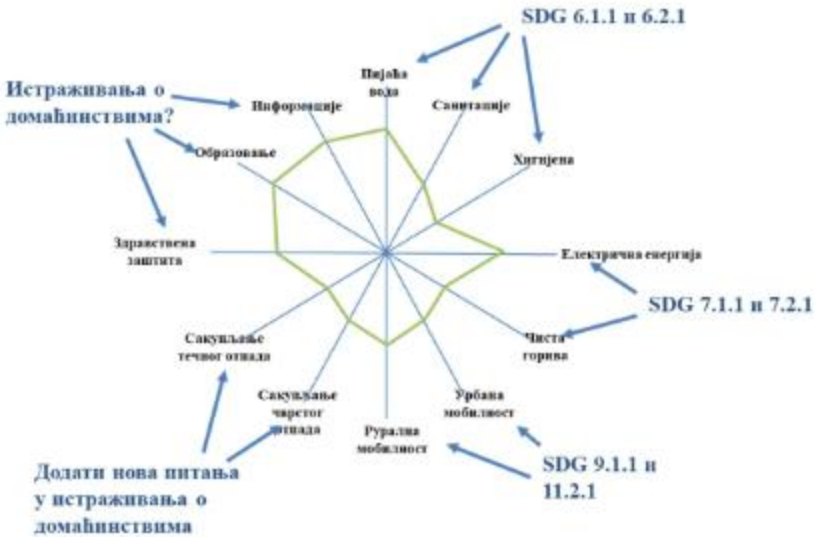
Јасно је да сваки од ових сервиса захтева **сет индикатора** који се базирају на одговарајућим дефиницијама - методолошком приступу. На основу истраживања тренутног статуса индикатора 1.4.1, коришћењем секундарних извора, може се

закључити да су прве две категорије услуга (основне инфраструктурне услуге и услуге које утичу на социјални живот) углавном методолошки покривене, као и да има и преклапања са другим индикаторима (Njiru, 2018) (Слика 7.3).

	Приступ услуги:	Показани SDG		Ниво
Физичка инфраструктура	**Безбедна и приступачна пијаћа вода**	6.1.1	Удео становништва које користи пијаћу воду из система којима се безбедно управља	I
	Безбедне санитације	6.2.1	Удео становништва које користи услуге санитације којима се безбедно управља, укључујући прање руку сапуном и водом	I
	Сакупљање отпада	11.6.1	Удео комуналног чврстог отпада који се редовно прикупља и који се на одговарајући начин одлаже у укупној количини генерисаног комуналног чврстог отпада, по градовима	II
	Мобилност и транспорт	11.2.1	Удео становништва које има одговарајући приступ јавном превозу, према полу, старости и инвалидитету	II
		9.1.1	Удео становништва које живи у кругу од 2 км од пута који је функционалан током целе године	II
	Модерна енергија	7.1.1	Удео становништва које има приступ електричној енергији	I
		7.1.2	Удео становништва које се примарно ослања на чиста горива и технологије	I
	Јавне површине	11.7.1	Просечан удео отворених јавних површина у изграђеним површинама просторима градова, према полу, старости и инвалидитету	III
Социјална инфраструктура	**ИКТ**	5.b.1	Удео особа које поседују мобилни телефон, према полу	I
		9.c.1	Удео становништва обухваћеног мобилном мрежом, према технологијама	I
	Здравље	5.6.1	Удео становништва обухваћеног мобилном мрежом, према технологијама	II
		5.6.2	Број држава које имају законе и прописе који гарантују пун и једнак приступ женама и мушкарцима старости 15 и више година, бризи за сексуално и репродуктивно здравље, информацијама и едукацији	III
	Образовање	4.1.1	Удео деце и младих: (а) у разредима 2/3; (б) на крају основног; и (ц) на крају нижег средњег образовања која постижу најмање минимални ниво достигнућа у (і) читању и (іі) математици, према полу	III

Слика 7.3 Покривеност основних услуга осталим SDG индикаторима (Njiru, 2018)

Потенцијални извори података за индикатор 1.4.1 се углавном референцирају на традиционална истраживања унутар система званичне статистике (базиране на домаћинствима: Анкета о потрошњи домаћинства, Анкета о радној снази и сл.), или су садржани у неким другим индикаторима (Слика 7.4) (Ndugwa, 2018).



Слика 7.4: Потенцијални извори података 1.4.1 индикатора (Ndugwa, 2018)

Из тог разлога, ради указивања на пун потенцијал предложеног *Big Data* модела, акценат у овој студији случаја даје се на трећу категорију услуга које утичу на квалитет живота. Полазећи од претпоставке да ће 1.4.1 бити композитни индикатор, који је представљен преко појединачних индикатора за сваку од три категорије основних сервиса, основни задатак истраживања се своди на дефинисање индикатора квалитета живота.

7.1.3 Идентификовање концепата (статистичких стандарда)

Као што је случај са дефинисаним индикаторима одрживог развоја, тако постоје и два релевантна статистичка концепта за истраживање квалитета живота: ЕУ приступ и УН приступ.

Према УН концепту услуге који утичи на квалитет живота (УКЖ) се могу пратити развојем индикатора који се односе на безбедност на јавним местима (*public safety*), урбано планирање (*urban planning*), културу и забаву, спорт и јавна места (*culture and entertainment, sport and public spaces*).

Након истраживања података из секундарних извора, евидентирани су документи и извори који се могу третирати релевантним за развој овог концепта, као што је за Нова урбана агенда (*The New Urban Agenda)*, која је усвојена на УН конференцији о Становању и одрживом урбаном развоју (Habitat III) одржаној 2016. године (UN, 2017).

На нивоу ЕУ два су релевантна истраживања за дефинисање и имплементацију индикатора квалитета живота:

- Европско истраживање о квалитету живота (*The European Quality of Life Survey – EQLS*) и
- Истраживање о приходима и условима живота (*EU statistics on income and living conditions - EU SILC*)

EQLS је основни алат за праћење и анализу квалитета живота у ЕУ и спроводио се 2003., 2007., 2011. и 2016. године. *EQLS* укључује материјалну и социјалну димензију и обухвата субјективне и објективне мере: ставове и преференције, као и ресурсе и искуства. Истраживање се базира на приступу да је „квалитет живота" широк појам и да обухвата не само индивидуално благостање, већ и квалитет јавних услуга и квалитет друштва у целини (Eurofound, 2013; Galonja & Šunderić, 2017; Kaliterna, et al., 2012).

У извештају *EQLS* истраживања из 2016. године, приказана је вишедимензионална перспектива европских грађана која обухвата различите области (Eurofound, 2017):

- Субјективно благостање;
- Животни стандард и аспекте депривације;
- Равнотежа између професионалног и породичног живота и брига о другима;
- Здравствена заштита, дуготрајна брига, брига о деци и друге јавне услуге; и
- Социјална несигурност, социјална искљученост и друштвене тензије, поверење и друштвену партиципацију и ангажовање заједнице.

EU SILC покрива објективне и субјективне аспекте социјалне инклузије и услова живота у монетарном и немонетарном смислу, како за домаћинства тако и за појединце. Користи се за праћење стратегије Европа 2020, посебно главног циља стратегије - смањења сиромаштва (Tinto, et al., 2018; OECD, 2017; EU SILC, 2018).

EU SILC истраживање не прати редовно сва питања која се тичу квалитета живота, већ ове аспекте прати у модулима (Čomić, 2018). Постоји иницијативна на нивоу ЕУ да се од 2023. године питања о квалитету живота интегришу у стандардни упитника *EU SILC-a* (Galonja & Šunderić, 2017).

На нивоу ЕУ индикатори квалитета живота су публиковани у Евростатовој онлајн публикацији „Квалитет живота у Европи – чињенице и погледи" (*Quality of life indicators - measuring quality of life*) која обухвата 8+1 димензију квалитета живота по мери грађана (Табела 7.1) (Eurostat, 2018).

Табела 7.1: Индикатори квалитета живота на нивоу ЕУ (Eurostat, 2018)

Рб	Димензија
1	Материјални услови живота
2	Продуктивна или главна активност везана за запошљавање
3	Здравље
4	Образовање
5	Одмор и друштвена интеракција
6	Економска и физичка сигурност
7	Управљачка и основна права
8	Природна и животна средина
9	Свеобухватно задовољство животом

За сваку димензију квалитета живота представљен је и анализиран скуп одабраних релевантних статистичких индикатора, као и временска компонента - трендови и разлика између земаља или демографских група.

У Србији се индикатори квалитета живота дефинишу на бази ЕУ приступа (Galonja & Šunderić, 2017). Предлог индикатора за Србију се базира на Евростатовој публикацији „Квалитет живота у Европи – чињенице и погледи" и *EQLS* (Табела 7.1).

Полазећи од резултата анализе ЕУ приступа у дефинисању индикатора квалитета живота у Републици Србији, закључак је да се овај приступ може применити у дефинисању УН индикатора 1.4.1. (UNSDSN, 2015). Према томе, за дефинисање УН индикатора 1.4.1 може применити микс приступ у компоновању композитног индикатора: за прве две категорије (основни инфраструктурни сервиси и сервиси

који утичу на социјални живот) дефинисаће се одговарајући појединачни индикатори, првенствено пратећи расположивост података из постојећих УН индикатора, док ће се приликом дефинисања индикатора за праћење квалитета живота искористи ЕУ приступ. Сем тога, дефинисани индикатор квалитета живота се може користити и самостално, за потребе праћења квалитета живота.

7.1.1 Провера расположивости података

У табели 7.2 дат је предлог индикатора и потенцијални извори података који је сачинио Тим за социјално укључивање и смањење сиромаштва, званични партнер Владе Републике Србије. Опис и дефиниције појединих индикатора квалитета живота су садржани у публикацији „Праћење социјалне укључености у Републици Србији" (Tim, 2017). Већина расположивих података за предложене индикаторе се добија из *EQLS* или истраживања која спроводи Републички завод за статистику Србије. Нека од ових истраживања су годишња (правосудна статистика), док су статистика избора, Анкета о коришћењу времена – РЗС, 2010, 2015 и *EU-SILC* модули из 2013. и 2015. године *ad-hoc* истраживања (модули) (Čomić, 2016). У *DevInfo* бази података се налазе подаци из различитих извора, најчешће УНИЦЕФ-а (DevInfo, 2018).

Табела 7.2: Предлог индикатора квалитета живота на националном нивоу у Републици Србији (Tim, 2017).

Рб.	Назив индикатора и димензија	Димензија	Дефиниција	Извор података и периодика доступности
1	Усклађеност професионалног и породичног живота	Квалитет живота	Удео лица који на питање „Да ли се ваше радно време уклапа са вашим породичним и друштвеним обавезама ван посла: врло добро, доста добро, не баш сасвим добро или се уопште не уклапа?" одговарају са „на баш сасвим добро" или „уопште се уклапа"	*EQLS* 2012, 2016. Подаци су доступни једном у четири године

2	Свеукупно задовољство животом	Квалитет живота	Просечна оцена добијена као одговор на питање „Кад све узмете у обзир, у којој мери сте задовољни својим животом данас?“, уз коришћење скале од 1 (врло незадовољан/незадовољна) до 10 (врло задовољан/задовољна) Удео лица која су одговорила да су прилично незадовољна својим животом данас (удео лица чији је одговор на скали од 0 до 10 три или мањи од три)	*EQLS* 2012, 2016. Подаци су доступни једном у четири године
3	Задовољство послом	Квалитет живота	Просечна оцена добијена као одговор на питање „Можете ли ми рећи колико сте задовољни Вашим тренутним послом?“, уз коришћење скале од 1 (врло незадовољан/незадовољна) до 10 (врло задовољан/задовољна) Удео лица која су одговорила да су прилично незадовољна својим послом данас (удео лица чији је одговор на скали од 0 до 10 три или мањи од три).	*EQLS* 2012, 2016. Подаци су доступни једном у четири године
4	Количина и квалитет слободног времена	Квалитет живота	Просечно време (у сатима) проведено у слободним активностима	Анкета о коришћењу времена – РЗС, 2010, 2015.
5	Просечан број сати плаћеног и просечан број сати неплаћеног посла	Родне неједнакости у расподели плаћеног и неплаћеног посла	Однос просечног времена (у сатима) проведеног у плаћеном послу код мушкараца наспрам жена. Однос просечног времена (у сатима) проведеног у неплаћеном послу код мушкараца наспрам жена.	Анкета о коришћењу времена – РЗС, 2010, 2015.
6	Постојање некога од кога би се могла тражити подршка у случају потребе Постојање некога са киме је могуће посаветовати се о личним стварима	Квалитет социјалне мреже	Удео лица која имају некога на кога се могу ослонити у случају подршке Удео лица која имају некога са киме се могу посаветовати о личним стварима	*EU-SILC* модули из 2013. и 2015. године

7	Степен поверења у друге људе	Квалитет социјалне мреже	Просечна оцена добијена као одговор на питање „Да ли мислите да се већини људи може веровати или да у контакту са људима морате бити изузетно опрезни?“, уз коришћење скале од 1 (која значи да морате бити изузетно опрезни) до 10 (која значи да се већини људи може веровати) Удео лица која су одговорила да се људима не може веровати као удео лица чији је одговор на скали од 0 до 10, три, или мањи од три.	*EQLS* 2012, 2016. Подаци су доступни једном у четири године
8	Степен поверења у институције (политички систем, правни систем, полицију)		Удео лица која су одговорила да уопште немају поверење у различите институције (удео лица која су на питање о степену поверења у различите институције одговорила да га уопште немају – тј. дала одговор три или нижи од три на скали од 1 до 10, где 1 представља „уопште немам поверење“, а 10 „имам пуно поверење“)	*EQLS* 2012, 2016. Подаци су доступни једном у четири године
9	Распрострањеност насиља 9а. у породици 9б. над полним слободама	Распрострањеност насиља	Број кривичних пријава поднет у случају насиља у породици у датој години Удео оптужених у односу на укупан број поднетих пријава насиља у породици Удео осуђених у односу на укупан број поднетих пријава насиља у породици Број кривичних пријава поднет у случају нарушавања полних слобода Удео оптужених у односу на укупан број поднетих пријава против полних слобода Удео осуђених у односу на укупан број поднетих пријава против полних слобода	Правосудна статистика – РЗС Правосудна статистика – РЗС
10	Излазност гласача	Партиципација у политичком животу	Проценат бирача гласалих на националним изборима у укупном броју бирача.	Изборна статистика – РЗС

	10а. на националним изборима 10б. на локалним изборима		Проценат бирача гласалих на изборима за одборнике скупштина општина и градова, у укупном броју бирача.	Изборна статистика – РЗС
11	Заступљеност жена на местима одлучивања у јединицама локалне самоуправе	Родна равноправност	Удео жена одборница, чланица општинских већа, градоначелница и/или председница општине у укупном броју одборника и градоначелника и/или председника општина	РЗС – *DevInfo* база

Из овог прегледа, очигледно је да за већину предложених индикатора важи да су базирани на истраживањима која се спроводе на европском нивоу сваке 4 године (*EQLS*) или више (Анкета о коришћењу времена – РЗС), што свакако не обезбеђује њихову ажурност. Ово је и закључак Европске комисије у чијим документима се наводи да је: „Недостатак благовремених информација о трендовима, посебно о сиромаштву, главна слабост у подацима која омета формулисање политике засноване на реалним чињеницама" (Mijatović, 2017). Најбољи доказ ове тврдње је новијег датума, из времена избијања светске економске кризе, када је постало очигледно да се не могу утврђивати ефекти краткорочних утицаја економских шокова, као и ефикасност одговора одређених политика само на основу редовних статистичких истраживања.

У том смислу *Big Data* приступ представља реалан потенцијал за прикупљање релевантних информација у много краћем или скоро реалном времену.

7.2 Фаза 2. Пројектовање

Наставак *Big Data* пројекта подразумева извођење следећих подфаза (везано за индикаторе 1-4, 6, 7 и 8 из табеле 7.2):

- Процена ризика извођења *Big Data* пројекта
- Пројектовање резултата
- Дефинисање променљивих
- Избор метода и инструмената прикупљања података
- Методологија узорка
- Методологија статистичке обраде

- Пројектовање производних система и токова рада

7.2.1 Процена ризика извођења *Big Data* пројекта

Пре него што се приступи даљем пројектовању, неопходно је урадити процену ризика извођења *Big Data* пројекта у домену коришћења података о личности и нарушавања приватности.

Након анализе упитника из контролне листе за процену ризика извођења *Big Data* пројекта (Табела 6.2) евидентирани су следећи ризици везани за истраживање:

→ Део 1. Везано за типове података који ће се прикупљати

1.2. Потенцијално коришћење псеудонимизованих подаци који не идентификују појединца директно, али који би се могли користити за издвајање јединствене особе применом постојећих и лако доступних средстава и технологија.

1.3. Потенцијално коришћење осетљиви (сензитивних) података.

→ Део 2. Везано за прикупљање података

2.3. и 2.4. Провера и обезбеђивање правног основа за прикупљање и обраду прикупљених података, како за институцију која спроводи истраживање тако и за достављаче података (провајдери мобилних услуга).

→ Део 3. Коришћење података

Неопходно је проверити да ли су добијене све регулаторне и друге потребне дозволе за наставак пројекта: коришћење података везано за кориснике мобилне телефоније – да ли су у складу са законима о телекомуникацијама; провера услова коришћења података на платформама друштвених медија и условима сагласности коришћења

→ Део 4. Процес комуникације

Везана за транспарентност и нивое транспарентности потребно је израдити план комуникација везан за коришћење података

→ **Део 6. Ризици и потенцијалне штете**

Ризике треба проценити одвојено од штета. Треба имати у виду да пристрасност у статистичком смислу представља скривени ризик који се може произвести као резултат коришћења података.

На крају, на основу одговора у одељцима 1-7, врши се процена вероватноћа наступања потенцијалног ризика, као и вероватноћа, величина и озбиљност потенцијалних штета, како би се утврдили да ли су ризици и штетне последице несразмерно високи у поређењу са очекиваним позитивним резултатима *Big Data* пројекта. Уколико је констатован потенцијални ризик и одлука је да се настави са имплементацијом пројекта, неопходно је да се предложе корективне акције за смањивање потенцијалних штетних ефеката.

7.2.2 Пројектовање резултата и дефинисање променљивих

Пројектовани резултати треба да буду што је могуће сличнији већ дефинисаним индикаторима. Дакле, циљ нам је да обезбедимо ажурне податке за изабране индикаторе из табеле 7.2 (циљамо што краћи период објављивања – дисеминације).

7.2.3 Избор метода и инструмената прикупљања података

Циљеви одрживог развоја захтевају годишње извештавање свих земаља потписница, на бази дефинисаних методолошких поступака и стандарда квалитета. То доводи до додатног оптерећења у домену националних статистичких институција које морају бити активно укључене у развој глобалних и националних оквира индикатора, кроз Међуагенцијску и експертску групу за индикаторе одрживог развоја (*Inter-agency and Expert Group on SDG Indicators*) а коју сазива Комисија за статистику УН (*UN Statistical Commission*) (UNECE, 2017).

Приликом разматрања коришћења *Big Data*, у претходном тексту смо навели да се (за сада) *Big Data* извори могу користити у комбинацији са традиционалним изворима података. Приликом коришћења званичних статистичких података, користе се стандардни инструменти и алати који су на располагању у постојећим примарним и секундарним истраживањима: анкетама и административним изворима.

→ **Административни подаци - регистри**

За велики број индикатора подаци се прикупљају из административних извора, које обично прикупљају ресорна министарства, а затим достављају статистичким заводима на даљу обраду и анализу. Стање у Србији у овој области карактерише недовољна употреба регистара. Разлози леже у инфраструктури, односно одсуству, пре свега регистра становништва, домаћинства и станова (кућа). Да би се прикупили годишњи подаци високог квалитета, Србија ће морати да ојача капацитете за обраду административних података и успостави системе интероперабилности који се заиста спроводе у пракси у складу са Европским оквирима интероперабилности (European Commission, 2017). Одсуство интероперабилности и заједничких стандарда у великој мери представља препреку за адекватну примену података из административних извора, што доводи до недовољне координације државних институција и статистичког система. Овиме се, на крају, доводи у питање примена заједничких решења у пружању јавних услуга (Milojković, 2012) .

→ **Анкете (првенствено у домаћинствима)**

Анкете у домаћинствима представљају најважнији традиционални примарни извор података за праћење индикатора одрживог развоја. Скоро свака земља у свету спроводи овакву врсту истраживања, без обзира на ниво развоја статистичког система и коришћења административних података. Анкете у домаћинствима су важан извор социо-економских података, посебно у земљама у којима су административни системи података недовољно развијени (као што је стање у Србији данас) или је неопходно мерити људска понашања и промене у ставовима (бизнис клима, нпр.)

Приликом дефинисања индикатора, методолозима су на располагању две врсте анкета у домаћинствима које се примењују у званичној статистици:

- Редовне анкете које су део статистичког система једне земље и спроводи их национални статистички институт према плану и програму статистичких истраживања, и

- Анкете које се спроводе под патронатом међународних организација, УН организација, Светске банке, Међународног монетарног фонда и сл. У пракси се најчешће спроводе истраживања као што су: *Multiple Indicator Cluster Surveys (MICS)* или *Demographic and Health Surveys (DHS), Living Standard Measurement Surveys (LSMS).*

Динамика спровођења ових анкета, које нису у плану и програму статистичких истраживања у Републици Србији није чврсто дефинисана. Последње истраживање о животном стандарду (*LSMS*), спровео је РЗС у сарадњи са Светском банком, 2008. године (LSMS, 2008). Са друге стране, УНИЦЕФ-ово Истраживања вишеструких показатеља (MICS) се спроводе у регуларној динамици на 4-5 година (RZS i UNICEF, 2014). Посебна вредност ових истраживања, између осталог, се састоји у обезбеђивању података који се не прикупљају кроз систем званичне статистике (с обзиром да нису део редовних статистичких истраживања), а односе се на рањиве групе становништва (Cappa, et al., 2018).

Дакле, уз неприкосновени квалитет и употребљивост прикупљених података остаје проблем динамике у којој се ова истраживања спроводе. Подаци прикупљени сваких четири или пет година, захтевају додатан рад на додатним пројекцијама, како би се обезбедила непристрасна оцена годишњег напретка и међународна упоредивост. Према томе, имплементацијом *Big Data* приступа, (барем теоријски) обезбеђују се велики скупови података, на континулани начин, што омогућава извештавања у месечној, чак и недељној периодици (уколико постоје такве потребе).

7.2.4 Методологија статистичке обраде

Пројектовање методологије (статистичке) обраде и анализе података односи се на најважнији корак у алгоритму - избор аналитичких метода (методе *Big Data* аналитике). Изабрана аналитичка метода је директно детерминисана расположивим изворима података, а она сама детерминише даљу имплементацију истраживања, нарочито фазе обраде и анализе података.

Кључно питање произашло из студије случаја је везано за избор методе за прикупљање и анализу података, што је у складу са налазима везаним за

дефинисање *Big Data* модела статистичких истраживања. Практично сва питања и дилеме везане за методолошки приступ у *Big Data* истраживању (моделу) преламају се у фази анализе (*Big Data* аналитика). Процес избора одговарајуће аналитичке методе и технике првенствено зависи од дефинисаног предмета и циљева истраживања и расположивих извора података.

У разматраном случају, **предмет истраживања** се односи на индикатор квалитета живота, као једне субкомпоненте индикатора одрживог развоја 1.4.1: „Услуге које утичи на квалитет живота". Овај индикатор се рачуна према ЕУ приступу за мерење квалитета живота, који одговара специфичностима Републике Србије.

Циљ истраживања је садржан у дефинисању одговарајућег индикатора за праћење квалитета живота у Републици Србији коришћењем иновативних метода базираних на *Big Data* концептима, а који ће довести до унапређења постојећих индикатора, пре свега у домену прецизности и брзине извештавања.

Након дефинисања предмета и циљева истраживања, алгоритам се наставља избором циљане **методе** *Big Data* аналитике, утврђивањем потенцијала за њену примену, испитивањем расположивости података (утврђивање *Big Data* извора) и избором одговарајуће технике *Big Data* аналитике (Слика 7.5).

Слика 7.5: Поступак избора одговарајуће технике *Big Data* аналитике

Избор циљане методологије зависи од расположивости оквира узорка из *Big Data* извора, чије постојање омогућава имплементацију метода случајног узорка, у складу са дефинисаним *Big Data* моделом – Слика 6.4. На слици 7.6 приказан је резултат имплементације овог корака за потребе дефинисања индикатора квалитета живота.

Слика 7.6: Алгоритам *Big Data* модела – располагање случајним узорком за индикатор одрживог развоја

За истраживање друштвених медија (без обзира на технолошку платформу која омогућава комуницирање – укључујући паметне телефоне), у овом тренутку, није могуће дефинисати оквир узорка у складу са теоријом узорка, док је за имплементацију анкетних истраживања на телефонској популацији (CATI) и општој популацији (CAPI) оквир узорка теоријски расположив.

На основу анализе предлога појединачних индикатора и потенцијалних извора података, садржаних у табели 7.2, закључено је следеће:

- Индикатори од 1-4, 6, 7 и 8 могу се пратити методом истраживања друштвених медија и одговарајућом *Big Data* аналитиком;
- Индикатори 5, 9, 10 и 11 могу се пратити имплементацијом метода анкетних истраживања на мобилним платформама, CATI техником на телефонској популацији која не користи паметне телефоне и традиционалним анкетама на делу опште популације – CAPI техником.

Методологија узорка за анкетна истраживања на мобилним платформама тек треба да се развије и није предмет истраживања које је обухваћено овом дисертацијом.

Ова врста анкетирања је у литератури наводи као реална опција у будућим истраживањима унутар статистичког система (Antoun, et al., 2015; Ohmori, et al., 2005; Poynter, 2015; Ricciato et al., 2016). У наставку рада ближе се одређује и прати истраживање друштвених медија као *Big Data* извора података за дефинисање индикатора квалитета живота.

У следећем кораку, а пре утврђивања расположивих извора, врши се анализа потенцијала за примену циљане методе истраживања друштвених медија у систему званичне статистике Републике Србије.

7.3 Анализа потенцијала за примену циљане методе истраживања друштвених медија у систему званичне статистике Републике Србије

Анализа потенцијала за примену циљане методе истраживања друштвених медија у систему званичне статистике Републике Србије базирана је на истраживањима Републичког завода за статистику Србије (RZS, 2018a).

→ Методолошке основе истраживања

Републички завод за статистику спровео је два истраживања о употреби информационо-комуникационих технологија у 2018. години. Прво се односи на домаћинства и појединце, а другим су обухваћена предузећа RZS (2018a).

Реализација: Истраживање је спроведено од 12. марта 2018. до 28. марта 2018.

Тип истраживања: Телефонски интервју.

Домен истраживања: Територија Републике Србије (без АП Косово и Метохија)

Величина узорка: 2.800 домаћинстава, 2.800 појединаца. Узорак је алоциран на подручја централне Србије (без Београда), АП Војводине и Београда, пропорционално броју домаћинстава. Обим узорка износи 2.800 домаћинстава и 2.800 појединаца

Тип узорка: Двофазни, стратификован узорак

Циљна популација:

- За домаћинства: циљна популација састоји се од свих домаћинстава с најмање једним чланом који има између 16 и 74 године живота

- За појединце: циљна популација састоји се од свих појединаца који имају између 16 и 74 године живота

Референтни период: Три месеца која су претходила телефонском интервјуисању

Метод прикупљања података: Телефонска анкета - CATI, било је дозвољено и посредно анкетирање (давање одговора уместо одсутног лица).

Истраживања су спроведена по *методологији Евростата* (Eurostat, 2018a).

Стопа одговора износила је 94,7%.

→ **Основне дефиниције популације**

Узимајући у обзир обухват наведеног извештаја и постављене циљеве истраживања, уведене су следеће дефиниције популације:

- **Општу популацију** Србије чине грађани Србије без КиМ. Укупни процењен број становника у Републици Србији (општу популација) у 2017. години је 7.020.858, од чега 51,3% чине жене (3.601.043), а 48,7% мушкарци (3.419.815). Настављен је тренд депопулације, што значи да је и коефицијент раста становништва, у односу на претходну (2016) годину, негативан и износи -5,3‰ (RZS, 2018).
- **Циљна популација домаћинства** састоји се од свих домаћинстава с најмање једним чланом старости између 16 и 74 године живота.
- **Циљна популација лица** састоји се од свих појединаца старости између 16 и 74 године живота. Величина циљне популације је мања од опште популације и износи приближно 5.300.000 лица.

Све остале дефиниције посматраних популација су ограничене наведеном циљном популацијом.

- **Онлајн (интернет) популацију** Србије чине грађане Србије који поседују интернет прикључак, без обзира на тип конекције и учесталост коришћења (унутар циљне популације).
- **Популација интернет корисника** чине грађане Србије који користе свакодневно користе интернет.

- **Офлајн популацију** чине грађани Србије који НЕ поседују интернет прикључак .
- **Популацију корисника друштвених медија** чине грађани Србије који користе друштвене медије (пре свега друштвене мреже: *Facebook* и *Twitter*, и који прате блогове).
- **Телефонску популацију** чине грађани Србије који поседују мобилни телефон.
- **Технолошки непокривену популацију** чине грађани Србије који не који поседују мобилни телефон и немају приступ интернету

→ **Резултати истраживања - домаћинства у Србији (2018):**

Истраживање показује да:

- 99,1% домаћинстава поседује ТВ
- 93,0% домаћинстава поседује мобилни телефон
- 47,6% домаћинстава поседује лаптоп
- 72,9% домаћинстава поседује интернет прикључак

- Интернет конекција
  - 67,5% корисника интернету приступа путем мобилних уређаја (телефон или таблет) и то је најзаступљенији тип конекције који, у односу на претходну годину, бележи пораст од 13,9%;
  - друга по заступљености је *ADSL* конекција коју употребљава 51,2% корисника.
- Широкопојасну интернет конекцију поседује 72,5% домаћинстава у Републици Србији.
- У последња 3 месеца:
  - рачунар је користило 70,7% лица, што чини повећање од 3% у односу на 2017. годину.
  - интернет користило 73,4% лица, док га 24,2% лица никада није користило.

- Преко 3.590.000 лица користи интернет сваког или скоро сваког дана - (92,2% онлајн популације Србије).

- Уређај који се најчешће користи за приступ интернету је мобилни телефон са 83,7%, следе персонални рачунар са 59,8% и лаптоп са 45,3%. 77,6%
- Куповину или поручивање робе путем интернета, у последња три месеца, обављало је 30,9% корисника интернета.
- 92,6% становништва користи мобилни телефон.
- 76,8% онлајн популације Србије узело је учешће на друштвеним медијима (мрежама као што су *Facebook* и *Twitter*), блоговима и сл. у последња три месеца
- 96,4% онлајн популације Србије од 16 до 24 године старости располаже налогом на друштвеним мрежама
- 53,3% онлајн популације Србије користи електронску пошту
- Преко 1.800.000 лица куповало је или поручивало робу/услуге путем интернета у последњих годину дана
- 21,3% интернет популације користило је *cloud* сервисе за складиштење или размену података.

→ **Разлике између онлајн и опште популације**

Најзначајнији јаз између домаћинства која поседују интернет прикључак у Србији, евидентиран је у односу на тип насеља (урбано и остало) у коме домаћинство живи и у зависности од материјалног статуса домаћинства.

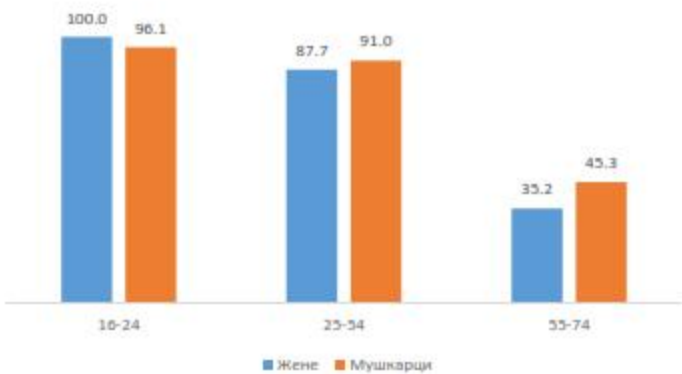
- Заступљеност интернет прикључака у градским насељима Србије износи 78,3%, наспрам 63,9% колико износи у осталим (сеоским и мешовитим) насељима Србије: 78,3%. Такође, у градским срединама је евидентирана стопа раста 5,4% у односу на претходну (2017.) годину, док тај раст у осталим деловима Србије у истом периоду износио 4,1%.

Материјални статус, мерен преко структуре домаћинстава према висини месечног прихода представља још један важан фактор тзв. дигиталног јаза (Marković, 2018):

- 87,8% домаћинства која имају месечни приход који премашује 600 евра поседују интернет прикључак
- 56,8% домаћинства која имају месечни приход до 300 евра поседују интернет прикључак

Напомена: просечни месечни приход домаћинстава у Србији у новцу и натури у трећем кварталу 2018. године, износили су 63.832 динара, што према просечном званичном средњем курсу Народне банке Србије, у истом периоду износи приближно 540 евра (RZS, 2018b). Истовремено, линија апсолутног сиромаштва 2017. години износила је 12.045 динара месечно по потрошачкој јединици, а потрошњу нижу од тог износа имало је 7,2% становника Републике Србије (Tim, 2018).

Такође, евидентна и очекивана разлика је и у уделу корисника интернета, према одређеним социо-демографским обележјима, нарочито нивоу образовања, старости и полу (Слика 7.7):



Слика 7.7: Корисници интернета према старости и полу (%) (RZS, 2018a)

Анализа испитаника према полу показује да је у последња три месеца 76,8% особа мушког пола, а 70,1% особа женског пола користило интернет. Такође, више особа женског пола (74,4%) у односу на мушки (66,4%) користе друштвене мреже.

Детаљнију анализу је могуће спровести на основу табеле 7.3.

Табела 7.3: Приступ информационо-комуникационим технологијама у Републици Србији (у %) (RZS, 2018а)

Одговор (домаћинства)	Приход			Регион			Тип домаћинства		Укупно
	До 300 евра	300-600 евра	Преко 600 евра	Централна Србија	Војводина	Београд	Градска	Остала	
А1_0 Уређаји заступљени у домаћинствима (вишеструки одговор)									
Персонални рачунар (PC)	54,8	80,6	87,9	69,0	69,3	81,1	78,2	61,8	72,1
Лаптоп	31,4	50,6	71,5	42,9	43,2	61,5	55,0	35,0	47,6
ТВ	99,1	98,9	99,5	99,4	98,3	99,3	98,8	99,6	99,1
Радио-пријемник	55,6	71,9	80,4	71,7	55,7	72,4	70,8	61,1	67,2
Мобилни телефон	88,5	94,8	96,7	92,3	91,9	95,5	95,3	89,0	93,0
А1: Да ли Ви или било ко из Вашег домаћинства има приступ интернету код куће?									
Да	56,8	80,6	97,8	69,3	70,7	82,2	78,3	63,9	72,9
Не	43,2	19,4	12,2	30,7	29,3	17,8	21,7	36,1	27,1
А2: Тип интернет конекције (вишеструки одговор)*									
DSL(ADSL)	53,3	51,1	50,5	58,4	53,0	37,8	47,7	58,4	51,2
Кабловски интернет	36,8	43,1	47,4	32,8	41,2	59,2	47,8	31,5	42,5
Мобилни телефон (таблет, USB) (путем 3G мреже)	60,8	68,1	73,9	62,4	68,3	74,7	68,7	65,0	67,5
Дајл-ап приступ путем телефонске линије или ISDN	0,0	0,8	0,2	0,4	0,3	0,5	0,5	0,2	0,4
Мобилни телефон (путем GPRS-а)	4,7	3,9	4,7	7,5	0,7	3,5	4,	5,5	4,5

*Подаци се односе н домаћинства која су на питање А1 одговорила „Да".

- **Закључци анализе**

У Србији се бележи континуирани напредак у информатизацији друштва, али и значајна разлика у односу на развијене земље и ЕУ мерено *DESI* индикатором који обухвата повезаност, људски капитал, употребу Интернета, интеграцију дигиталних технологија и дигиталне јавне услуге (DESI, 2018). На слици 7.8 приказане су просечне вредности категорија *DESI* за Србију у односу на просек ЕУ и земље у окружењу (Ratel, 2017).

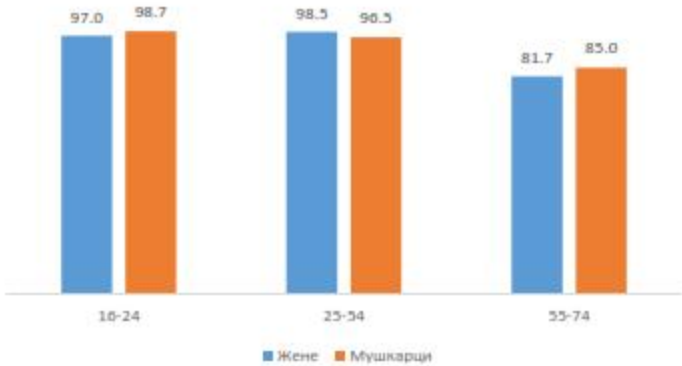
Слика 7.8: Просечне вредности категорија *DESI* за Србију у односу на просек ЕУ и земље у окружењу (Ratel, 2017)

Највеће заостајање је евидентирано у категорији „повезаност", која представља ниво инфраструктуре неопходне за имплементацију дигиталне економије и друштва и нуди информације о врстама и квалитету приступа интернету, као и његовој приступачности. Слично је са нивоом развоја јавних дигиталних сервиса и коришћењу интернета у целини, док су нешто бољи резултати у домену људског капитала и интеграције дигиталних технологија.

Ови резултати указују на озбиљан потенцијал али и високи степен ограничења када је реч о имплементацији онлајн истраживања у Републици Србији.

Када је реч о сентимент анализи, чија се примена предлаже у конкретној студији случаја, важне су следеће пропорције:

- 92,6% становништва користи мобилни телефон (преко 4.910.000 лица)



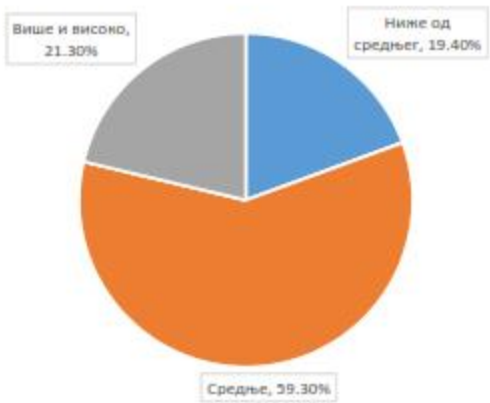
Слика 7.9: Корисници мобилне телефоније према годинама старости и полу, у Србији (у %) (RZS, 2018а).

- 77.6% онлајн популације Србије (преко 3.000.000 лица) користи паметне телефоне (*smartphone*) у приватне сврхе (Табела 7.4), што чини око 42% опште популације.

Табела 7.4: Структура корисника паметних телефона у Србији (у %) (RZS, 2018а).

Старост						Пол		Образовање			Радни статус				Укупно
16-24	25-34	35-44	45-54	55-64	65-74	Мушки	Женски	Ниже од средњег	Средње	Више и високо	Запослен	Незапослен	Студент	Остали	
89,5	86,8	76,6	76,8	58,7	57,2	79,5	75,6	68,8	79,2	81,8	81,9	75,3	91,3	66,0	77,6

- Интернет користи 73,4% лица (преко 3.890.000 лица у последња три месеца)



Слика 7.10: Образовна структура онлајн популације у Србији (RZS, 2018а).

- Уређај који се најчешће користи за приступ интернету је мобилни телефон са 83,7%, следе персонални рачунар са 59,8% и лаптоп са 45,3%.

- 76,8% онлајн популације Србије (преко 2.900.000 лица) узимало је учешће на друштвеним медијима (мрежама као што су *Facebook* и *Tweeter*), блоговима и сл. у последња три месеца (Табела 7.5 и Табела 7.6). Ова популација се великој мери поклапа са популацијом корисника паметних телефона (*smartphone*) – разлика је у 200.000 лица.

Табела 7.5: Структура онлајн популације Србије која узима учешће на друштвеним мрежама према годинама старости и полу, (у %) (RZS, 2018а).

Одговор (лица)	Године						Пол	
	16-24	25-34	35-44	45-54	55-64	65-74	Мушки	Женски
Учешће у друштвеним мрежама (*Facebook, Tweeter*)	96,4	85,7	74,2	56,9	38,7	33,4	66,4	74,4

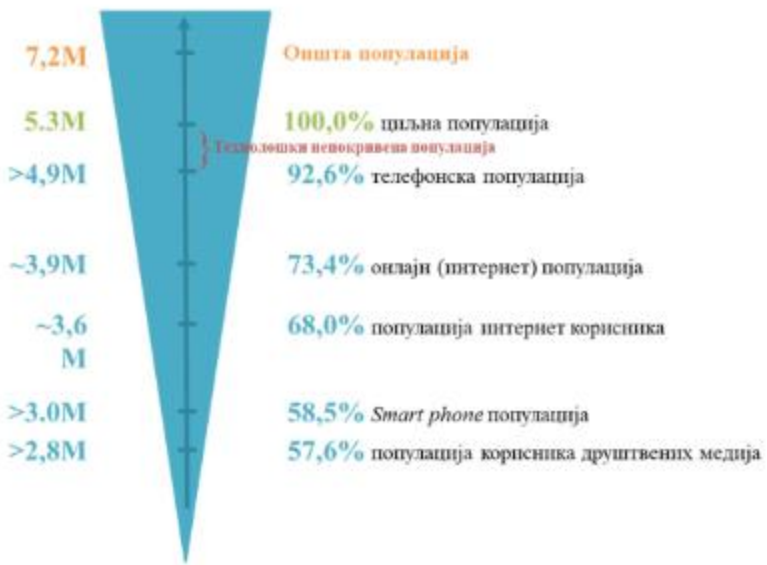
Напомена: 96,4% онлајн популације Србије од 16 до 24 године старости има налог на друштвеним мрежама

Табела 7.6: Структура онлајн популације Србије која узима учешће на друштвеним мрежама према образовању и радном статусу (RZS, 2018a).

Одговор (лица)	Образовање			Радни статус				Укупно
	Ниже од средњег	Средње	Више и високо	Запослен	Незапослен	Студент	Остали	
Учешће у друштвеним мрежама (*Facebook, Tweeter*)	69,8	72,2	65,4	70,2	72,9	98,7	58,9	70,3

Напомена: 98,7% студентске онлајн популације има налог на друштвеним мрежама.

Сумарни графикон популације Србије приказан је на слици 7.11:



Слика 7.11: Сумарни приказ популације Србије према пенетрацији ИКТ

На основу спроведене анализе може се закључити да се онлајн истраживања у Србији могу имплементирати на онлајн популацији, и да се закључци могу доносити само на овом делу опште популације. Поставља се питање да ли је, и у којој мери, могуће закључивање на нивоу опште популације. Аналогија за

дефинисање и утврђивање корективног фактора за тотални обухват се може успоставити на примени и модификовању методолошких поступака за примену телефонских анкета, из времена док је разлика између телефонске и опште популације била на нивоу разлике између онлајн и опште популације (Groves &. Nicholls, 1986; Kissinger, 1999; Lavrakas, 1987; Watson et al., 1995). Такође, треба узети у обзир да је оквиром узорка обухваћена популација од 16-74 године и да се за потребе конкретног истраживања сентимента овај оквир мора променити померајући обе границе годишта.

- Уколико се користи друштвени медији као извор података, они се могу уопштити на нешто више од ¾ онлајн популације Србије (76,8%) што чини преко 2.800.000 лица (корисници друштвених медија). За њих се може применити предложени *Big Data* модел базиран на анализи сентимента.

Остатак онлајн популације се може искористити за друге видове онлајн истраживања (укупну онлајн популацију чини преко 3.890.000 лица).

92,6% становништва користи мобилни телефон (преко 4.910.000 лица) и они чине телефонску популацију Србије.

- Један део телефонске популације Србије (око 200.000 лица) чине **корисници паметних телефона који нису корисници друштвених медија** и они могу бити покривен коришћењем метода анкетирања на бази апликација технологија паметних телефона (Mavletova & Couper, 2016; Miller, 2017; Ohmori et. al; 2005; Poynter, 2015; Simas & Wilson, 2018).
- Већински остатак телефонске популације може се покрити класичним CATI (*Computer-assisted telephone interviewing*) приступом. У циљу истраживања, из ове популације треба искључити 2.800.000 лица (корисници друштвених медија), тако да је оквир за CATI: 4,91M – 0,2M – 2.8M = 1.9M лица (M – милион).
- 7,4% становништва из циљане популације (нешто преко 390.000 лица) остаје ван домашаја ИКТ. Они се могу покрити коришћењем традиционалних теренских истраживања – CAPI (*Computer-assisted personal interviewing*) приступом.

7.4 Сентимент анализа

Након дефинисања предмета и циљева истраживања, избора истраживања друштвених медија као циљане **методе** *Big Data* аналитике и утврђивања потенцијала за примену ове методе, алгоритам приказан на слици 7.5 се наставља испитивањем расположивости података (утврђивање *Big Data* извора), како би се приступило избору одговарајуће **технике** *Big Data* аналитике.

Проучавањем литературе и анализом имплементираних случајева из статистичке праксе дошли смо до закључка да је за праћење индикатора од 1-4, 6, 7 и 8 сентимент анализа, као технике аналитике друштвених медија, најпримеренија.

Користили смо се методом бенчмарк анализе (*benchmarking*), из разлога што не располажемо сетом одговарајућих података великог волумена (V). У уводној фази истраживања спровели смо анализу документационих извора (секундарно истраживање онлајн извора), анализу цитираности и на крају бенчмарк анализу.

Радове из области сентимент анализе (СА) смо класификовали у 2 области:

- Радове који се генерално баве СА,
- Радови који развијају апликативну страну СА у домену званичне статистике.

Пошли смо од општег, ка посебном, изучавајући прво генерални приступ сентимент анализи (на примеру *The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation*, Zimbra et al., 2018), да бисмо на крају анализирали специфичну студију СА у домену званичне статистике: *Social media sentiment and consumer confidence* (Daas & Puts, 2014).

Најједноставнија дефиниција сентимент анализе је да је то техника за обраду текста која је у стању да класификује мишљење (ставове) о дефинисаној појави углавном у три категорије (тростепена скала): позитивно, неутрално, негативно, мада ова скала може бити и петостепена (Abbasi et al. 2018). СА се бави претраживањем и проучавањем текста из кога се идентификују и издвајају субјективне информације из онлајн извора, чиме се омогућава аналитичарима да разумеју друштвени сентимент (однос, став, мишљење) према посматраној појави (Zimbra, at al. 2018).

У највећем броју случајева СА је ограничена на основне анализе сентимента и њене метрике засноване су на бројању – фреквенцијама дефинисаних модалитета на тростепеној или петостепеној скали. Са недавним напретком вештачке интелигенције, нарочито техника дубоког учења (*deep learning*), способност алгоритама за текст аналитику је значајно побољшана (Gupta, 2018).

Сваки вид комуникације пружа прилику за приступ и разумевање перспектива корисника о темама од интереса, а оне садрже информације које могу објаснити и предвидети пословне и друштвене феномене.

Непрекидно се генеришу огромне количине текстова из различитих медија које креирају њихови корисници кроз различите канале, као што су:

- Традиционални медији – онлајн издања (Balahur et al. 2013; Pang & Lee, 2008; Raina, 2013),
- Друштвени медији (Asur & Huberman, 2010; Balahur, 2013; Fan & Gordon, 2014; Godbole et al. 2007; Hutto & Gilbert, 2014; Paltoglou & Thelwall, 2012, Thelwall et al., 2011)
- Маркетинг извори (Cambria et al. 2013; Chamlertwat et. al, 2012; Pontiki et. al. 2015; Schweidel & Moe, 2014),
- *Web* форуми и дискусионе групе (Abbasi et al. 2008; Li & Wu, 2013; Yang et al. 2013),

Сентимент анализу теоријски можемо разматрати као да је реч о анализи јавног мњења о одређеној појави (утврђивање сентимента на тростепеној скали).

Значај СА се најбоље може сагледати кроз примере његове примене. У литератури се као најчешће коришћени сервис за СА наводи *Twitter*. На тржишту се могу пронаћи бесплатни системи за *Twitter* сентимент анализу (ТСА), развијени од стране академске заједнице, као и комерцијални алати који се базирају на месечној претплати.

Комерцијални ТСА алати се могу груписати у две основне категорије (Abbasi et al. 2008):

- Самостални комерцијални алати – углавном се базирају на традиционалном (статистичком) приступу (Cambria et al. 2013; Daas & Puts, 2014; Pak & Paroubek, 2010; Pang & Lee, 2008)
- ТСА као стандардни део онлајн софтвера за пословну интелигенцију - базирају се на примени машинског учења и осталим техникама пословне аналитике, који захтевају развој модела заснованог на учењу заснованом на образовном сету. (Boiy & Moens, 2008; Neethu & Rajasree, 2013) Thelwall, 2011; Zimbra et al. 2018)

СА се дефинитивно може посматрати као (нехијерархијска) класификација – одређени број испитаника се класификује у дисјунктне скупове популације:

- корисници друштвених медија (0,1)
- корисници одређених сервиса – друштвених медија (*Facebook*, *Twitter* итд.) (0,1)
- испитаници према ставовима (+, 0, -)

Свака од наведених класа се може и треба додатно агрегирати према одређеним атрибутима - социо-економским карактеристикама.

При томе посебну пажњу захтевају грешке у класификовању, које су неизбежне у свим врстама класификовања, тако и овде стага је неопходно акценат ставити на евалуацију добијених резултата (Hassan et al. 2013; Zimbra, et al. 2018).

7.5 Закључна разматрања студије случаја

У конкретном истраживању – дефинисање индикатора квалитета живота (1-4, 6, 7 и 8), за прикупљање података предлаже се комбинована метода која се састоји из: анализа сентимента (*Big Data* приступ), апликације паметних телефона, CATI и CAPI (Табела 7.7).

Табела 7.7: Методе за прикупљање података за индикаторе (1-4, 6, 7 и 8) квалитета живота

Бр.	Метода	Популација	Предложена динамика спровођења	Кумулатив популације (у %)
1	Анализа сентимента	Популација корисника друштвених медија	Недељно	57.6
2	Апликације паметних телефона	Део телефонске популације који поседује паметни телефон	Месечно	58.5
3	CATI	Остатак телефонске популације (без Бр. 1 и 2)	Квартално	92.6
4	CAPI	Лица ван домашаја ИКТ	Полугодишње	100

Дакле, што се тиче брзине извештавања на месечном нивоу се може обезбедити покривеност од приближно 60% циљане популације, а на кварталном око 93%. Такође, за брзо извештавање, након првог полугођа, могу се утврдити пондери за допринос сваке од посматраних популација. На тај начин, са фиксирањем пондера, прво за полугодишње, а затим и квартално истраживање могуће је добити флеш оцене индикатора на месечном нивоу. Ове оцене се након кварталног истраживања поново израчунавају, па након полугодишњег се израчунавају коначне оцене индикатора. Овим се база пондера увећава, а самим тим и смањује одступање флеш оцене од коначне вредности индикатора.

Такође, у ближој будућности се, може очекивати повећање учешћа популација 1 и 2, чиме ће флеш оцене бити додатно прецизније.

Највећи изазов у имплементацији *Big Data* пројекта представља фаза прикупљања података, односно предложени метод за дефинисање индикатора квалитета живота, који се базира на употреби сентимент анализе. Један од начина аквизиције података је набавка (куповина) података од специјализованих (приватних) компанија које се овим баве, као што је наведено у студији случаја истраживања пословне климе (Daas & Puts, 2014). Једно од кључних питања које се овде поставља је

успостављање оптималног баланса између квалитета расположивих података и питања заштите приватности (Abowd, 2017). И док се имплементацијом контролне листе врши процена ризика извођења *Big Data* пројекта (фаза 2. Пројектовање) и у великој мери решава проблем заштите приватности, питање квалитета остаје отворено.

На крају, на основу резултата спроведене анализе студије случаја може се закључити да у Републици Србији постоји солидан потенцијал за примену *Big Data* модела и сентимент анализе у дефинисању и праћењу индикатора квалитета живота и одрживог развоја – али за сада само на популацији која је присутна на друштвеним медијима (види ИКТ истраживање).

За покривање опште популације је неопходно спровести модификовано телефонско истраживање (CATI базиран на паметним телефонима), класични CATI и традиционалну анкету (за покривање оних који су ван домашаја технологија, поштујући основне премисе одрживог развоја – нико не сме да остане необухваћен (*Leaving no one behind*) (SDG, 2016).

Правци даље акције у овој области се могу сублимирати на следеће препоруке:

- За индикаторе квалитета живота, спровести базично (нулто) традиционално истраживање, правећи 4 популациона стратума (1. друштвени медији, 2. паметни телефони, 3. остали телефони, 4. без телефона).
- Утврдити пондере за сваки популациони стратум.
- Урадити пилот истраживања за стратуме 1-3 и након извршене евалуације ревидирати пондере.
- Утврдити допринос прва два стратума укупном КРЕТАЊУ појаве (сентимента).
- Утврдити допринос првог стратума укупном КРЕТАЊУ појаве (сентимента).
- Дефинисати флеш процену КРЕТАЊА појаве (сентимента).
- Једном годишње поновити комплетно истраживање ради евалуације и ревидирања пондера.

За очекивати је убудуће повећање утицаја прва два, односно првог стратума. На тај начин би се омогућило ефикасније праћење политика у Србији.

8 Научни и стручни доприноси

8.1 Научни доприноси

Најзначајни научни доприноси дисертације су:

- Унапређење методолошког поступка за спровођење истраживања у домену званичне статистике, коришћењем *Big Data* концепта.
- Дефинисање *Big Data* модела процеса статистичког истраживања и
- Унапређење методологије за мониторинг циљева одрживог развоја.

Појава *Big Data* узроковала је преиспитивање научне и стручне јавности, не само у домену до скоро важеће (традиционалне) методологију научних истраживања, већ и шире научне заједнице. *Big Data* концепт је увео нове парадигме научног сазнања и закључивања. Радећи са подацима великог обима, истраживачи су у позицији да откривају одређене законитости и сазнања само на основу увида у податке који су им на располагању (концепт инфографике). Додатно „рударење" података (*data mining*), им омогућава имплементацију аналитичких метода без тестирања традиционалних статистичких хипотеза. На тај начин су се створили услови за имплементацију савремених информатичких концепата (као што су вештачка интелигенција, *blockchain* технологија, и сл.) у многим сферама научних истраживања. Са друге стране, опоненти и критичари ове парадигме за *Big Data* наводе да представља праксу без теорије ("*The no theory thesis*"), док неки аутори чак иду до тога да наводе да *Big Data* представља крај теорије у науци (Leonelli, 2014; Mazzocchi, 2015).

Традиционална методологија статистичког истраживања која се примењује у системима званичне статистике базира се на прикупљању података употребом анкетних истраживања, коришћењу административних извора података, регистара, пописа и сл. Основни недостатак ових метода је у томе што прикупљања података, њихова обрада и објављивање резултата истраживања траје дуже него што би корисници то желели. Такође, све је присутнија теза да је неопходно растеретити даваоце података, смањити њихово оптерећење, без обзира да ли се ради о правним или физичким лицима.

У дисертацији је анализиран потенцијал коришћења велике количине података (*Big Data*) у систему званичне статистике. Указано је на главне изазове у имплементацији *Big Data* концепта и предложене су основне одреднице будућег правца развоја модела истраживања коришћењем *Big Data* концепта у званичној статистици. Једна од основних замерки које се односе на *Big Data*, са методолошког становишта је да овако прикупљене информације нису добијене на основу података који су прикупљани у статистичке сврхе. У дисертацији је изложена аналогија са употребом административних података – регистара, који представљају тежњу да постану доминантан извор података у систему званичне статистике. Из тог разлога, полазну основу предложеног методолошког поступка представљају модели процеса статистичких истраживања.

Утврђена су кључна ограничења употребе нових технологија као суплемент традиционалним статистичким истраживања која се у најкраћем могу свести на закључак да истраживања базирана на *Big Data* концептима у овом тренутку представљају један од додатних извора података у систему званичне статистике. Разматрана је веза између традиционалног концепта статистичког истраживања и примене *Big Data* концепта и сагледана улога ИКТ у савременом приступу који подразумева интернет и друштвене медије као једне од основних *Big Data* извора.

У дисертацији су истражени и идентификовани основни недостаци и препреке које је неопходно отклонити за имплементацију и даљи развој *Big Data* концепта у систему званичне статистике у Републици Србији, који се могу сублимирати у следећем:

- Одсуство стратешког и правног основа за коришћење *Big Data*, што између осталог, узрокује непостојање оквира академско-приватног-јавног партнерства.
- Недовољно отварање података (имплементација *Open data* у већој мери),
- Непостојање методолошког оквира,
- Недостатак стручног кадра („стручњаци и научници за податке“) и
- Недовољно развијена ИКТ инфраструктура.

У дисертацији су такође анализирани главни проблеми у области статистичких истраживања и предложено је методолошко решење за имплементацију *Big Data* концепта у домену званичне статистике. Тежиште рада је на дефинисању основних методолошких корака који подразумевају идентификацију *Big Data* ресурса, њихову типологију, идентификацију и услове коришћења.

Кроз анализу и преглед постојећих пилот-студија и решења, базираних у области примене *Big Data* ресурса, указано је на тренутно стање у области званичне статистике на међународном нивоу. Потенцијална имплементација предложеног методолошког поступка, базираног на имплементацији *Big Data* модела, приказана је на моделу статистике одрживог развоја у Републици Србији, и то по фазама имплементације, закључно са анализом и евалуацијом основних предуслова. Посебна пажња усмерена је на процену ризика извођења *Big Data* пројекта у домену коришћења података о личности и нарушавања приватности.

Услед комплексности и осетљивости система званичне статистике, који се базира на чврстим методолошким упориштима, датим кроз међународне препоруке а мереним за свако спроведено истраживање, кроз извештаје о квалитету њиховог спровођења, до сада није било много радова који су са методолошког аспекта истраживали ову област. У том смислу материја изложена у докторској дисертацији представља посебну вредност.

На крају, један од научних доприноса дисертације представља унапређење методологије за мониторинг циљева одрживог развоја. Кроз студију случаја која обухвата Републику Србију извршена је анализа индикатора одрживог развоја, по УН и ЕУ приступу. За изабрани индикатор 1.4.1: „Удео становништва које живи у домаћинствима која имају приступ основним услугама", предложен је методолошки поступак за дефинисање индикатора животног стандарда, применом дефинисаног *Big Data* модела. Имплементација предложеног решења омогућава да се унапреди процес мониторинга одрживог развоја у наредном периоду, пре свега у скраћеној периодици извештавања.

8.2 Стручни доприноси

Основни стручни доприноси дисертације се огледају у могућности примене предложене методологије у различитим областима које покрива званична статистика:

- националани рачуни
- економске статистике
- друштвене статистике.

Осим тога, предложена методологија истраживања се може користити и у другим областима, као што су: медицинска истраживања, истраживања јавног мњења, маркетиншка истраживања, истраживања друштвених медија и сл.

Анализа дефинисаних предусловима за развој методологије која омогућава имплементацију *Big Data* модела процеса статистичког истраживања у Републици Србији указује да је већина ових предуслова делимично испуњено.

За једну област *Big Data* – онлајн истраживања, у Републици Србији, извршена је анализа потенцијала примене истраживања друштвених заједница и интернет истраживања уопште. Указано је на потенцијал, ограничења и очекивани правац развоја методологије у области, правећи аналогију са традиционалним телефонским истраживањима.

8.3 Друштвени допринос

Са становишта друштвене корисности, резултати истраживања могу имати вишеструке импликације:

- Резултати истраживања помоћи ће да се анализира проблематика даљег унапређења техника и метода статистичких истраживања у домену званичне статистике.
- Резултати истраживања омогућавају да се анализира потенцијал примене *Big Data* концепта на нивоу целе државне управе и целокупне администрације у Републици Србији.

- Резултати истраживања указују на неопходне предуслове, пре свега законодавне и стратешке, који морају бити испуњени за потпуну имплементацију предложеног концепта.
- Резултати истраживања доприносе да се прецизније одреде потребни временски, технолошки, финансијски и људски ресурси, неопходни за увођење истраживања базираног на *Big Data* концепту у програм и план званичне статистике.
- Резултате истраживања, пре свега предложену методологију, могу користити и друге институције, привредна друштва и појединци који се баве статистичким истраживањима, како би успешно имплементирале истраживања базирана на *Big Data* концепту.

9 Будућа истраживања

И даље су отворена многа питања које се налазе пред доносиоцима одлука када је реч о имплементацији *Big Data* у званичној статистици:

Везано за технологије:

- Да ли су *Big Data* прекретница за прелазак са одржавање базе података које су у власништву, на обраду у *Cloud-y*?
- Ако статистички завод жели да користи *Big Data* да ли мора да користи *Cloud* технологије, *Elastic Map Reduction, Hadoop* и *Big Data Analytics* алате развијене у комерцијалне сврхе или може да се окрене сопственим решењима или бесплатним академским алатима?
- Колико је *Cloud* окружење безбедно за складиштење високо поверљивих података?
- У којој мери ће *Cloud* технологија бити исплатива за модернизацију статистичких производа и начина обраде?

Везано за методологију:

- Које су основна методолошка начела на којима се заснива *Big Data*?
- По чему са обрада и анализа *Big Data* разликује од обраде и анализе иначе великих база података које се користе у званичној статистици?
- Како званична статистичка заједница може интегрисати *Big Data* са подацима из пописа, подацима из анкета или других административних података?

Везано за организацију:

- Како да се уведу промене у оквиру званичне статистичке заједнице, на глобалном, регионалном и националном нивоу, како би се преиспитали и иновирали статистички производи и процеси у производњи званичне статистике са појавом *Big Data*?
- У којој мери се знања и вештине које су неопходни за обраду и анализу *Big Data* разликују од знања и вештине неопходних за обраду и анализу података из традиционалних извора? Да ли је могуће преквалификовати

кадар у званичним статистикама како би били обучени да користе *Big Data* или се мора запослити нови стручни кадар (научници/стручњаци за податке) или се ова активност може поверити некој приватној компанији ван куће?

Везано за коришћење:

- Како, где и када можемо искористити *Big Data* за подршку доносиоцима одлука и јавних политика?
- Које ће бити последице (позитивне и негативне) коришћења *Big Data* за подршку доносиоцима одлука и јавних политика?
- Да ли су статистичке организације у земљама у развоју у стању да прате рад модерних организација када су у питању нови изазови у *Big Data*?

Када је реч о Републици Србији акценат треба ставити на отклањање идентификованих недостатака, које представљају препреку за даљу имплементацију и развој *Big Data* концепта, не само у систему званичне статистике и истраживања, већ генерално, на нивоу шире друштвене заједнице:

1. Дефинисање стратешког и правног основа по угледу на земље које су највише постигле у овој области (у дисертацији су укратко цитирани примери најуспешнијих) је први корак који ће отворити податке и омогућити стварање партнерстава.
2. Оснивање партнерстава (приватно-јавна-академска) која ће омогућити спој науке и праксе и спречити одлив мозгова, како из научно-истраживачких институција у приватни сектор, тако и одлазак у иностранство.
3. Додатно отворити податке – проширити обухват и сетове отворених података (*open data*) што ће омогућити студентима, истраживачима и научним радницима да уче и имплементирају нове технологије на реалним подацима и решавају практичне проблеме.
4. Иновирати образовни систем кроз увођење курикулума које ће покривати *Big Data*, науку о подацима, пословну аналитику, вештачку интелигенцију и сл. на свим нивоима студија, по угледу на најуспешније светске универзитете. То захтева реинжењеринг наставног процеса тако да се акценат ставио на рад са великим количинама података и повезивањем са

приватним компанијама и државним институцијама (реализацијом партнерстава).

5. Направити технолошку (*Big Data*) инфраструктуру која ће омогућити практичну имплементацију *Big Data* пројеката. За почетак, препорука је да се користе готова инфраструктурна решења која су на располагању од стране ЕУ и другим међународних организација. Тако се може користити CEF *Big Data Test* инфраструктура, описана у дисертацији, које егзистира у виртуелном окружењу и омогућава регистрованим корисницима да експериментишу са *Big Data* технологијама стављајући им на располагање и одређене скупове података и аналитичка решења (BDTI, 2018).

Отклањањем ових недостатака у Републици Србији ће се створити услови за пуну имплементацију *Big Data* у систему званичне статистике, статистичким истраживањима уопште (у медицини, маркетингу, јавном мњењу и сл.) и другим областима (економији и друштву уопште). При томе, наука мора да прати праксу и да покуша са формализацијом постигнутих достигнућа у области *Big Data*.

10 Закључак

Резултати приказани у овој докторској дисертацији указују на значај и актуелност предложене теме имајући у виду улогу и значај статистичких истраживања и званичне статистике уопште. Добијени закључци указују *на неопходност што хитнијег увођења Big Data концепта у свакодневну истраживачку праксу у системима званичне статистике*. Такође, може се закључити да су статистичка истраживања, базирана на *Big Data* концептима, отворена за даље унапређење, пре свега у домену научне методологије, пратећи убрзани развоје науке и технологије. С обзиром на актуелност теме и улогу статистичких истраживања и званичне статистике уопште, могућности примене предложеног методолошког концепта су велике.

У анализи студије случаја описан је *Big Data* модел процеса статистичког истраживања базиран на *Big Data* концептима (*Big Data* модел), који се разликује од модела процеса статистичког истраживања заснованог на традиционалним моделима. Приказана је методологија на којој се *Big Data* модел заснива и основни кораци у његовој имплементацији.

Предложени *Big Data* модел је дефинисан да прати стандардизоване кораке статистичког пословног модела (GSBPM), који се примењује у статистичким заводима широм света, конкретно прилагођен је Моделу процеса статистичког истраживања који се користи у РЗС.

Овако дефинисани модел отворен је за примену Дата концепта и представља један од главних доприноса дисертације. Предложени модел не садржи појединачна технолошка решења и алате, већ је базиран на ближем методолошком одређењу. Приказане су и анализиране фазе животног циклуса истраживања *Big Data* модела. Ово решење се базира на имплементацији Контролне листе за процену ризика извођења *Big Data* пројекта у домену коришћења података о личности и нарушавању приватности у свим фазама истраживања.

Производни процес добијања резултата истраживања у традиционалном и *Big Data* моделу се практично паралелно одвија, с тим што традиционалне методе и технике

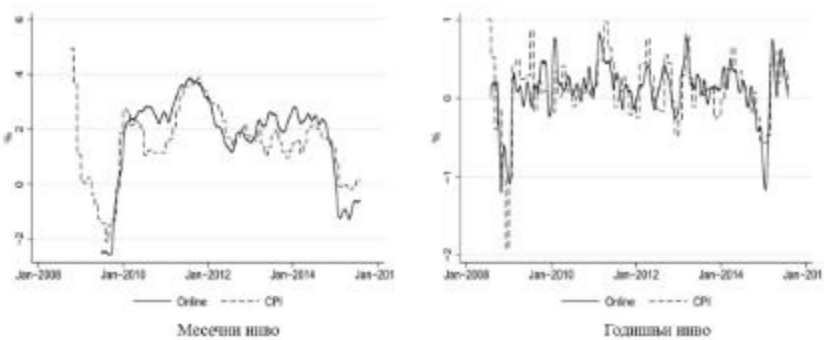
имају предност над *Big Data* истраживањима уколико постижу исте задатке и циљеве који су дефинисани у првом кораку истраживања. Главни разлог је у томе што традиционална статистичка методологија омогућава контролу квалитета добијених резултата, применом стандардизованих образаца, док код *Big Data* истраживања то није случај.

Из тог разлога *Big Data* истраживања имају ограничену употребу која се базира на *cost/benefit* анализи. Основно питање је да ли ће тако добијени резултати потенцијално донети више штете или користи корисницима (стејкохолдерима). За сада, универзални алат за процену ризика коришћења *Big Data* истраживања у процесу доношења одлука, као и стандардизовани извештаји квалитета не постоје, и на корисницима је да одлуче да ли, и у којој мери треба да користе ове резултате. При томе је неопходно да, приликом дисеминације резултата, сви корисници буду информисани о ограничењима и потенцијалној пристрасности до које је можда дошло приликом производње одређених информација – резултата.

Типичан пример који отвара простор за употребу *Big Data* концепта а оставља простор за дилеме је компарација индекса цена у САД: *Big Data* модел (онлајн) и традиционални модел (офлајн) (Cavallo & Rigobon, 2016). *Big Data* модел представља *Daily online price index (DOPI)*, док *CPI* (*Consumer Price Index*) представља офлајн индекс (Слика 10.1).

Слика 10.1: Компарација DOPI и CPI (Cavallo & Rigobon, 2016)

Слични резултати се добијају ако се ови индекси користе као мера инфлације на месечном и годишњем нивоу (Слика 10.2).



Слика 10.2: Компарација мере инфлације на бази DOPI и CPI на месечном и годишњем нивоу (Cavallo & Rigobon, 2016)

У овом примеру за контролу квалитета може се користи оцена статистичког параметра – коефицијента детерминације израчуната у претходном периоду.

Питање је да ли је кориснима прихватљива разлика. Шта им је важније: велика прецизност или период и време извештавања? Време извештавања (дисеминације) код CPI може бити мерено у недељама након референтног периода, до код DOPI извештавање може бити дневно – скоро у реалном времену). Такође, на основу ових графикона може се приметити да је тренд појаве приближно исти, што је понекад важније од апсолутног нивоа мерене појаве.

Резултати изложени у докторској дисертацији потврдили су постављене хипотезе истраживања.

Главна хипотеза, која је тестирана у раду: „Имплементацијом концепта *Big Data,* унапређује се процес статистичких истраживања, отварају се могућности примене нових метода и техника истраживања, олакшава се процес анализе великих количина података и побољшавају укупне перформансе система званичне статистике" је потврђена кроз анализу случајева из праксе који се односе на примену *Big Data* у званичној статистици (поглавље 4 дисертације) и кроз имплементацију студије случаја (поглавље 7).

Посебне хипотезе које се односе на имплементацију *Big Data* концепта за развој корпуса нових метода и техника, као и израду индикатора званичне статистике, као и појединачне хипотезе дефинисане у уводном поглављу, такође су доказане у тексту дисертације на јасан и недвосмислен начин.

11 Референтна литература

Abbasi, A., H. Chen, A. Salem (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. ACM Trans. Info. Syst. 26, 3 (2008)

Abbasi, A., A. Hassan, M. Dhar (2014). Benchmarking Twitter Sentiment Analysis Tools, Conference: 9th Language Resources and Evaluation Conference, Reykjavik, Iceland, доступно на: https://pdfs.semanticscholar.org/d0a5/21c8cc0508f1003f3e1d1fbf49780d9062f7.pdf

Abowd, J. (2017). How Will Statistical Agencies Operate When All Data Are Private?". Journal of Privacy and Confidentiality 7 (3). https://doi.org/10.29012/jpc.v7i3.404.

Акциони план (2018). Акциони план за спровођење иницијативе Партнерство за отворену управу у Републици Србији за период 2018-2020. године, доступно на: https://www.srbija.gov.rs/dokument/45678/strategije.php

Aker A, D., V. Kumar, G. Day (2012). Marketing Research, Wiley Publishing.Inc.

AGIMO (2013). Australian Public Service Big Data Strategy, Department of Finance, Australian Government, Commonwealth of Australia 2013, https://www.finance.gov.au/sites/default/files/Big-Data-Strategy.pdf

Ahas, R., Silm, S., Järv, O., Saluveer E., Tiru, M. (2010). Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones, Journal of Urban Technology, 17(1): 3-27.

Ahas, R., J. Armoogum, S. Esko, M. Ilves, E. Karus, J-L. Madre, O. Nurmi, F. Potier, D. Schmücker, U. Sonntag, M. Tiru, (2014). Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics, Consolidated Report, The European Commission, Eurostat, Publications Office of the Europe an Union, Luxembourg, доступно на: https://ec.europa.eu/eurostat/documents/747990/6225717/MP-Consolidated-report.pdf

Ahn, J-I., Y-Ja Hwang (2013). Production of Official Statistics by Using Big Data, Meeting on the Management of Statistical Information Systems (MSIS 2013) (Paris, France, and Bangkok, Thailand, 23-25 April 2013), http://www.unescap.org/sites/default/files/13-Production%20of%20Official%20Statistics%20by%20Using%20Big%20Data,Kostat.pdf

Alvi, M. H. (2016). A Manual for Selecting Sampling Techniques in Research, University of Karachi, Iqra University, Online at https://mpra.ub.uni-muenchen.de/70218/ MPRA Paper No. 70218, posted 25 March 2016

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, Discoveries magazine-16.07, dostupno na https://www.wired.com/2008/06/pb-theory/

Antoun, C. (2015). Who Are the Internet Users, Mobile Internet Users, and Mobile-Mostly Internet Users?: Demographic Differences across Internet-Use Subgroups in the U.S. In: Toninelli, D, Pinter, R & de Pedraza, P (eds.) Mobile Research Methods: Opportunities and Challenges of Mobile Research Methodologies, Pp. 99–117. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bar.g. License: CC-BY 4.0.

Asur, S., B. A. Huberman (2010). Predicting the Future with Social Media Proceeding WI-IAT '10 Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, Pages 492-499

Bacaer, N (2011). A Short History of Mathematical Population Dynamics, DOI 10.1007/978-0-85729-115-8 2, Springer-Verlag London

Balahur, A., R. Steinberger, M. Kabadjov, V. Zavarella, E. v. d. Goot, M. Halkia, B. Pouliquen, J. Belyaeva (2013). Sentiment Analysis in the News, Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010), pp. 2216-2220. Valletta, Malta, 19-21 May 2010

Balahur, A. (2013). Sentiment Analysis in Social Media Texts, Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media

Analysis, pages 120–128, Association for Computational Linguistics Atlanta, Georgia, 14 June 2013., http://www.aclweb.org/anthology/W13-1617

Batrinca, B., P. C. Treleaven (2015). Social media analytics: a survey of techniques, tools and platforms, AI & Soc (2015) 30:89–116, DOI 10.1007/s00146-014-0549-4

Barcellan, R. (2013). Multipurpose Price Statistics, Ottawa Group, European Commission, Eurostat, http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/8bdac0e73d96c891ca257bb00002fdb4/$FILE/OG%202013%20Barcellan%20-%20Multipurpose%20Price%20Statistics.pdf

BDTI (2018). Big Data Test Infrastructure, The Connecting Europe Facility (CEF), European Commission, European Commission, DG INFORMATICS (DIGIT), https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/Big+Data+Test+Infrastructure

Beręsewicz, M., R.Lehtonen, F. Reis, L. D. Consiglio, M. Karlberg (2018). An overview of methods for treating selectivity in Big Data sources, The European Commission, Eurostat, doi: 10.2785/312232, доступно на: https://ec.europa.eu/eurostat/documents/3888793/9053568/KS-TC-18-004-EN-N.pdf/52940f9e-8e60-4bd6-a1fb-78dc80561943

Beyer, M.A, Thoo, E and Zaidi, E. (2018): Magic Quadrant for Data Integration Tools, Gartner reprint, Gartner, доступно на: https://b2bsalescafe.files.wordpress.com/2018/08/2018-gartner-magic-quadrant-for-data-integration-tools-july-2018.pdf

Beevolve (2017). An Exhaustive Study of Twitter Users Across the World, Beevolve, http://www.beevolve.com/twitter-statistics/

Blazquez, D, J. Domenech (2017). Big Data sources and methods for social and economic analyses, Technological Forecasting and Social Change, Volume 130, May 2018, Pages 99-113, доступно на: https://doi.org/10.1016/j.techfore.2017.07.027

BLS (2018). Employment Projections, Fastest growing occupations, The Bureau of Labor Statistics, https://www.bls.gov/emp/tables/fastest-growing-occupations.htm

Boettcher, I. (2015). Automatic price collection on the Internet (web scraping) New Techniques and Technologies for Statistics (NTTS) Conference 2015, https://ec.europa.eu/eurostat/cros/system/files/Boettcher_Automatic%20price%20collection%20on%20the%20Internet.pdf

Bok, B., D. Caratelli, D. Giannone., A. M. Sbordone, A. Tambalotti (2018). Macroeconomic Nowcasting and Forecasting with Big Data, Annual Review of Economics, Volume 10, https://doi.org/10.1146/annurev-economics-080217-053214

Boiy, E., M-F. Moens (2008). A machine learning approach to sentiment analysis in multilingual Web texts, Information Retrieval Journal, 12: 526. https://doi.org/10.1007/s10791-008-9070-z

Box, G. E. P.; Draper, N. R. (1987). Empirical Model-Building and Response Surfaces, John Wiley & Sons.

Bret, V. (2006). Magic Ink Information Software and the Graphical Interface, Publisher: Worrydream.com 2006, доступно на http://worrydream.com/MagicInk/

Brown, B., Sikes J. & Willmott, P. (2013). Bullish On Digital: McKinsley Global Survey Results, McKinsey&Co

Buelens, B., Daas, P., Burger, J., Puts, M. and van den Brakel, J. (2014). Selectivity of Big Data , Internal report, Statistics Netherlands, Heerlen, The Netherlands.

Cai, L., Y. Zhu (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14: 2, pp. 1-10, DOI: http://dx.doi.org/10.5334/dsj-2015-002

Cambria, E., B. Schuller, B. Liu, H. Wang; C. Havas (2013). Statistical Approaches to Concept-Level Sentiment Analysis, IEEE Intelligent Systems, Volume: 28, Issue: 3, May-June 2013

Cappa, C.; D. Mont, M. Loeb, C. Misunas, J. Madans, T. Comic, de F. Castro (2018). The development and testing of a module on child functioning for identifying children with disabilities on surveys. III: Field testing, Disability and health journal, ISSN: 1876-7583, Vol: 11, Issue: 4, Page: 510-518, Publisher(s): Elsevier BV, Publication Year: 2018, PMID: 30049638, DOI: 10.1016/j.dhjo.2018.06.004 (IF 1.863)

Capponi, L., (2011). Roman Egypt, Bloomsbury Publishing

Carlson, M., Nyquist, H. and Villani, M. (eds) (2010). Official Statistics – Methodology and Applications in Honour of Daniel Thorburn. Available at officialstatistics.wordpress.com.

Cavallo, A, R. Rigobon (2016). The Billion Prices Project: Using Online Prices for Measurement and Research. Journal of Economic Perspectives 30.2 (2016): 151–178.

Cavanillas, J. M., 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

Cavallo, A. (2012). Online and official price indexes: Measuring Argentina's inflation. Journal of Monetary Economics, доступно на: http://dx.doi.org/10.1016/j.jmoneco.2012.10.002

Cavallo, A. and Rigobon, R. (2016). The Billion Prices Project: Using Online Prices for Measurement and Research, JEL-Codes: E3, F3, F4., MIT Sloan. http://bpp.mit.edu/datasets/ .

CBS (2015). A first for Statistics Netherlands: launching statistics based on Big Data, Statistics Netherlands, https://www.cbs.nl/NR/rdonlyres/4E3C7500-03EB-4C54-8A0A

CBS (2016). CBS launching Center for Big Data Statistics, Statistics Netherlands (CBS), https://www.cbs.nl/en-gb/our-services/innovation/nieuwsberichten/big-data/cbs-launching-center-for-big-data-statistics

Cervera, J. L. Ed. (2014). Big Data in official statistics, ESS Big Data Event Rome2014 – Technical Event Report,

https://ec.europa.eu/eurostat/cros/system/files/Big%20Data%20Event%202014%20-%20Technical%20Final%20Report%20-finalV01_0.pdf

Chamlertwat, W., P. Bhattarakosol, T. Rungkasiri, C. Haruechaiyasak (2012). Discovering Consumer Insight from Twitter via Sentiment Analysis, Journal of Universal Computer Science, vol. 18, no. 8 (2012), 973-992, http://jucs.org/jucs_18_8/discovering_consumer_insight_from/jucs_18_08_0973_0992_chamlertwat.pdf

Chaudhuri, A., H. Stenger (2005). Survey Sampling: Theory and Methods, Second Edition, CRC Press Textbook, ISBN 9780824757540

Cheng, J. H-W. (2014). Big Data for Development in China, UNDP China Working Paper, http://www.undp.org/content/dam/china/docs/Publications/UNDP-CH-UNDP%20Working%20Paper_Big%20Data%20for%20Development%20in%20China_Nov%202014.pdf

Chizawsky LL, C.A, Estabrooks, AE Sales (2011). The feasibility of Web-based surveys as a data collection tool: a process evaluation. Appl Nurs Res 2011

Choi, H., H.R. Varian (2009). Predicting the present with Google Trends, Google Research Blog доступно на: http://static.googleusercontent.com/media/www.google.com/fr//googleblogs/pdfs/google_predicting_the_present.pdf

Cox, M., D. Ellsworth (1997). Application-controlled demand paging for out-of-core visualization, Proceeding VIS '97 Proceedings of the 8th conference on Visualization '97, Pages 235-ff. Phoenix, Arizona

Coy, P. (2017). How to Tame Google, Facebook, Amazon, and Apple, Bloomberg, доступно на: https://www.bloomberg.com/news/features/2017-11-29/how-to-tame-google-facebook-amazon-and-apple

Čomić, T (2016). Challenges and experiences in reporting on SDGs on people with disabilities - Serbian Perspective, ESA/STAT/AC.324/7, https://unstats.un.org/sdgs/files/meetings/sdg-seminar-seoul-

2016/7_Challenges_and_Experiences_in_reporting_on_SDGs_on_people_with_disabilities_Serbian_perspective.pdf

Čomić, T., A. Đoković, D. Vukmirović (2017). Potencijal primene Big Data koncepta u domenu zvanične statistike, INFOTEH-JAHORINA Vol. 16, March 2017.

Čomić, T. (2018). Inclusion of production for own consumption in the household disposable income concept: Impact on the income distribution and on key EU income-based indicators, Net-SILC3 International Conference, Comparative EU Statistics on Income and Living Conditions, Athens, 19-20 April 2018

Daas, P.J.H., Puts, M.J., Buelens, B. and van den Hurk, P.A.M. (2013). "Big Data and official statistics", Paper for the 2013 "New techniques and technologies for statistics" conference, Brussels, Belgium

Daas, P. J. H., Puts, M., Tenneks, M. and Priem, A. (2014a). 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: http://www.pietdaas.nl/beta/pubs/pubs/07a2_daas_puts_tennekes_priem_e.pdf (страница посећена 6.12.2016).

Daas, P.J.H., Puts, M.J. (2014). Social media sentiment and consumer confidence, Statistics paper series no 5 / september 2014, European Central Bank, https://www.ecb.europa.eu/pub/pdf/scpsps/ecbsp5.en.pdf

Daas, P.J.H., Puts, M.J., Buelens, B. and van den Hurk, P.A.M. (2015). Big Data as a Source for Official Statistics. Journal of Official Statistics 31(2), pp. 249-262. http://dx.doi.org/10.1515/JOS-2015-0016

Damin, D., C. Jinjing National Bureau of Statistics (2014). Big Data and Official Statistics in China. Working Paper. For Meeting on the Management of Statistical Information Systems (MSIS 2014), http://www.unescap.org/sites/default/files/1-Big%20Data%20and%20Official%20Statistics%20in%20China.pdf

Derbeko, P., Dolev, S. E. Gudes (2016). Concise Essence-Preserving Big Data Representation, 2016 IEEE International Conference on Big Data (Big Data), Washington D.C., USA

Desai, A., B. S-S. Nuño (2015). Korea shows how to use Big Data for development, The World Bank blog, http://blogs.worldbank.org/voices/korea-shows-how-use-big-data-development

DESI (2018). The Digital Economy and Society Index, European Commission, Digital Single Market, https://ec.europa.eu/digital-single-market/en/desi

Deetjen, U., E. T. Meyer, E., R. Schroeder (2015). Big Data for Advancing Dementia Research: An Evaluation of Data Sharing Practices in Research on Age-related Neurodegenerative Diseases", OECD Digital Economy Papers, No. 246, OECD Publishing, Paris.

DevInfo (2018). DevInfo Србија, http://devinfo.stat.gov.rs. www.devinfo.org

Diebold, F. X. (2018). The Origin(s) and Development of "Big Data ": The Phenomenon, the Term, and the Discipline, Unpublished manuscript, Dep. Econ., Univ. Penn., Philadelphia, JEL codes: C81, C82 https://www.sas.upenn.edu/~fdiebold/papers/paper112/Diebold_Big_Data.pdf

Drovandi, C. C., Holmes, C., McGree, J. M., Mengersen, K., Richardson, S., & Ryan, E. G. (2017). Principles of Experimental Design for Big Data Analysis. Statistical science : a review journal of the Institute of Mathematical Statistics, 32(3), 385-404.

Doyukai, K. (2011). Vision of Japan 2020, Japan Association of Corporate Executives, https://www.doyukai.or.jp/en/policyproposals/2010/pdf/110111a.pdf

Dumbill, E. (2013). Making Sense of Big Data . Big Data . Vol. 1. Iss. 1.

Ernst & Young (2016). Audio analytics: new opportunities in litigation and investigation, Ernst & Young LLP., доступно на: https://www.ey.com/Publication/vwLUAssets/ey-audio-analytics/$FILE/ey-audio-analytics.pdf

ESS (2015). ESS handbook for quality reports, Eurostat, European Union, доступно на: https://ec.europa.eu/eurostat/documents/3859598/6651706/KS-GQ-15-003-EN-N.pdf

ESSnet (2016). List of available Big Data sources in the domain(s), ESSnet Big Data Specific Grant Agreement No 1 (SGA-1), Work Package 7, ESSnet Big Data Project

Eurofound (2013). Third European Quality of Life Survey – Quality of life in Europe: Subjective well-being, Publications, Office of the European Union, Luxembourg, доступно на: https://www.researchgate.net/profile/Saamah_Abdallah/publication/312601809_Quality_of_Life_in_Europe_Subjective_Well-being/links/588609274585150dde4a82cd/Quality-of-Life-in-Europe-Subjective-Well-being.pdf

Eurofound (2017). European Quality of Life Survey 2016: Quality of life, quality of public services, and quality of society, Publications Office of the European Union, Luxembourg, доступно на: https://www.eurofound.europa.eu/sites/default/files/ef_publication/field_ef_document/ef1733en.pdf

European Commission (2017). New European Interoperability Framework, Publications Office of the European Union, Luxembourg, ISBN 978-92-79-63756-8 doi:10.2799/7868, https://ec.europa.eu/isa2/sites/isa/files/eif_brochure_final.pdf

Eurostat (2012). Consumer prices research, The European Commission, Eurostat, Publications Office of the Europe an Union, Luxembourg, доступно на: https://ec.europa.eu/eurostat/documents/272892/272992/Consumer_prices_research_2012.pdf/84d7c2f9-59fe-446d-9377-12632686377b

Eurostat (2013). Scheveningen Memorandum on "Big Data and Official Statistics" adopted by the ESSC, The European Commission, Eurostat, https://ec.europa.eu/eurostat/documents/42577/43315/Scheveningen-memorandum-27-09-13

Eurostat (2015). ESS handbook for quality reports, The European Commission, Eurostat, Publications Office of the Europe an Union, Luxembourg, доступно на: https://ec.europa.eu/eurostat/documents/3859598/6651706/KS-GQ-15-003-EN-N.pdf

Eurostat (2017). The European Commission, Eurostat, https://ec.europa.eu/eurostat

Eurostat (2017a). The European Commission, Eurostat European statistics Code of Practice — revised edition 2017, https://ec.europa.eu/eurostat/web/products-catalogues/-/KS-02-18-142

Eurostat (2018). Quality of life indicators - measuring quality of life, The European Commission, Eurostat online publication, https://ec.europa.eu/eurostat/statisticsexplained/index.php?title=Quality_of_life_indicators_-_measuring_quality_of_life#B1_dimensions_of_quality_of_life

Eurostat (2018a). ICT usage in households and by individuals (isoc_i), The European Commission, Eurostat, Publications Office of the Europe an Union, Luxembourg, доступно на: https://ec.europa.eu/eurostat/cache/metadata/fr/isoc_i_esms.htm

Eurostat (2018b). Sustainable development in the European Union, Monitoring report on progress towards the SDGS in an EU context, 2018 edition, The European Commission, Eurostat, Publications Office of the Europe an Union, Luxembourg, доступно на: https://ec.europa.eu/eurostat/documents/3217494/9237449/KS-01-18-656-EN-N.pdf/2b2a096b-3bd6-4939-8ef3-11cfc14b9329

EU SILC (2018). EU statistics on income and living conditions (EU-SILC) methodology, The European Commission, Eurostat, https://ec.europa.eu/eurostat/statistics-explained/index.php/EU_statistics_on_income_and_living_conditions_(EU-SILC)_methodology

Evans J.R., A., Mathur (2005). The value of online surveys, New York, USA, Internet Research, Vol. 15 No. 2,

Fan, W., M. D. Gordon (2014). The power of social media analytics, Communications of the ACM, Volume 57 Issue 6, June 2014, Pages 74-81

Fan, J., F. Han, H. Liu (2014). Challenges of Big Data analysis, National Science Review, Volume 1, Issue 2, 1 June 2014, Pages 293–314, https://doi.org/10.1093/nsr/nwt032

Fisher, R.A. (2006). History of Statistics, JOC/EFR Marcx. Доступно на: http://www-history.mcs.st-andrews.ac.uk/Extras/Fisher_Statistics_History.html

Forrester (2016). Forrester's TechRadar, Big Data, Q1 '16, Forrester Research

Fox, N.C., R.C. Petersen (2013). The G8 Dementia Research Summit—a starter for eight?, The Lancet, Volume 382, Issue 9909, P1968-1969, DECEMBER 14, 2013, DOI:https://doi.org/10.1016/S0140-6736(13)62426-5

Galonja, A, Ž. Šunderić (2017). Praćenje socijalne uključenosti u Republici Srbiji – Indikatori kvaliteta života, Tim za socijalno uključivanje i smanjenje siromaštva, Vlada Republike Srbije, доступно на: http://socijalnoukljucivanje.gov.rs/wp-content/uploads/2017/10/Pracenje_socijalne_ukljucenosti_u_Republici_Srbiji_trece_dopunjeno_izdanje_Indikatori_kvaliteta_zivota.pdf

Gandomi, A. H, Murtaza (2015). Beyond the hype: Big Data concepts, methods, and analytics, International Journal of Information Management Volume 35, Issue 2, April 2015, Pages 137-144

Gandomi, A., M. Haider (2015). Beyond the hype: Big Data concepts, methods, and analytics, International Journal of Information Management, Volume 35, Issue 2, April 2015, Pages 137-144, https://doi.org/10.1016/j.ijinfomgt.2014.10.007

GDPR (2016). General Data Protection Regulation, REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL Of 27 April 2016

GSBPM (2013). Generic Statistical Business Process Model, Version 5.0, December 2013, UNECE, https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0

Godbole, N., M. Srinivasaiah, S. Skiena (2007). Large-Scale Sentiment Analysis for News and Blogs, ICWSM'2007 Boulder, Colorado, USA, http://www.uvm.edu/pdodds/files/papers/others/2007/godbole2007a.pdf

Goes, P. B. (2014). Big Data and IS Research, MIS Quarterly Vol. 38 No.3/ September 2014

Google (2016). Google trends, https://www.google.rs/trends (страница посећена 1.10.2016)

Groves, R. M., L. Nicholls (1986). The Status of Computer-Assisted Telephone Interviewing: Part II - Data Quality, Journal of Official Statistics; Stockholm Vol. 2, Iss. 2.

Gupta, S. (2018). Sentiment Analysis: Concept, Analysis, and Applications, Tutorial, DZone community, AI Zone, https://dzone.com/articles/sentiment-analysis-concept-analysis-and-application

GWG (2018). Big Data Project Inventory, UN Global Working Group, United Nations Statistics Division, https://unstats.un.org/bigdata/inventory/

Hackl, P. (2016). Big Data : What can official statistics expect?, Statistical Journal of the IAOS 32 (2016) 43–52 43, DOI 10.3233/SJI-160965, IOS Press

Heidemann, J., M. Klier, F. Probst (2012). Online social networks: A survey of a global phenomenon, Computer Networks, 56 (18), pp. 3866-3878

Hajirahimova, M. S., A. S. Aliyeva (2015). Big Data strategies of the world countries Institute of Information Technology of ANAS Baku (2015) Conference Paper, Conference: Национальный Суперкомпь-ютерный Форум (НСКФ-2015), Россия, Переславль-Залесский, 24-27 ноябрь, https://www.researchgate.net/publication/306097650_Big_Data_strategies_of_the_world_countries

Hakeem A. et al. (2012). Video Analytics for Business Intelligence. In: Shan C., Porikli F., Xiang T., Gong S. (eds) Video Analytics for Business Intelligence. Studies in Computational Intelligence, vol 409. Springer, Berlin, Heidelberg

Hammer, C. L, D. C. Kostroch, G. Quirós, and STA Internal Group (2017). Big Data: Potential, Challenges, and Statistical Implication, IMF - Statistics Department,

JEL Classification Numbers: E0, Y2, Z0,
https://www.imf.org/~/media/Files/Publications/SDN/2017/sdn1706-bigdata.ashx

Ханић Х. (2003). Истраживање тржишта и маркетинг информациони систем, Економски факултет Београд

Hastie T., Tibshirani R., J. Friedman (2013). The Elements of Statistical Learning: Data Mining, Inference and Prediction Springer Series in Statistics (3rd), Springer

Hey, T. (2008). The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research, Redmond, Washington доступно на: https://www.fh-potsdam.de/fileadmin/user_upload/fb-informationswissenschaften/bilder/forschung/tagung/isi_2010/isi_programm/TonyHey_-__eScience_Potsdam__Mar2010____complete_.pdf

Hey, T., S. Tansley, K. Tolle (2009). Jim Gray on escience: a transformed scientific method, Emergence of a Fourth Research Paradigm, The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research, Redmond, Washington, ISBN978-0982544204. доступно на:
https://www.immagic.com/eLibrary/ARCHIVES/EBOOKS/M091000H.pdf

Holbrook, A. L., J. A. Krosnick, A., Pfent (2008). The Causes and Consequences of Response Rates in Surveys by the News Media and Government Contractor Survey Research Firms, Advances in Telephone Survey Methodology, Edited by James M. Lepkowski, Clyde Tucker, J. Michael Brick, Edith de Leeuw, Lilli Japec, Paul J. Lavrakas, Michael W. Link, and Roberta L. Sangster, John Wiley & Sons, Inc., доступно на: https://pprg.stanford.edu/wp-content/uploads/2007-TSMII-chapter-proof.pdf

Hurwitz, J.S.,Nugent, A. (2013). Big Data for Dummies, John Wiley & Sons, New Jersey

Hutto, C. J., E. Gilbert (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text, Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media,

https://www.researchgate.net/publication/275828927_VADER_A_Parsimonious_Rule-based_Model_for_Sentiment_Analysis_of_Social_Media_Text

IAP (2017). Supporting the Sustainable Development Goals, A Guide for Merit-Based Academies, InterAcademy Partnership 2017, https://royalsociety.org/~/media/about-us/international/supporting-sustainable-development-goals-iap.pdf?la=en-GB

IBM (2018). Big Data analytics, IBM, https://www.ibm.com/analytics/hadoop/big-data-analytics

IAEG-SDGs (2018). Tier Classification for Global SDG Indicators, Inter-agency and Expert Group on Sustainable Development Goals, https://unstats.un.org/sdgs/iaeg-sdgs/tier-classification/

IDS (2015). Inicijativa Digitalna Srbija - Digitalni manifest, доступно на: https://www.dsi.rs/wp-content/uploads/2018/04/Manifest-SRB.pdf

ILOSTAT (2018). International Labour Organization Organization, Statistics, https://www.ilo.org/ilostat

Интернет 1: https://www.google.rs/trends, (Страница посећена 11.12.2016)

Интернет 2: https://dumps.wikimedia.org/other/pagecounts-ez/, (Страница посећена 11.12.2017)

Интернет 3: Портал отворених података, https://data.gov.rs/sr/discover/ (Страница посећена 21.8.2018)

Интернет 4: Hal Varian and the sexy profession, https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1740-9713.2011.00476.x (Страница посећена 11.11.2017)

Интернет 5: https://qz.com/1296930/hal-varian-googles-chief-economist-thinks-the-world-needs-more-data-scientists/, (Страница посећена 11.08.2018)

Интернет 6: Top 53 Big Data platforms and Big Data analytics software, https://www.predictiveanalyticstoday.com/bigdata-platforms-bigdata-analytics-software/ (Страница посећена 01.07.2018)

Интернет 7: The 8 Best Data Visualisation Tools In 2017, https://www.bernardmarr.com/default.asp?contentID=1137, (Страница посећена 13.04.2018)

Jin, X ., B.W. Wah, X., Cheng, Y. Wang (2015). Significance and Challenges of Big Data Research, Big Data Research, Volume 2, Issue 2, June 2015, Pages 59-6

Jug, M., Vaccari, C., Virgillito, A. (2016). A Shared Computation Environment for International Cooperation on Big Data, доступно на: https://ec.europa.eu/eurostat/cros/system/files/Jug-etal_NTTS15-Sandbox-Final.pdf

Kalil, T. (2012). Big Data is a big deal, the White Hous, Blog, https://obamawhitehouse.archives.gov/blog/2012/03/29/big-data-big-deal; https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf

Kaliterna, LJ., J. Burusic, M. Tadić (2012). Indikatori kvalitete življenja/Quality of life indicators, In book: Psihologija u zaštiti mentalnog zdravlja, Chapter: Indikatori kvalitete življenja, Psihologija u zaštiti mentalnog zdravlja, Publisher: Zavod za javno zdravstvo „Sveti Rok" Virovitičko-podravske županije, Editors: Božičević, Viktor; Brlas, Siniša ; Gulin, Marina, pp.437-444

Kaur, A., C. Deepti (2016). Comparison of Text Mining Tools, 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Page(s):186-192 https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7784950

Kish, L. (1965) Survey Sampling, Wiley: New York.

Kissinger, P., J. Rice, T. Farley, S. Trim, K. Jewitt, V. Margavio, D.H. Martin (1999). Application of Computer-assisted Interviews to Sexual Behavior Research, American

Journal of Epidemiology, Volume 149, Issue 10, 15 May 1999, Pages 950–954, https://doi.org/10.1093/oxfordjournals.aje.a009739

Koulopoulos, T. (2018). Alphabet, Apple, Microsoft, And Facebook Are Monopolies: So What? Inc., доступно на: https://www.inc.com/thomas-koulopoulos/its-time-to-reevaluate-if-alphabet-apple-microsoft-facebook-are-monopolies.html

Kumar, V, D. A. Aaker, G.S. Day (2013). Essentials of Marketing Research, John Wiley & Sons Inc., New York

Laney, D. (2001). Application delivery strategies, Meta Group Inc. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lazer, D. M., R. Kennedy, G. King, A. Vespignani (2014). The Parable of Google Flu: Traps in Big Data Analysis', Science, 343(6176), pp. 1203–1205, DOI: 10.1126/science.1248506. DOI: 10.1126/science.1248506

Lavrakas, P. J. (1987). Applied social research methods series, Vol. 7. Telephone survey methods: Sampling, selection, and supervision. Thousand Oaks, CA, US: Sage Publications, Inc.

Letouzé, E., Jütting, J. (2015). Official Statistics, Big Data and Human Development, The Data-Pop Alliance, Harvard Humanitarian Initiative, the MIT Media Lab, and the Overseas Development Institute

LSMS (2008). Studija o životnom standardu Srbija 2002 – 2007, Urednici: Dragan Vukmirović, Rachel Smith Govoni, Republički zavod za statistiku. (2008), доступно на: http://socijalnoukljucivanje.gov.rs/wp-content/uploads/2014/07/Studija-o-zivotnom-standardu-Srbija-2002-%E2%80%93-2007.pdf

Japan (2013). Japan: Big Data for a bigger economy, British Embassy Tokyo, https://opentoexport.com/article/japan-big-data-for-a-bigger-economy/

Kitchin, R. (2014) Big Data, new epistemologies and paradigm shifts, Big Data & Society, April–June 2014: 1–12, DOI: 10.1177/2053951714528481, bds.sagepub.com

Korea (2014). Republic of Korea National Action Plan on Open Government Partnership, https://www.opengovpartnership.org/sites/default/files/140624_OGP_Action_Plan_Republic_of_Korea

Labrinidis, A. H.V. Jagadish (2012). Challenges and opportunities with Big Data, Proceedings of the VLDB Endowment, 5 (12) (2012), https://cra.org/ccc/wpcontent/uploads/sites/2/2015/05/bigdatawhitepaper.pdf

Lankow, J., J. Ritchie, R. Crooks (2012). Infographics: The Power of Visual Storytelling, John Wiley & Sons, Inc. New Jersey

Lavrakas, P. J. (2008). Encyclopedia of Survey Research Methods, SAGE publications, DOI: http://dx.doi.org/10.4135/9781412963947.n469

Li, N., D. D. Wu (2013). Using text mining and sentiment analysis for online forums hotspot detection and forecast, Decision Support Systems Volume 48, Issue 2, January 2010, Pages 354-368, https://doi.org/10.1016/j.dss.2009.09.003

Leonelli, S. (2014). What difference does quantity make? On the epistemology of Big Data in biology. SAGE, Big Data &Society, 2014 1: 1–11, доступно на: https://ore.exeter.ac.uk/repository/bitstream/handle/10871/16163/BigData%26Society_Final_April2014.pdf?sequence=2&isAllowed=y

Lyko, K. Nitzschke, M., Ngonga Ngomo, A-C (2016). Big Data Acquisition, chapter 4 in New Horizons for a Data-Driven Economy, A Roadmap for Usage and Exploitation of Big Data in Europe, Springer Open, доступно на: https://link.springer.com/content/pdf/10.1007%2F978-3-319-21569-3.pdf

Majooni, A., M. Masood, A. Akhavan (2018). An eye-tracking study on the effect of infographic structures on viewer's comprehension and cognitive load, Information Visualization, Vol 17, Issue 3, 2018, https://doi.org/10.1177/1473871617701971

Malle, J-P (2013). Big Data : farewell to Cartesian thinking, Paris Innovation Review, http://parisinnovationreview.com/articles-en/big-data-farewell-to-cartesian-thinking

Marković, N. (2018). Digitalni jaz: Fenomen koji preti, PC-Press, https://pcpress.rs/digitalni-jaz-fenomen-koji-preti/

Mavletova, A., Mick P. Couper, M. P (2016). Device use in Web Surveys: The effect of differential incentives, International Journal of Market Research (IJMR), https://doi.org/10.2501/IJMR-2016-034

Mazzocchi, F. (2015). Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science, EMBO reports 16, 1250-1255, DOI 10.15252/embr.201541001| Published online 10.09.2015, http://embor.embopress.org/content/16/10/1250

McAfee, A, E. Brynjolfsson, (2012). Big Data : The Management revolution, Harvard Business Review, 10, 2012, https://hbr.org/2012/10/big-data-the-management-revolution

McCarthy, W., F. (1989). Evaluation of Computer Assisted Telephone Interviewing as a Survey Methodology by Means of Cost Models and Mathematical Programming, Springer-Verlag, New York

Metadata (2018). Metadata - ESS Reference metadata reporting standards, Eurostat, https://ec.europa.eu/eurostat/data/metadata

Михаиловић, Д. (2004). Методологија научних истраживања, Факултет организационих наука, Београд

Miller, P. V (2017). Is There a Future for Surveys?, Public Opinion Quarterly, Volume 81, Issue S1, 1 April 2017, Pages 205–212, https://doi.org/10.1093/poq/nfx008

Mijatović, B. (2017). Praćenje socijalne uključenosti u Republici Srbiji – Indikatori finansijskog siromaštva i nejednakosti, Tim za socijalno uključivanje i smanjenje siromaštva, Vlada Republike Srbije, http://socijalnoukljucivanje.gov.rs/rs/category/dokumenta/

Milojković, J. (2012). Interoperabilnost u elektronskom poslovanju statističkih sistema, doktorska disertacija, Fakultet organizacionih nauka, Univerzitet u Beogradu

Minelli, M. & Chambers, M. (2013). Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses. Wiley Publishing Inc.

Moreno, A., T. Redondo (2011). Text Analytics: the convergence of Big Data and Artificial Intelligence, International Journal of Interactive Multimedia and Artificial Intelligence, Vol I, Number 4

Neethu, M., S, R. Rajasree (2013). Sentiment analysis in twitter using machine learning techniques, 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), IEEE, DOI: 10.1109/ICCCNT.2013.6726818,

Ndugwa, R. P (2018). Indicator 1.4.1 "Proportion of population living in households with access to basic services", Global Urban Observatory Unit, Research and Capacity development Branch, UN-Habitat, https://unstats.un.org/sdgs/files/meetings/webex-6sep2018/1.%20UN-Habitat%201.4.1%20Presentation.pdf

Njiru, E. (2018). Introducing Indicator 1.4.1, Global Urban Observatory, UN-HABITAT, 26th – 29th March 2018 Bangkok, Thailand, https://www.unescap.org/sites/default/files/Indicator%201.4.1_Basic%20Services.pdf

Nordbotten, S. (2008). The Use of Administrative Data in Official Statistics – Past, Present, and Future – With Special Reference to the Nordic Countries, Keynote speech at the 2008 Conference of the International Association of Official Statistics, October 14, 2008 in Shanghai, доступно на: www.nordbotten.com/articles/Adm_data.pdf

O'Connor, B., Balasubramanyan, R., Routledge, B.R. and Smith, N.A. (2010). "From tweets to polls: linking text sentiment to public opinion time series", Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 23-26 May, Washington DC, USA.

OECD/ITF (2015). Big Data and Transport: Understanding and Assessing Options, OECD and International Transport Forum, Paris

OECD (2013). Exploring Data-Driven Innovation as a New Source of Growth: Mapping the Policy Issues Raised by "Big Data ", OECD Digital Economy Papers, No. 222, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k47zw3fcp43-en

OECD (2016). Skills for a Digital World, Policy brief on the future of work, OECD, https://www.oecd.org/els/emp/Skills-for-a-Digital-World.pdf

OECD (2017). How's life? 2017: Measuring well-being, OECD Publishing, Paris. http://dx.doi.org/10.1787/how_life-2017-en

OECD (2018). OECD Statistics, https://stats.oecd.org/

OECD-UNDP (2018). Monitoring guide For National Co-ordinators from Participating Governments, OECD-UNDP Joint Support Team, Global Partnership for Effective Development Co-operation, http://effectivecooperation.org/pdf/2018_Monitoring_Guide_National_Coordinator.pdf

Ohmori, N., Nakazato, M., and Harata, N. (2005). GPS Mobile Phone–Based Activity Diary Survey. Proc., Eastern Asia Society for Transportation Studies, Vol. 5, pp. 1104–1115

Open Data Strategy (2014). Open Data Strategy 2014-2016, Department for Business, Innovation and Skills, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/330382/bis-14-946-open-data-strategy-2014-2016.pdf

Open Data Strategy (2017). Open Data. Strategy. 2017 – 2022, Open Data Unit, Department of Public Expenditure & Reform, Dublin, Ireland https://data.gov.ie/uploads/page_images/2018-03-07-114306.063816Final-Strategy-online-version1.pdf

Ospina, A.V. (2018). Big Data for resilience storybook: Experiences integrating Big Data into resilience programming. Winnipeg: International Institute for Sustainable Development. Retrieved from www.iisd.org

Pääkkönen, P., D. Pakkala (2015). Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems, Big Data Research 2 (2015) 166–186

Paltoglou, G., M. Thelwall (2012). Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media, ACM Transactions on Intelligent Systems and Technology (TIST) archive Volume 3 Issue 4, September 2012, Article No. 66,

Pang, B., L. Lee (2008). "Opinion Mining and Sentiment Analysis", Foundations and Trends® in Information Retrieval: Vol. 2: No. 1–2, pp 1-135. http://dx.doi.org/10.1561/1500000011

Pak, A., P. Paroubek (2010). Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of LREC 2010 (pp. 1320–1326). Paris: EuropeanLanguageResourceAssociation. RetrievedNovember5, 2010, from http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf

Perduca, V. (2014). Improving prediction of unemployment statistics with Google trends: preliminary experiments, The European Commission, Eurostat, https://ec.europa.eu/eurostat/cros/system/files/UnemploymentFrance_bigdata.pdf

Pentland, A. (2013). The Data Driven Society", Scientific American 309, pp. 78-83

Podoski K. B., Žilionis V. (2008). Scientific and technological advance: globalization and impact of national policies. Global Academic Society Journal: Social Science Insight, Vol. 1, No. 3, pp. 15-24. ISSN 2029-0365.

Poynter, R. (2015). The Utilization of Mobile Technology and Approaches in Commercial Market Research. In: Toninelli, D, Pinter, R & de Pedraza, P (eds.) Mobile Research Methods: Opportunities and Challenges of Mobile Research Methodologies, Pp. 11–20. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bar.b. License: CC-BY 4.0

Pontiki M., D. Galanis, H. Papageorgiou, S. Manandhar, I. Androutsopoulos (2015). SemEval-2015 Task 12: Aspect Based Sentiment Analysis, Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 486–495, Denver, Colorado, June 4-5, 2015. c 2015 Association for Computational Linguistics, http://www.aclweb.org/anthology/S15-2082

Попис (2011). Домаћинства према броју чланова, Републички завод за статистику, Београд, http://pod2.stat.gov.rs/ObjavljenePublikacije/Popis2011/Knjiga10.pdf

Portal (2017). Nacionalni portal za praćenje indikatora, Projekat Jačanje statističkog sistema Srbije poboljšanjem metodologija i standarda kroz primenu, dobre prakse, http://europa.rs/wp-content/uploads/2017/12/Nacionalni-portal-za-pracenje-indikatora-brosura.pdf

Preis, T, Helen Susannah Moat, H. Eugene Stanley and Steven R. Bishop (2012). Quantifying the Advantage of Looking Forward. Scientific Reports. 2: 350. doi:10.1038/srep00350. PMC 3320057, Freely accessible. PMID 22482034.

Puts, M., P. Daas, M. Tennekes (2015). High frequency road sensor data for official statisitics, New Techniques and Technologies for Statistics (NTTS) Conference 2015, https://ec.europa.eu/eurostat/cros/content/high-frequency-road-sensor-data-official-statisitics-marco-puts-piet-daas-martijn-tennekes_en

Puza, B. and O'Neill, T. (2006) .Selection Bias in Binary Data from Volunteer Surveys, The Mathematical Scientist, 31, pp. 85–94.

Radenković, B., Despotović-Zrakić, M., Bogdanović, Z., Barać, D., Labus, A. & Bojović, Ž. (2017). Internet inteligentnih uređaja. Fakultet organizacionih nauka Univerzitet u Beogradu.

Raina, P. (2013). Sentiment Analysis in News Articles Using Sentic Computing, 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW), pp. 959-962. doi: 10.1109/ICDMW.2013.27

Ratel (2017). Pregled tržišta telekomunikacija i poštanskih usluga u republici srbiji u 2017. godini, Regulatorna agencija za elektronske komunikacije i poštanske usluge – RATEL, Beograd, https://www.ratel.rs/uploads/documents/empire_plugin/5bd194d2428d3.pdf

Ray, S. (2017). Essentials of Machine Learning Algorithms (with Python and R Codes), Analytics Vidhya, https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

Reinhard, T. (2010). Complexity Management in Graphical Models, Doctoral thesis, The Faculty of Economics, Business Administration and Information Technology of the University of Zurich, https://files.ifi.uzh.ch/rerg/amadeus/staff/reinhard/diss-final.pdf

Ricciato, F., P. Widhalm, M. Craglia, F. Pantisano (2016). Estimating Population Density Distribution from Network-based Mobile Phone Data, Joint Research Centre, доступно на: https://ec.europa.eu/jrc/en/publication/estimating-population-density-distribution-network-based-mobile-phone-data

Riley, A., D. Smith, D. (2013). Big Changes Will Deliver a Big Future: What Marketing Decision-Makers Expect Their Customer Insight Teams to Deliver, ESOMAR, Best Paper Award, 66th Congress Thing Big, Istanbul

Rubin, D.B. (1976). Inference and Missing Data, Biometrika, 63(3), pp. 581–592.

Russom, P. (2011). Big Data Analytics, Best Practices Report, Fourth quarter 2011, TDWI, доступно на: https://tdwi.org/research/2011/09/best-practices-report-q4-big-data-analytics.aspx?tc=page0&tc=assetpg

Ruths, D., J. Pfeffer (2014). Social Media for Large Studies of Behavior, Science, 346(6213), pp. 1063–1064, DOI: 10.1126/science.346.6213.1063. DOI : 10.1126/science.346.6213.1063

РЗС (2012). Model procesa statističkog istraživanja, Republički zavod za statistiku Srbije, ver. 1.0, http://stat.gov.rs, 2012

РЗС (2017). Zakon o zvaničnoj statistici, Republički zavod za statistiku, Beograd, http://www.stat.gov.rs/media/2271/zakon_o_statistici.pdf

РЗС (2018). Procene-stanovnistva, Republički zavod za statistiku, Beograd, http://www.stat.gov.rs/sr-Latn/oblasti/stanovnistvo/procene-stanovnistva

РЗС (2018a). Upotreba informaciono-komunikacionih tehnologija u Republici Srbiji, 2018, Republički zavod za statistiku, Beograd, ISSN 1820-9084, dopstupno na: http://publikacije.stat.gov.rs/G2018/Pdf/G201816013.pdf

РЗС (2018b). Prihodi u novcu i u naturi i lična potrošnja domaćinstava, III kvartal 2018, Republički zavod za statistiku, Beograd, http://www.stat.gov.rs/sr-Latn/vesti/20181214-prihodi-u-novcu-i-u-naturi-i-licna-potrosnja-domacinstava-iii-kvartal-2018

РЗС і UNICEF (2014). Istraživanje višestrukih pokazatelja položaja žena i dece u Srbiji 2014 і Istraživanje višestrukih pokazatelja položaja žena i dece u romskim naseljima u Srbiji 2014, Konačni izveštaj. Beograd,Srbija: Republički zavod za statistiku і UNICEF, доступно на: https://www.unicef.org/serbia/sites/unicef.org.serbia/files/2018-08/MICS5.pdf

SAS (2018). Big Data - What it is and why it matters, SAS Institute, https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

Samuel, S. J, R.V.P, Koundinya, K. Sashidhar, C.R. Bharathi (2015). A survey on Big Data and its research challenges, ARPN Journal of Engineering and Applied Sciences. VOL. 10, NO. 8, MAY 2015

Schweidel, D. A., W. W. Moe (2014). Listening In on Social Media: A Joint Model of Sentiment and Venue Format Choice. Journal of Marketing Research: August 2014, Vol. 51, No. 4, pp. 387-402. https://doi.org/10.1509/jmr.12.0424

SD in EU (2017). Sustainable development in the European Union, Monitoring report on progress towards the SDGS in an EU Context, Statistical Books, Eurostat, Доступно на: https://ec.europa.eu/eurostat/documents/3217494/8461633/KS-04-17-780-EN-N.pdf/f7694981-6190-46fb-99d6-d092ce04083f, ISBN 978-92-79-72287-5 doi:10.2785/237722

SDG (2016). Sustainable Development Goals. United Nations. https://sustainabledevelopment.un.org/topics/sustainabledevelopmentgoals

Shan, C. F. Porikli, T. Xiang, S. Gong (Eds.) (2012). Video analytics for business intelligence, Springer, Berlin, Heidelberg (2012). pp. 309-354

Shmueli, G., O. R. Koppius (2011). Predictive Analytics in Information Systems ResearchMIS Quarterly Vol. 35, No. 3 (September 2011), pp. 553-572, Management

Information Systems Research Center, University of Minnesota, DOI: 10.2307/23042796 https://www.jstor.org/stable/23042796

Simas, M. G., L. Wilson (2018). Integrating Smartphone Apps into Small Regional Household Travel Surveys, Transportation Research Board, 97th Annual Meeting, Washington DC

Smiciklas, M. (2012). The Power of Infographics: Using Pictures to Communicate and Connect with Your Audience, QUE Publishing, Copyright by Pearson education, inc, доступно на: http://ptgmedia.pearsoncmg.com/images/9780789749499/samplepages/0789749491.pdf

Smith, T.M.F. (1983). On the Validity of Inferences from Non-Random Samples, Journal of the Royal Statistical Society (Series A): Statistics in Society, 146(4)

SNU (2013). Big Data Institute, Seoul National University, http://bdi.snu.ac.kr/eng/

Stat. Neth. (2013a). "Consumer confidence", Statistics Netherlands web page, available at http://www.cbs.nl/enGB/menu/themas/dossiers/conjunctuur/publicaties/conjunctuurbericht/inhoud/ conjunctuurklok/toelichtingen/ck-03.htm

Stat. Neth. (2018). "Consumer confidence", Statistics Netherlands web page, available at https://www.cbs.nl/en-gb/news/2018/51/dutch-consumers-again-less-positive

Stat. Neth. (2013b). "Seven in ten internet users active on social media", Statistics Netherlands web magazine, 4 October, available at http://www.cbs.nl/en-GB/menu/themas/vrije-tijdcultuur/publicaties/artikelen/archief/2013/2013-3907-wm.htm

Stodden, V. (2017). Reproducibility Enhancement Principles, Using Big Data : The Ethics, Dilemmas, and Possibilities for Educational Opportunity AERA Presidential Session; Invited Speaker Session, AERA Annual Meeting, San Antonio

Стратегија (2015). Стратегија развоја електронске управе у Републици Србији за период од 2015-2018. године и Акциони план за спровођење стратегије за период

од 2015-2016. године, Министарство за државну управу и локалну самоуправу, Влада Републике Србије, доступно на: http://www.mduls.gov.rs/doc/Strategija%20razvoja%20eUprave%20sa%20AP%202015-2018.pdf

Tam, S-M, Clarke, F. (2015). Big Data, Official Statistics and Some Initiatives by the Australian Bureau of Statistics, International Statistical Review (2015), 0, 0, 1–13 doi:10.1111/insr.12105

Tam, S-M, Clarke, F. (2015a). Big Data, Statistical Inference and Official Statistics, Research Paper, Australian Bureau of Statistics, http://www.ausstats.abs.gov.au/Ausstats/subscriber.nsf/0/015937BADB90186BCA257E0B000E428A/$File/1351055054_mar%202015.pdf

Thelwall, M. K. Buckley, and G. Paltoglou (2011). Sentiment in Twitter events. J. Amer. Soc. Info. Sci. Technol. 62, 2 (2011), 406–418, http://www.uvm.edu/pdodds/files/papers/others/2011/thelwall2011a.pdf

Tim (2018). Ocena apsolutnog siromaštva u Srbiji u 2017 godini, Tim za socijalno uključivanje i smanjenje siromaštva, Vlada Republike Srbije, доступно на: http://socijalnoukljucivanje.gov.rs/rs/ocena-apsolutnog-siromastva-u-srbiji-u-2017-godini/

Tinto A., F. Bacchini, B. Baldazzi, A. Ferruzza, J. A. Van den Brakel, R.M.A. Willems, N. Rosenski, T. Zimmermann, Z. Andrási, M. Farkas, Z. Fábián (2018). Report on international and national experiences and main insight for policy use of well-being and sustainability framework, Horizon 2020 - Research and Innovation Framework Programme, Work Package 1: Analysis of the frameworks on well-being and sustainability at national and international level for policy making, Deliverable 1.1, www.makswell.eu

TSM (2018). Fermat and Pascal on probability, The Story of Mathematics, доступно на: https://www.york.ac.uk/depts/maths/histstat/pascal.pdf

Tufte, E.R (2001). The Visual Display of Quantitative Information, Graphics, Second Edition, ISBN13: 9780961392147)

Ularu, E. G., F. C. Puican, A. Apostu, M. Velicanu (2012). Perspectives on Big Data and Big Data Analytics, Database Systems Journal vol. III, no. 4/2012

UNDP, UN Global Pulse, (2016). A Guide to Data Innovation for Development: From Idea to Proof of Concept, доступно на: http://www.sustainablesids.org/wp-content/uploads/2018/02/UNDP-UN-Global-Pulse-A-Guide-to-data-innovation-for-development-2016.pdf

UNECE (2013). What Does "Big Data " Mean For Official Statistics? (2013), United Nations Economic Commission for Europe, Conference of European Statisticians, http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170622

UNECE (2014). The Role of Big Data in the Modernisation of Statistical Production, United Nations Economic Commission for Europe, доступно на: https://statswiki.unece.org/display/bigdata/2014+Project

UNECE (2014a). The United Nations Economic Commission for Europe, Experiment report: Job Vacancies, https://statswiki.unece.org/display/bigdata/Experiment+report%3A++Job+Vacancies

UNECE (2014b). A Suggested Framework for the Quality of Big Data, United Nations Economic Commission for Europe, доступно на: https://statswiki.unece.org/.../Big%20Data%20Quality%20Framework%20-%20final-.

UNECE (2015). The United Nations Economic Commission for Europe, Experiment report: Canadian Smart Meter Data, https://statswiki.unece.org/display/bigdata/Experiment+report%3A++Canadian+Smart+Meter+Data#c5a2a378b50f4a84b573ef0684973b78

UNECE (2017). Conference of European Statisticians, Statisticians' Road Map on Statistics for Sustainable Development Goals, First Edition, Steering Group on Statistics for SDGs, The United Nations Economic Commission for Europe, доступно на:

https://www.unece.org/fileadmin/DAM/stats/publications/2017/ECECESSTAT20172.pdf

UNECE (2018). The United Nations Economic Commission for Europe, https://www.unece.org

UNECE (2018a). Big Data Projects, The United Nations Economic Commission for Europe, https://statswiki.unece.org/display/bigdata/Big+Data+Projects

UNSDSN (2015). Indicators and a Monitoring Framework for the Sustainable Development Goals, Launching a data revolution for the SDGs, A report by the Leadership Council of the Sustainable Development Solutions Network, Revised working draft (Version 7), March 20, 2015 http://unsdsn.org/wp-content/uploads/2015/03/150320-SDSN-Indicator-Report.pdf

UN (2017). New Urban Agenda, United Nations, ISBN: 978-92-1-132731-1 доступно на: http://habitat3.org/wp-content/uploads/NUA-English.pdf

UNSTATS (2014). Big Data and modernization of statistical systems, United Nations Statistics Division Report of the Secretary-General, Statistical Commission, Forty-fifth session, 4-7 March 2014

UNSTATS (2015). UN Fundamental Principles of Official Statistics – Implementation guidelines, United Nations Statistics Division, https://unstats.un.org/unsd/dnss/gp/fp-english.pdf

UNSTATS (2017). The United Nations Statistics Division, https://unstats.un.org/home/

UNSTATS (2018). Work Plans for Tier III Indicators, United Nations, Department of Economic and Social Affairs, https://unstats.un.org/sdgs/tierIII-indicators/files/Tier3-01-04-01.pdf, pristup izvoru: 12.10.2018)

Vaccari, C. (2014). Big Data in Official Statistics, PhD Thesis, Università degli studi di camerino, School of advanced studies, https://www.researchgate.net/profile/Carlo_Vaccari2/publication/264052337_Big_Data_in_Official_Statistics_-_PhD_Thesis_in_Computer_Science_-

_University_of_Camerino/links/00b4953ccc5d35fea1000000/Big-Data-in-Official-Statistics-PhD-Thesis-in-Computer-Science-University-of-Camerino.pdf

Vale, S. (2013). Classification of Types of Big Data, Big Data in Official Statistics, UNECE, https://statswiki.unece.org/display/bigdata/Classification+of+Types+of+Big+Data

Varian H.R. (2014). Big Data : new tricks for econometrics, Journal of Economic Perspectives, Volume 28, Number 2. Spring 2014, https://pubs.aeaweb.org/doi/pdf/10.1257/jep.28.2.3

Varian, H. (2018). Artificial Intelligence, Economics, and Industrial Organization, Chapter in forthcoming NBER book, The Economics of Artificial Intelligence: An Agenda, edited by Joshua Gans and Avi Goldfarb. The University of Chicago Press, https://www.nber.org/chapters/c14017.pdf

Vellaipandiyan, S. (2013). Big Data framework for national E-governance plan, Eleventh International Conference on ICT and Knowledge Engineering, Bangkok, https://www.researchgate.net/publication/269406616_Big_data_framework_for_national_E-governance_plan

Вукмировић, Д. (1994). Генератори апликација у истраживању маркетинга-BLAISE систем, Маркетинг, Vol. XXV Број 3, Београд, 1994. (стр. 16 - 20) ISSN 0354-3471 UDK 338.45:629.12

Вукмировић, Д., Б. Раденковић, З. Радојичић (1995). CATI метод у истраживању маркетинга, Simorg '95, Зборник радова, Златибор

Vukmirović, A. (2017). Model infrastrukture za internet marketing istraživanja u elektronskom poslovanju, Doktorska disertacija, Fakulet organizacionih nauka, Beograd

Vuković, N., D. Vukmirović (2004). Statistika, Fakultet organizacionih nauka, Beograd

Wang, C., Chen, M. H., Schifano, E., Wu, J., & Yan, J. (2016). Statistical methods and computing for Big Data. Statistics and its interface, 9(4), 399-414., doi: 10.4310/SII.2016.v9.n4.a1

Watkins, E., Marius, S. (2014). Kant's Philosophy of Science", The Stanford Encyclopedia of Philosophy (Fall 2014 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2014/entries/kant-science/

Watson, E.K., D, W. Firman, A. Heywood, A. C. Hauquitz, I. Ring (1995). Conducting regional health surveys using a computer-assisted telephone interviewing method, Public Health, Volume 19, Issue 5, https://doi.org/10.1111/j.1753-6405.1995.tb00419.x

Yan,T., F. Kreuter, R.Tourangeau (2012). Evaluating Survey Questions: A Comparison of Methods, Journal of Official Statistics, Vol.28, No.4, 2012. pp. 503–529

Yang, Y., W. Duan, Q. Cao (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach, Decision Support Systems, Volume 55, Issue 4, November 2013, Pages 919-926

Zikopoulos, P., Eaton, Ch., deRoos, D., Deutsch, T. & Lapis, G. (2012). Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Dhe State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation The McGraw Hill Companies

Zimbra, D., A. Abbasi, H. Chen (2018). The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation, ACM Trans. Manage. Inf. Syst. 9, 2, Article 5 (August 2018), 29 pages https://doi.org/10.1145/3185045

12 Списак слика

13 Списак табела

14 Основни биографски подаци о кандидату

Тијана Чомић (рођ. Милојевић) рођена је у Крагујевцу 8. септембра 1984. Основну школу и Прву крагујевачку гимназију завршила је у Крагујевцу. Уписала је 2003. године Економски факултет Универзитета у Београду, где је дипломирала 2007. године на смеру Статистика, информатика и квантитативна анализа. На истом факултету завршила је и мастер студије, опција: економетрија. Мастер рад на тему „Одређивање маргиналних ефеката фактора сиромаштва применом модела са квалитативном зависном променљивом" одбранила је 2009. године са оценом 10. Од 2013. уписана је на Факултету организационих наука Универзитета у Београду, као студент докторских студија на смеру Операциона истраживања.

Од 2007. до 2016. године била је запослена у Републичком заводу за статистику на позицији статистичара-аналитичара. Била је руководилац групе за Анкету о приходима и условима живота.

У сарадњи са Бечким институтом радила је као део тима који је истраживао проблем неједнакости и јавних политика. Рад је презентован на неколико међународних скупова организованих од стране GDN. Од стране УНИЦЕФ-а била је ангажована на додатној обради података MICS-а. Тренутно је ангажована на пројекту NetSILC3 који има за циљ унапређење методологије EU-SILC-а, на нивоу Европске уније. Током 2017. године била је ангажована од стране регионалне канцеларије УНИЦЕФ-а, као консултат на пословима упоређивања европских истраживања са MICS истраживањем у циљу употпуњавања доступности индикатор одрживог развоја. Тренутно је од стране регоиналне канцеларије УНИЦЕФ-а ангажована као регионални консултат MICS-а за истраживања на домаћинстцима.

Од децембра 2016. ради као независни експерт у области статистичког праћења индикатора одрживога развоја и животног стандрада, на пројектима које спроводи УНИЦЕФ, Европска комисија, UNDP, итд.

14.1 Списак радова

Током досадашњег рада Тијана Чомић је објавила више радова у земљи и иностранству и учествовала на више међународних и домаћих скупова и конференција.

Радови од националног значаја (домаће конференције - монографије)

1. Jovičić M., **T. Milojević** (2009). Ocena efekata zamrzavanja penzija i zarada u javnom sektoru na siromaštvo, Ekonomska politika srbije u 2010. godini - Ka novom modelu makroekonomske stabilnosti, Redaktori: S. Stamenković i B. Živković. Naučno društvo ekonomista sa Akademijom ekonomskih nauka i Ekonomski fakultet u Beogradu, CIDEF. ISBN 978-86-403-1048-2

2. Stamenković, S., M. Kovačević, B. Živković, V. Vučković, **T. Milojević** (2011). Ekonomska politika srbije u 2011: Početak ili kraj razvojnog modela? Nova strategija razvoja privrede Srbije – Izazovi ekonomske politike u 2011. godini, Redaktori: Jurij Bajec и Miomir Jakšić, Naučno društvo ekonomista Srbije sa Akademijom ekonomskih nauka i Ekonomski fakultet u Beogradu.

Радови од интернационалног значаја (међународне конференције)

1. Jovičić, M., & **Milojević, T**. (2010). Poverty and inequality changes in Serbia as the result of global instability. The 11-th Bi-Annual Conference of the EACES, August 26-28, 2010, ComparingResponses to Global Instability. Tartu, Estonia. ISBN978-9985-4-0613-7.
   http://ec.ut.ee/eaces2010/artiklid/Jovicic%20Milojevic-Poverty%20and%20inequality.pdf

2. **Čomić, T**. (2016): Challenges and experiences in reporting on SDGs on people with disabilities - Serbian Perspective, International Seminar on Data for Sustainable Development Goals: Data Disaggregation, United Nations Statistics Division (UNSD), Statistics Korea (KOSTAT), Seoul, Republic of Korea, 3-4 November 2016, ESA/STAT/AC.324/7

3. **Čomić, T.**, A. Đoković, D. Vukmirović (2017): Potencijal primene Big Data koncepta u domenu zvanične statistike, Infoteh-Jahorina, Vol. 16, March 2017.

4. **Čomić, T.** (2018): Inclusion of production for own consumption in the household disposable income concept: Impact on the income distribution and on key EU income-based indicators, Net-SILC3 International Conference, Comparative EU Statistics on Income and Living Conditions, Athens, 19-20 April 2018

Радови од интернационалног значаја (међународни часопис)

1. Cappa, C.; D. Mont, M. Loeb, C. Misunas, J. Madans, **T. Comic**,; de F. Castro (2018).The development and testing of a module on child functioning for identifying children with disabilities on surveys. III: Field testing, Disability and health journal, ISSN: 1876-7583, Vol: 11, Issue: 4, Page: 510-518, Publisher(s): Elsevier BV, Publication Year: 2018, PMID: 30049638, DOI: 10.1016/j.dhjo.2018.06.004 (IF 1.863; 5-Year Impact Factor: 2.129)

Посебне публикације:

1. Ivančev, O., M Jovičić and **T. Milojević** (2010). Income Inequality and Social Policy in Serbia, The Vienna Institute for International Economic Studies, Balkan Observatory Working Paper No. 86, October 2010

2. **Милојевић, Т.**, Ђ. Вуксановић (2009). Анализа сиромаштва - интеграција избеглих лица, Анализа карактеристика сиромаштва у Србији, Уредник: Јелена Марковић, Влада Републике Србије, Тим потпредседника Владе за имплементацију Стратегије за смањење сиромаштва

Пројекти:

1. IPA 2015 Multi-beneficiary statistical cooperation programme, Non-key expert for the Pilot project 7.2 Sustainable Development Goal (SDGs) indicators in selected countries in Western Balkan, GOPA. (2018 - on-going),

2. (on-going) UNICEF Regional Office, Geneva, regional MICS consultant, in charge for survey methodology. (2017 - on-going),

3. Household economic Survey, Kingdom of Saudi Arabia, consultant for data analysis and report writing, GOPA. (2017 - on-going),

4. UNICEF Kazakhstan, Review and critically assess work done for identification and collection of data sources used for the production of the Statistical Yearbook on children of Kazakhstan; Refine a list of core indicators for monitoring the situation of children in the country; Establish SDG baselines on selected core children-related SDG indicators at the national and disaggregated levels. For disaggregated levels, further analysis of the 2015-16 MICS as well as calculation of confidence intervals may be necessary; Provide support in user-friendly visualization and presentation of data and deliverables, helping to produce web-page on children. (2017).

5. UNICEF Turkmenistan, Consolidation, further analysis of MICS5, disaggregation of the existing data on children and women to serve as SDG baselines and inform the upcoming Situation Analysis of Children and Women in Turkmenistan, preparation of dashboards for regional analysis in Turkmenistan. (2017).

6. Uzbekistan National Nutrition Survey 2017, Technical support in the capacity building of national team on household mapping and listing, Uzbekistan, UNICEF CO. (2017).

7. Conducting baseline assessment for child-related SDG-s in the region; Enhance the potential for convergence between MICS and EU-SILC/LFS; Revision of MICS6 survey instruments for additional indicators; Support RO in revision of CSPs for different countries. (2016-2017).

8. Mapping of the baseline data availability for CEE/CIS Region for children related SDGs and enhance the potential for convergence between MICS and EU-SILC/LFS. (2016).

9. Revision of social inclusion indicators, Social Inclusion and Poverty Reduction Unit, Government of the Republic of Serbia (SIPRU) and UNICEF. (2016-2018).

10. Net-SILC3, Eurostat and LISER. (2016-2018).

11. Strengthening the Serbian Statistical System by upgrading methodologies and standards and by the appliance of good practice, Contract no: 371-594, IPA 2012 National. (2016 - on-going).

12. Field Testing of the UNICEF/WG Module on Child Functioning and Disability, SORS and UNICEF. (2015-2016).

13. Secondary analyses of data from the survey on income and living conditions (SILC), Social Inclusion and Poverty Reduction Unit, Government of the Republic of Serbia (SIPRU). (2015).

14. Additional Analysis of MISC data – Poverty,UNICEF. (2015).

15. MICS5 (Multi Indicator Cluster Survey – Round 5), SORS and UNICEF. (2013-2014).

16. Economic Impact of Social Enterprises, SORS, EC, Grupa 484, SeCons. (2013-2014).

17. Survey on Corruption and Crime affecting the Business Sector in the Western Balkans, SORS and UNODC. (2012-2014).

18. 2012 – 2015 Survey on Income and Living Conditions - Serbia, SORS, EC and WB. Main responsibilities: Coordinator of the Project team and Chief methodologist: Questionnaire development and testing; Methodology development; Organisation of the fieldwork; Calculation of the indicators. (2012-2015).

19. Prati pare, CRTA. (2012).

20. Citizen budget, under USAID Business Enabiling Project in Serbia, BIRN. (2012).

21. Државни динар по становнику, Министраство економије и регионалног развоја. (2012).

22. Business Tendency Survey, SORS and SIDA

23. Fiscal Monitor, Centar za finansije. (2011).

24. MICS4 (Multi Indicator Cluster Survey – Round 4), SORS and UNICEF. (2010-2012).

25. DevInfo Serbia, SORS and UNICEF. (2012).

26. Household survey on experience of corruption and other forms of crime in countries/territories of the Western Balkans, SORS and UNODC. (2010-2011).

27. Inequality and public policy, WIIW. (2008-2010).

28. Innovation Study, World Bank (2010).

Прилог 1.

Изјава о ауторству

Потписана Тијана Чомић

Број индекса 5045/2013

Изјављујем

да је докторска дисертација под насловом

Унапређење званичне статистике применом *Big Data* концепта

- резултат сопственог истраживачког рада,
- да предложена дисертација у целини ни у деловима није била предложена за добијање било које дипломе према студијским програмима других високошколских установа,
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио интелектуалну својину других лица.

Потпис докторанда

У Београду, _____

_____

Прилог 2.

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора: Тијана Чомић

Број индекса: 5045/2013

Студијски програм:

Наслов рада: **Унапређење званичне статистике применом Big Data концепта**

Ментор: Проф. др Зоран Радојичић

Потписана Тијана Чомић

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла за објављивање на порталу **Дигиталног репозиторијума Универзитета у Београду**.

Дозвољавам да се објаве моји лични подаци везани за добијање академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис докторанда

У Београду, _____

_____

Прилог 3.

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Унапређење званичне статистике применом Big Data концепта

која је моје ауторско дело.

Дисертацију са свим прилозима предала сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигитални репозиторијум Универзитета у Београду могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучила.

1. Ауторство

2. Ауторство - некомерцијално

3. Ауторство – некомерцијално – без прераде

4. Ауторство – некомерцијално – делити под истим условима

5. Ауторство – без прераде

6. Ауторство – делити под истим условима

Потпис докторанда

У Београду, _____

_____