СТОПАНСКА АКАДЕМИЯ „Д. А. ЦЕНОВ“

КАТЕДРА „МАТЕМАТИКА И СТАТИСТИКА”

Галя Живкова Статева

ПРОБЛЕМИ ПРИ СТАТИСТИЧЕСКИЯ АНАЛИЗ НА „ГОЛЕМИ ДАННИ“ (BIG DATA)

А В Т О Р Е Ф Е Р А Т

Дисертация за присъждане на образователна и научна степен „доктор“ по докторска програма „Статистика и демография“

Научен ръководител:
Проф. д-р Поля Ангелова

Свищов
2017

Дисертационният труд е в обем от 208 страници, от които 174 страници основен текст. В структурно отношение включва въведение, изложение в две глави, заключение, списък с използвана литература – 148 източника (39 на кирилица и 109 на латиница), 4 приложения и декларация за оригиналност. В основния текст на дисертационния труд са включени 8 таблици и 16 фигури.

Откритото заседание на научното жури за защита на дисертационния труд ще се състои на 20.12.2017 г. от 11.30 часа в Заседателна зала Ректорат на СА „Д. А. Ценов" – Свищов.

Материалите по защитата са на разположение на интересуващите се в офис „Докторантура и академично развитие" на СА „Д. А. Ценов" – Свищов.

I. ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

1. Актуалност на темата

Процесите на глобализация и технологизация във всички сфери на обществения живот и динамично променящите се информационни потоци, поставят нови акценти в социално-икономическото развитие на обществото. Иновациите, основаващи се на данни, разкриват нови възможности за развитие на икономиката и справяне със социалните предизвикателства. Приоритетът за интелигентен растеж чрез изграждане на икономика, основаваща се на знания и иновации в контекста на Стратегия „Европа 2020“ – водеща инициатива „Програма в областта на цифровите технологии за Европа“, изисква да бъдат постигнати устойчиви икономически и социални ползи чрез създаване на истински единен пазар на онлайн съдържание и услуги.

В нашия модерен свят все повече данни се генерират от световната интернет мрежа и се произвеждат от електронни сензори и устройства, които са навсякъде около нас. Обемът на данните, тяхното разнообразие и високата скорост, с която се създават, води до създаване на концепцията „големи данни“ или Big Data[1]. „Големите данни“ са съвършено нови източници за официалната статистика, с характеристики, различни от тези на традиционните източници. Промяната в характера на данните, тяхната наличност, процеса по събиране и разпространение е фундаментална. Тази промяна по същество е промяна на парадигмата на конвенционалното статистическо изследване.

Пред научната общност има много възможности, свързани с изучаването на Big Data, но едновременно с това съществуват някои основни предизвикателства, които трябва да бъдат решени, за да може пълният потенциал на тези данни да се реализира в официалната статистическа практика. Уникалността на „големите данни“ като съвършено нови източници на данни за официалната статистика поражда много въпроси относно тяхното качество, управление и приложение, което определя актуалността на настоящото

[1] В дисертацията терминът „Big Data” се използва едновременно на английски език и възможно най-адекватния български еквивалент „големи данни”, наложил се през последните години в литературата. Английският термин придобива все по-голяма популярност в различни български източници.

изследване. Официалната статистика притежава опита и възможностите да изведе проблемите и начините за използването на Big Data за информационно осигуряване на изцяло „бели полета“ в социално-икономическите изследвания, за тематично допълване на конкретни наблюдения и др. Това обуславя и практическата полезност от разработването на настоящия дисертационен труд.

Авторът на дисертационния труд има възможността и привилегията да участва в международни проекти на Евростат по проблемите на „големите данни“, което го мотивира при избора на темата, и е убеден, че с разработването на настоящото изследване ще допринесе за по-нататъшното развитие на Big Data проекта в българската статистическа практика.

2. Обект и предмет на изследването

Обект на настоящия дисертационен труд са „големите данни“ (Big Data) като възможен източник на статистическа информация.

Предмет на дисертацията са подходите и възможностите при използването на Big Data за информационно осигуряване, изследователски и научни цели при анализа на явленията и процесите в обществото и решаването на обществено значими практически задачи.

3. Цел и задачи на дисертационния труд

Целта на изследването е да се изяснят теоретичните основи и практическите аспекти на „големите данни“ с оглед изграждане на цялостна концепция за възможностите на приложението на Big Data в българската статистическа практика в съответствие с националните особености и реализирането на техния пълен потенциал.

За реализирането на целта се поставят следните **изследователски задачи**:

Първо, да се дефинира обектът на изследване в контекста на статистическия познавателен хоризонт чрез представяне на еволюцията в теоретичните постановки за същността на Big Data.

Второ, чрез обзор на съществуващите в литературата класификации да се обосноват видовете източници на „големи данни“, които могат да се използват в статистическата практика.

Трето, да се анализира промяната на парадигмата за класическото статистическо изследване чрез изясняване на общите черти и различията между официалните източници на информация и Big Data.

Четвърто, да се изяснят и изведат фундаменталните проблеми в отделните етапи на обработката на Big Data.

Пето, да се обосноват възможните подходи за подобряване на качеството в процеса на производство на официални статистически данни от източници на Big Data и обследване на ефекта от тяхното използване.

Шесто, да се направи характеристика на най-често използваните технологии за обработка на „големите данни“.

Седмо, да се представят резултатите от проведено емпирично изследване за прилагане на техниките за използване на „големите данни“ върху съвкупността от единиците на регулярното статистическо изследване „Използване на ИКТ от предприятията“.

4. Изследователска теза

Изследователската теза на автора е, че чрез разработване на цялостна теоретично и емпирично обоснована концепция за приложението на източниците на „големи данни“ и очертаване на пътищата за тяхното комбинирано използване с официалната статистическа информация, ще се подпомогне намирането на адекватно решение на предизвикателствата на информационното общество в условията на глобализиращия се свят.

5. Методология на изследването

Методологията на изследването за постигане на целта и изпълнението на дефинираните задачи в дисертационния труд включва статистическия подход при изучаване на масовите явления, интердисциплинарния подход, индуктивния и дедуктивния метод, метода на анализа и синтеза, сравнителния метод. В

процеса на работа са приложени статистическите методи за извадкови изследвания и проверка на статистически хипотези. Програмната осигуреност на изследването включва както общодостъпни ИТ техники за работа с „големи данни“, така и авторски разработен софтуер за реализацията на емпиричното изследване.

6. Ограничителни условия

В настоящето изследване са въведени следните **ограничителни условия**:

- Изследователският обхват на изследването се фокусира върху „големите данни“ преди всичко като източник на информация за официалната статистика, тъй като проблематиката, свързана с Big Data, е водеща тенденция в развитието на информационните технологии, маркетинговите проучвания, иновациите в индустрията и други области на науката и практиката.
- Теоретичните постановки, на които се базира дисертационният труд, в преобладаващата си част са от чуждоезикови източници, поради обстоятелството, че концепцията Big Data в България тепърва прохожда и българските автори, работещи в тази област, аналогично се позовават на оригиналните произведения.
- Постигнатите резултати, направените изводи и заключения относно възможността за приложение на „големите данни“ в официалната статистическа практика се базират на проведеното емпирично изследване, тъй като то е първото и единствено за сега, осъществено в рамките на Националния статистически институт.

7. Апробация на изследването

Дисертационният труд е обсъждан на заседания на катедра „Математика и статистика” при СА „Д. А. Ценов” – Свищов. Отделни части от дисертационния труд са представени на национални и международни научни форуми и са публикувани в специализирани научни издания.

II. СТРУКТУРА И СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

Дисертационният труд е в обем от 208 страници, от които 174 страници основен текст. В структурно отношение включва въведение, изложение в две глави, заключение, списък с използвана литература – 148 източника (39 на кирилица и 109 на латиница), 4 приложения и декларация за оригиналност. В основния текст на дисертационния труд са включени 8 таблици и 16 фигури.

Структурата на дисертационния труд е следната:

Въведение

Глава първа. Теоретико методологически проблеми на приложението на „големите данни“ (Big Data) в процеса на статистическия анализ.

1.1. Същност и еволюция на Big Data

1.2. Основни източници на Big Data

1.3. Big Data и парадигмата за класическото статистическо изследване

1.4. Методологически проблеми при обработката на Big Data

1.5. Оценка на качеството на Big Data в официалната статистическа практика

1.6. Big Data технологии

Глава втора. Приложни аспекти на „големите данни“ в официалната статистика (Емпирично изследване „Извличане на информация от интернет за характеристики на предприятията (web scraping)“)

2.1. Концептуална рамка на емпиричното изследване

2.2. Технологична среда за приложението на „web scraping“

2.3. Практическа реализация на пилотните „сценарии“

Заключение

Използвана литература

Приложения

Декларация за оригиналност

III. КРАТКО ИЗЛОЖЕНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

Глава първа
ТЕОРЕТИКО-МЕТОДОЛОГИЧЕСКИ ПРОБЛЕМИ НА ПРИЛОЖЕНИЕТО НА „ГОЛЕМИТЕ ДАННИ“ (BIG DATA) В ПРОЦЕСА НА СТАТИСТИЧЕСКИЯ АНАЛИЗ

Изложението в първа глава е посветено на изясняването на теоретичните основи и методологическите аспекти на „големите данни“ в контекста на приложението им в официалната статистическа практика.

В **първи параграф** е направена характеристика на съвременната информационна и технологична среда, която обуславя възникването и развитието на концепцията за „големите данни”. Информационната експлозия или т. нар „потоп от данни“, намира израз в различни области на обществения живот. Посочени са редица примери и факти, които ясно очертават основните сфери и източници, генериращи „големи данни“ и обуславят възникването на феномена „Big Data“. Терминът е многостранно описание на богат и сложен набор от характеристики, практики, техники, етични въпроси и резултати, който първоначално произхожда от физическите науки и астрономията. Естеството на Big Data показва, че те не трябва да се разглеждат като технологичен, а по-скоро като съдържателен феномен от обширно количество необработена информация, пресичаща общественото пространство и генерираща се чрез бизнеса и държавните организации.

Представена е еволюцията на теоретичните концепции за Big Data и въз основа на критичен преглед на развиваните в литературата постановки за тяхното естество, е изведена тяхната същност. Голяма част от дефинициите за Big Data се фокусират преди всичко върху техните източници, но разнообразният им характер затруднява дефинирането на термина. Класическата дефиниция в Уикипедия гласи „Big Data е събиране на голям и комплексен набор от данни, който е трудно да бъде обработен, използвайки традиционните средства за управление на база данни или традиционните софтуерни приложения

за обработка на данни“[2]. Фирмата Opentracker, специализирана в уеб проследяване и предлагане на множество уеб услуги и аналитични разработки, е събрала и систематизирала повече от 30 дефиниции за Big Data от различни автори[3]. В този контекст са и дефинициите на българските автори, работещи в областта на големите данни[4].

Основен критерий при дефинирането на същността на „големите данни“ в литературата са техните обем, разнообразие и скорост. Този общ методологически подход е основание да се направи следното обобщение за същността на „големите данни“: ***Big Data е понятие, което описва голям обем на висикоскоростен набор от постоянно променящи се данни, изискващи модерни средства и технологии за обработване, съхранение, разпространение, управление и анализ на информацията.***

За да се обоснове използването на „големите данни“ като източник на информация за различни статистически, икономически и социални изследвания, са изяснени техните характеристики. Обемът (Volume), скоростта (Velocity), разнообразието (Variety) и истинността (Veracity), които представят т.нар.. ***„дефиниция за четирите V-та“***[5], се допълват с характеристиките

[2] Big data. https://en.wikipedia.org/wiki/Big_data.

[3] Definitions of Big Data. http://www.opentracker.net/article/definitions-big-data; Demystifying Big Data: A Practical Guide To Transforming The Business of Government. (2012). Prepared by TechAmerica Foundation’s Federal Big Data Commission. file:///D:/Demistyfying%20Big%20Data.pdf; Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Hung Byers, A. (2011). Big data: The next frontier for innovation, competition and productivity. Report McKinsey Global Institute. https://bigdatawg.nist.gov/pdf/MGI_big_data_full_report.pdf и др.

[4] Попов, В. (2014). Big Data предизвикателство пред системите за управление на съдържанието. В: Информационните технологии в бизнеса и образованието - сб. с докл. от междунар. научна конфер., посвет. на 45 годиш. от създаването на кат. "Информатика" в ИУ Варна. Наука и икономика, Варна, с. 351-358; Попов, В. (2016). Анализ на Big Data - методи, технологии и инструменти. В: Предизвикателства пред информационните технологии в контекста на "Хоризонт 2020" - юб. науч. конфер., 7-8 окт. 2016 г., Сб. с докл. Свищов, АИ Ценов, с. 101-107; Михайлова, А. (2014). Големи данни: размерът има значение. Мегавселена, 20 февруари. http://megavselena.bg/golemi-danni-razmeryt-ima-znachenie; Сгурев, В., Дрангажов, С. Problems of the Big Data and Some Applications. Международна конференция на тема: Big Data, Knowledge and Control Systems Engineering (BdKCSE’2014), София 2014, http://conference.ott-iict.bas.bg/wp-content/uploads/2014/01/BdKCSE2014; Хаджичонев, Х. (2017). Големите данни: Голямата дигитална бизнес промяна. Твоят бизнес, онлайн списание за предприемчивите българи, 18 април. http://www.tbmagazine.net/statia/golemite-danni-golyamata-digitalna-biznes-promyana-prva-chast.html и др.

[5] Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety. META Group, 6. February. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf;Crawford, K. (2011). Six Provocations for Big Data. Social Science Research Network: A Decade in Internet Time, Symposium on the Dynamics of the Internet and Society. http://softwarestudies.com/cultural_analytics/Six_Provocations_for_Big_Data.pdf.

променливост (Variability) и сложност (complexity)[6], което ги превръща в **Big Data 6V**. Направено е обстойно описание на посочените характеристики, в резултат на което е разработено в табличен вид дескриптивно представяне (описание, атрибути, управление) на най-важните характеристики на Big Data – обем, скорост, разнообразие и истинност.

Обосновано е значението и необходимостта от въвеждането на „големите данни“ в обхвата на официалната статистика, като са изведени някои ефективни и новаторски начини относно тяхното приложение. За по-пълното изясняване на феномена Big Data са посочени няколко примера от световната статистико-икономическата практика: Онлайн наблюдение на цените за получаване на *експресна оценка на инфлацията* в САЩ; наблюдение на *пътния трафик и идентифициране на инфраструктурни проблеми* в Холандия; изследване на *социално-медийното настроение* в Холандия и др.

Във **втори параграф** е направена характеристика и обобщена класификация на основните източници на „големи данни“. Обоснована е необходимостта от усъвършенстване на вече създадените или разработване на нови наръчници и класификации на източниците на Big Data. Това допринася за вземането на информирани решения за действие и успех на пилотните проекти на европейско и национално ниво и осъществяване на мониторинг по отношение на тяхната автентичност, качество и достоверност.

Натрупаният опит по отношение на Big Data в европейски и световен мащаб дава възможност да се разработят различни класификации за източниците на „големите данни“. Представени са последователно няколко възможни таксономии на източниците на Big Data, разработени от водещи компании[7] и международни организации[8]. Обобщавайки съществуващите групировки, авторът на дисертационния труд стига до заключението, че за нуждите на настоящото изследване и решаването на бъдещи практически задачи, като

[6] Amodeo, L. (2014). Big Data 6V: volume, variety, velocity, variability, veracity, complexity. The Journal insights, 24. deciembre. https://wydata.wordpress.com/2014/12/24/big-data-volume-variety-velocity-variability-veracity-complexity/.

[7] Devlin, B., Rogers, S., Myers, J. (2012). Big Data Comes of Age. Enterprise management associates (EMA) и 9sight Consulting Research, November. http://9sight.com/pdfs/Big_Data_Comes_of_Age.pdf.

[8] Big Data and the SDGs. (2016). United Nations Global Pulse. https://www.slideshare.net/unglobalpulse/big-data-and-the-sdgs.

водеща може да се приеме класификацията на типовете източници на Big Data, разработена от специално създадената за целта работна група към Статистическата комисия на ООН[9].

В посочената разработка релефно могат да се открояят ***три основни групи източници*** на Big Data:

Първа група: Социални мрежи (информация, генерирана от населението). Тази група съдържа информация, която по същество е запис на човешкия опит. В миналото е описвана в книги и произведения на изкуството, фотографии, аудио- и видеозаписи. В наши дни информацията е почти изцяло дигитализирана и се съхранява навсякъде – от персоналните компютри до социалните мрежи. Данните са неструктурирани и са трудно обработваеми и управляеми. По-конкретно тук се отнасят:

- социални мрежи: Facebook, Twitter, Linkedin и др.;
- електронни блогове;
- лични документи;
- снимки: Instagram, Flickr, Picasa и др.;
- видеоклипове: Youtube и др.;
- интернет търсачки;
- данни от мобилни телефони: текстови съобщения;
- електронни карти;
- електронна поща.

Втора група: Традиционни бизнес системи и уебсайтове (данни, генерирани от информационни системи). Тази група съдържа информация, която е в резултат от стопанска дейност, като регистриране на потребители, производство на продукти, заявки, финансови трансакции и други. Данните са ***структурирани в определен формат***, предимно табличен, дефинирани са връзките между тях и метаданните за тяхното съдържание. Съхраняват се в релационни бази данни или системи (някои източници, принадлежащи към този

[9] Vale, S. (2013). Classification of Types of Big Data. https://statswiki.unece.org/display/bigdata/Classification+of+Types+of+Big+Data.

клас, могат да попаднат и в категорията „Административни данни“). Тук се включват:

- данни, генерирани от държавни агенции, например медицински записи;
- данни, генерирани от бизнеса, в т.ч.:
  - ✓ търговски трансакции;
  - ✓ банкови записи/борсови операции;
  - ✓ електронна търговия;
  - ✓ кредитни карти.

Трета група: Интернет на нещата (данни, генерирани от машини/сензорни устройства). Тази група съдържа информация, генерирана от машини и сензорни устройства, които измерват и записват събитията и ситуациите във физическия свят. Данните са ***добре структурирани и са подходящи за компютърна обработка***, но размерът и скоростта им са извън възможностите на традиционните методи за обработка на данни. Към тази група се отнасят:

- Данни от сензори:
  - ✓ постоянни сензори, в т.ч.:
    - домашни автоматизирани системи;
    - сензори за времето/замърсяването;
    - пътни сензори/уебкамери;
    - температурни сензори;
    - охранителни/наблюдателни видеокамери.
  - ✓ Мобилни сензори (проследяващи устройства), в т.ч.:
    - локация на мобилни телефони;
    - навигационни системи за автомобили;
    - сателитни снимки/изображения.
- Данни от компютърни системи:
  - ✓ логове;
  - ✓ уеблогове.

Класифицирането на източниците на Big Data е важно, тъй като по този начин се определя стратегията за достъп до тези източници, събиране на

информация от тях, като се дефинират познавателни задачи за бизнес процесите в различни сфери на обществения живот. Едновременно с това, новите източници на данни формират нова информационна среда за управление на финансовите потоци, инвестициите, работната ръка и обмяната на материални ресурси, стоки и услуги.

Изложението в **трети параграф** е посветено на Big Data и промяната на парадигмата за класическото статистическо изследване. Извършен е сравнителен анализ на етапите на провеждане на едно изследване, базирано на данни от официалната статистика и от източници на Big Data. Констатирано е, че производственият процес в Big Data света е много по-различен в сравнение със статистическия производствен процес.

Анализирани са начините, по които „големите данни“ могат да бъдат използвани за производство на официална статистика, а именно:

- замяна изцяло на статистическите източници, основани на общи дефиниции, класификации и т.н., което е малко вероятно в обозримото бъдеще;
- частична замяна на статистическите източници, като информацията се допълва чрез съчетаване на данни от различни източници;
- осигуряване на напълно нови статистически числа, които могат да допълват и се интегрират с наличната статистическа информация, което е значително по-добрият начин за тяхното съвместно използване.

Първите два начина вероятно биха могли да доведат до намаляване на разходите и натоварването на респондентите, но това от своя страна ще доведе до нови задачи за адаптиране, съчетаване и хармонизиране на различни структури от данни към вече утвърдени и общоприети статистически концепции, дефиниции и класификации. Логично погледнато, „големите данни“ не могат да заменят напълно или частично статистическите източници в краткосрочен план и това би било твърде скъпо по отношение на времевите, финансовите и човешки ресурси. Наред с това, между статистическите данни и „големите данни“ се наблюдават моментни процеси на конвергенция, които са необходими за управление на бизнеса. Фирмите от частния сектор, произвеждащи статистика

на основата на „големите данни“, следват третия път и не се сблъскват с подобни проблеми.

Акцентирано е върху необходимите трансформации, които трябва да осъществят националните статистически институти, за да използват „големите данни“ като източник на информация. *Първо***,** ще се изисква промяна в начина, по който е организиран традиционният производствения процес, и модернизация на статистическите производствени системи с цел повишаване тяхната ефективност и гъвкавост. На *второ място*, Big Data със сигурност ще донесе нови задачи и отговорности на националните статистически организации, свързани преди всичко с гарантиране качеството на статистическите данни, произведени от източниците на Big Data чрез механизъм за акредитация и сертификация.

Изяснени са насоките, в които могат да се търсят ефектите и рисковете от приложението на Big Data в официалната статистика. *Ефектите* от използването на „големите данни“ за аналитични цели могат да се търсят основно в две направления: *намаляване натовареността на респондентите и намаляване цената на информационния продукт*. Едно от най-важните предимства на Big Data е, че те вече съществуват под някаква форма и поради тази причина изследванията на базата на Big Data не изискват допълнителни усилия и ресурси за нов процес на събиране на тези данни в зависимост от специфичните цели.

Рисковете от използването на източници на Big Data за производство на официална статистика могат да бъдат групирани в пет основни категории въз основа на резултати от проведено от Евростат онлайн проучване[10]:

- рискове, свързани с достъпа до данни;
- рискове, свързани със законодателството;
- рискове, свързани с конфиденциалността и сигурността на данните;
- рискове, свързани с уменията;
- други рискове, посочени от респондентите.

[10] UNSD (2016). Report of the Big Data Survey 2015. Statistical Commission, Background document, 47 session, March 8 – 11. https://unstats.un.org/unsd/statcom/47th-session/documents/BG-2016-6-Report-of-the-2015-Big-Data-Survey-E.pdf.

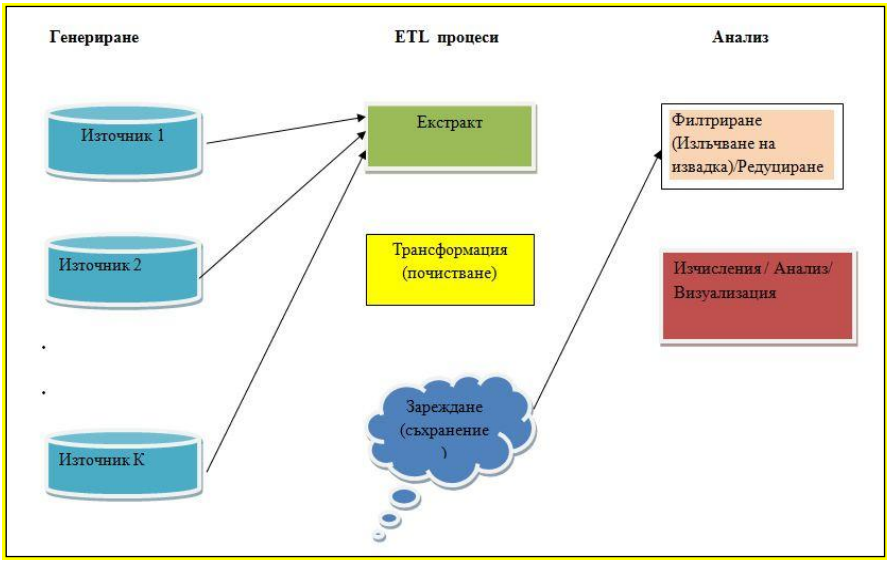
В резултат на направен сравнителен анализ на характеристиките на традиционните статистически изследвания и тези, базиращи се на източници на Big Data, в табличен вид са представени особеностите на различните източници – основни (пълни и извадкови изследвания), вторични (административни данни) и третични (Big Data).

В **четвърти параграф** са разгледани основни методологически проблеми при обработката и анализа на „големите данни“. Представени са дефинираните в научната литература[11] генерични стъпки в процеса на приложение на Big Data. Тези етапи са следните:

- *Генериране* – инцидентно или целенасочено получаване на данни от даден източник;
- *Екстрахиране / Трансформиране / Зареждане (ETL)* – обединяване на всички данни в хомогенна изчислителна среда на три етапа:
  - ✓ *екстрахиране* – данните се събират от техните източници, „раздробени“, валидирани и съхранявани;
  - ✓ *трансформиране* – данните са преведени, кодирани, прекодирани, агрегирани/дезагрегирани и/или редактирани;
  - ✓ *зареждане* – данните са интегрирани и съхранени в база данни.
- *Анализ* – данните се превръщат в информация чрез процес, включващ:
  - ✓ *филтриране* (*излъчване на извадка*)/*редуциране* – нежеланите характеристики и съдържание се заличават, някои характеристики се комбинират с цел получаване на нови, премахват се или се добавят елементи на данните, за да са по-лесно обработваеми при следващите процеси;
  - ✓ *изчисления/анализ/визуализация* – данните се анализират и/или представят за интерпретация и извличане на информация.

На фигура 1 са представени процесите за обработка на Big Data и движението на потоци от данни по описаните по-горе етапи.

[11] Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J. (2015). AAPOR Report on Big Data. https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/images/BigDataTaskForceReport_FINAL_2_12_15_b.pdf.



Фиг. 1. Карта на процеса на обработка на Big Data

Направен е анализ на относимостта на посочените етапи при обработката на Big Data към етапите на традиционното статистическо изследване и са посочени възможните грешки. *Етапът на генериране* е аналогичен на наблюдението, като първа фаза на традиционното статистическо изследване, поради което и *грешките* в този етап са донякъде аналогични на грешките, допускани при набирането на първичните данни – погрешни или непълни.

ETL процесите могат да съответстват на операции от различните етапи при обработка на данните в едно статистическо изследване. Те могат да включват създаване на метаданни, свързване на записи, кодиране на променлива, редактиране и интегриране на данни (свързване и обединяване на записи и файлове в различни системи). От своя страна, грешките на етапа Екстракт/ Трансформиране/Зареждане включват: спецификация на грешка (вкл. грешки в метаданните); грешки на свързването; грешки при кодирането; грешки при редактирането и интегрирането и други.

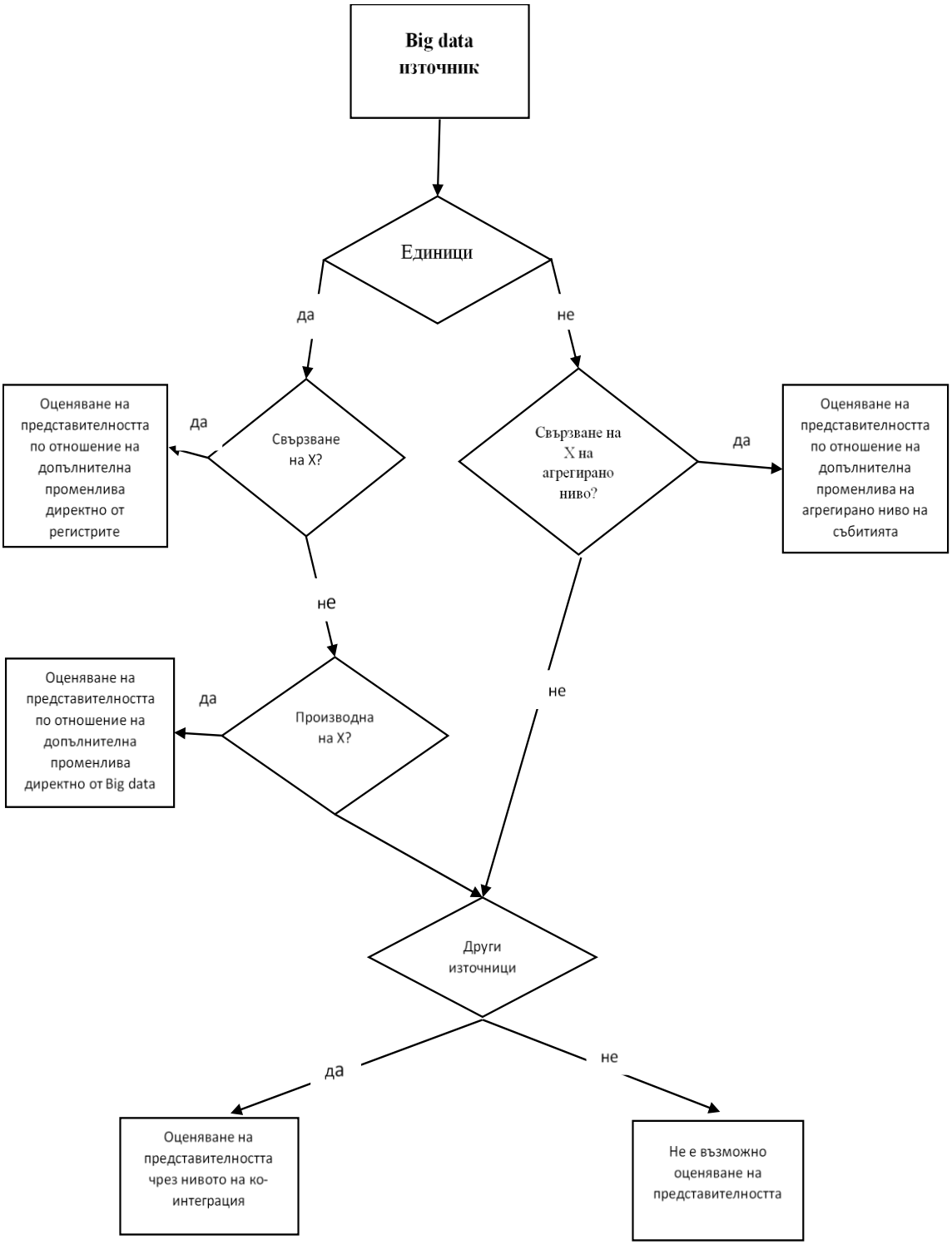
На третия етап от обработката на Big Data – *анализа*, съществува риск от натрупване на шум, фалшиви корелации и случайна ендогенност, които могат да се усложняват от извадкови и неизвадкови грешки. Във връзка с извадковите грешки, данните могат да бъдат филтрирани, включени в извадката, или редуцирани, за да образуват представителни и лесно управляеми данни. Тези

процеси могат да изискват допълнителни трансформации, изискващи значителни усилия и време за подготовка и съчетаването на данните от различни източници.

Акцентирано е на трудностите и предизвикателствата, свързани със собствеността на „големите данни“, тяхната поверителност, необходимите умения за интегрирането им, тъй като такива се проявяват независимо от етапа, на който се намира обработката на „големите данни“. Специално място е отделено на квалификацията на експертите и ресурсите, необходими за работа с Big Data. Работата и решаването на проблеми с Big Data изискват екип от минимум четири типа експерти: *експерт в специфичната област,* за която се отнасят данните, *изследовател (методолог), компютърен специалист* и *системен администратор.* Изяснени са уменията и компетенциите, които трябва да притежава всеки един от тях. Отбелязани са и съответните технологични изисквания, които автоматично се превръщат в поредното предизвикателство, свързано с новите източници на данни.

Подробно е разработен проблемът с грешките, произхождащи от комплексния процес на обработка, необходим за производството на статистически информационен продукт от Big Data. Едно от основните предизвикателства при използването на „големи данни“ е тяхната потенциална селективност. Отделните източници се различават значително по механизмите, чрез които се генерират, но едновременно с това имат общи черти по отношение на репрезентативните извадкови изследвания и стратегията за събиране на данни. Когато „големите данни“ се обсъждат в контекста на официалната статистика, често се критикуват методите за тяхното събиране, при които липсва прилагане на вероятностния подход за излъчване на извадка. По този начин получените резултати от Big Data източници имат значителни отклонения и не са представителни за изучаваната съвкупност. На фигура 2 е показана една възможна методология, прилагана от холандските статистици, работещи върху методологичните проблеми на Big Data[12].

[12] Buelens, B., Daas, P., Burger, J., Puts, M., Brakel, J. (2014). Selectivity of Big data. Discussion Paper, №11, Statistics Netherlands. file:///D:/2014-11-x10-pub.pdf.



Фиг. 2. Диаграма за оценяване на селективността на Big Data източниците

В **пети параграф** са представени и теоретично обосновани подходите за оценка на качеството на Big Data. Обсъдени са предложения и възможни решения за повишаване качеството на „големите данни“ при използването им в официалната статистическа практика въз основа на натрупания практически опит в областта на Big Data и са посочени предизвикателствата пред националните статистически институти на европейските държави в този контекст.

Един от възможните подходи за оценка на качеството на „големите данни“ е свързан с *разширяване обхвата на използваните статистически методи.* Много от вече утвърдените методи са предназначени за класическо статистическо изследване, но в действителност повечето изследвания използват за рамка на съвкупността данни, получени от административни източници. Логично следствие от преследваните цели за намаляване на тежестта на респондентите и постигане на максимална ефективност е търсенето на алтернатива, каквато са и „големите данни“. Към така наречената „коминна“ организация на статистиката, при която, всеки от статистическите процеси се изпълнява независимо един от друг, все повече се налага и „интегративната“ статистика на базата на множество източници и прилагани методи извън традиционната извадкова теория, които се прилагат за работа с Big Data.

По отношение на ***ограничителните условия за качеството*** в официалната статистика могат да бъдат разграничени две нива. *Първо,* качеството на статистическите данни е пряк резултат от прилаганите методи и тяхната параметризация. *Второ*, разработени са рамки за качество в Европейската статистическа система[13], които се прилагат за дефиниране на критерии за качеството и оценка на информационните продукти, на които трябва да отговарят и статистиките, получени чрез Big Data.

Предизвикателствата при оценка на качеството на Big Data могат да се търсят в няколко посоки, тъй като към настоящия момент използваемостта на Big Data като надежден източник за производство на официални статистически

[13] European Statistics Code of Practice. Eurostat Quality Assurance Framework. http://ec.europa.eu/eurostat/web/quality/overview.

данни е в тестови етап и е обект на пилотни проекти на европейско и национално ниво. За илюстрация са посочени няколко примера от европейската практика, свързани с различни видове източници на Big Data, които изискват иновативни решения и формират интересни методологични казуси. Проучени са различни възможни решения на базата на резултати от тестови пилотни проекти, които зависят от начина и предназначението на използване на Big Data[14].

В **шести параграф** е направена обща характеристика на най-често използваната ИТ инфраструктура за техническото обезпечаване на работата с Big Data, която включва специализирани ИТ средства, системи и техники, включително специализиран софтуер. Разгледани са основните техники за извличане на данни от интернет за нуждите на статистически и други изследвания, като специално внимание е отделено на същността на извличане на уеб страница (*„web crawler“*) и извличане на данни („*web scraping“*) [15]. Посочени са и някои рискове при прилагането на тези техники, като нарушаване правата на собственика на информацията и/или потребителските споразумения за използването на уебсайтове.

Направена е характеристика на най-често използваните технологии за обработка на „големи данни“:

- **Масивна паралелна обработка (MPP).** Архитектурата на релационна база данни за масивна паралелна обработка разпространява данни по редица независими сървъри или възли по прозрачен начин за потребителите. За работа с Big Data често се използват аналитични MPP системи, обикновено наричани бази данни "*shared-nothing*", тъй като възлите, които съставляват клъстера работят самостоятелно, комуникират чрез мрежа, но не споделят ресурси на диска или паметта[16]. MPP базите имат предимства, които им позволяват да се мащабират само чрез добавяне на хардуер и използване на стандартния SQL (Structured Query Language), така че да могат лесно да се

[14] Eurostat (2014). The Role of Big Data in the Modernisation of Statistical Production. 2014 UNECE Project. https://statswiki.unece.org/download/attachments/90636852/Big%20data%20project%20outline%20with%20sandbox%20annex.docx?version=1&modificationDate=1385740790968&api=v2.

[15] Hemenway, K., Calishain, T. (2003). Spidering Hacks. Cambridge, Massachusetts: O'Reilly. ISBN 0-596-00577-6.

[16] Fernández, F. (2010). Parallel and Distributed Computational Intelligence, ISBN 3-642-10674-9.

интегрират с ETL (Екстракт/Трансформиране/Зареждане) инструменти за визуализация и показване, без да се изискват нови умения.

- **Нерелационни бази данни (NoSQL – Not only Structured Query Language, NoSQL)**. NoSQL база данни е подход за управление и проектиране на база данни, която е приложима за много големи масиви от разпределени данни. NoSQL обхваща широк спектър от технологии и архитектури, предназначени да разрешат проблемите на мащабируемостта и представянето на Big Data, с които релационните бази данни не могат да се справят успешно[17]. NoSQL е особено полезна за достъп и анализ на огромно количество неструктурирани данни или данни, които се съхраняват от разстояние на няколко виртуални сървъри. Макар да е вярно, че някои NoSQL системи са изцяло нерелационни, други просто избягват избрана релационна функционалност, като например фиксирани таблични схеми и свързани операции. Съществува голямо разнообразие на различни категории NoSQL бази данни, по-важните от които са:[18] системи „ключ-стойност“; колонно-ориентирани бази данни; документно-ориентирани бази данни; графични системи.

- **Map-Reduce**. Map-Reduce е програмиращ модел и асоциативно приложение за обработка и генериране на големи обеми от данни с паралелен, дистрибутивен алгоритъм за един клъстер.[19] Потребителите определят свързваща функция, която обработва двойка ключ/стойност, за да генерира набор от междинни двойки ключ/стойност, както и да редуцира функцията, обединяваща всички междинни стойности, свързани с един и същ междинен ключ. Програми, написани в този функционален стил, автоматично се преобразуват във форма, която позволява работа с голяма група паралелни компютърни машини. Системата за хода на времето се грижи за детайлите на разделяне на входните данни, разписание на изпълнението на програмата сред набор от компютърни машини, транспортни аварии на машините, както и

[17] Leavitt, N. (2010). Will NoSQL Databases Live Up to Their Promise? IEEE Computer, february. http://leavcom.com/pdf/NoSQL.pdf.
[18] NoSQL DEFINITION. http://nosql-database.org/
[19] Dean, J., Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters. OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco. https://www.usenix.org/legacy/publications/library/proceedings/osdi04/tech/full_papers/dean/dean_html.

управление на необходимата вътрешно машинна комуникация. Това позволява на програмистите, без никакъв опит с паралелни и дистрибутивни системи, лесно да оползотворяват ресурсите на всяка голяма разпределителна система.

- **Екосистема Hadoop.** През последните години програмисти от цял свят изграждат и тестват разнообразни ИТ инструменти с цел адаптиране и повишаване практическата използваемост на системата Hadoop. Като резултат от тяхната работа е разработен списък на компонентите, необходими за изграждане и управление (около основната ос *Hadoop - MapReduce*) на Big Data приложенията в реалния свят[20].

- **Облачни ИТ технологии.** Облачните технологии предлагат набор от възможности за анализ на големи данни както в публичните, така и в частните настройки на облака, както от гледна точка на инфраструктурата, така и от гледна точка на аналитичните цели. Относно инфраструктурата, *Cloud* предоставя възможности за управление и достъп до много големи масиви от данни, както и за поддържане на мощни инфраструктурни елементи на сравнително ниска цена. Виртуалната, адаптируема, гъвкава и мощна природа на облака със сигурност се поддава на огромната и променяща се среда на Big Data. Архитектурата на облачните технологии се състои от редица виртуални машини, които са идеални за обработка на много големи набори от данни, доколкото обработката може да бъде разделена на множество паралелни процеси и това го прави идеална компютърна среда за Big Data.

- **Визуализиране на Big Data.** Визуалният анализ осигурява технология, която съчетава креативните страни на човека и електронната обработка на данни, т.е. визуализацията се превръща в среда на полуавтоматичен аналитичен процес, при който хората и машините си сътрудничат, използвайки съответните си способности за най-ефективни резултати. По отношение на прилагането на визуализиращи ИТ средства за Big Data, възникват някои въпроси, свързани с основните характеристики на големите данни. Слабо са развити методите за моделиране и визуализиране на

[20] Borthakur D. et al. (2011). Apache Hadoop goes realtime at Facebook. Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, pp. 1071-1080. http://ebookbrowsee.net/realtimehadoopsigmod2011-pdf-d146414448.

полу- или неструктурирани данни, ограничени са възможностите при работа с „големи данни“, тъй като софтуерните производители са склонни да се съсредоточават само върху малък брой "стандартни" техники за визуализация, като по този начин се забавя интеграционния процес на новаторските техники. Съществува необходимост от проектиране на персонализируеми функции за визуализация, като по този начин потребителят има свобода да променя настройките на визуалните параметри и повече възможности за съществени, аналитични изводи от визуално представената информация.

В **обобщение** може да се отбележи, че теоретичната концепция за същността, характерните чести и особености на „големите данни“ непрекъснато се обогатява и развива в съответствие с динамично развиващите се информационни и комуникационни технологии. Все още обаче възможностите за приложението на Big Data като източник на информация за официалната статистика не са достатъчно проучени и апробирани. Осигуряването на информация с високо качество е един от най-важните елементи на официалната статистика. Статистическата информация, която ще се произвежда от източниците на „големи данни“, е необходимо да е пригодна и да отговаря на стандартите за качество на Европейската статистическа система, за да бъде използваема от потребителите. Характеристиките на „големите данни“, включително техният обем, скорост и разнообразие, оказват влияние върху ИТ системите и инфраструктурата. Бъдещите ИТ инфраструктури ще бъдат определени предимно от новите бизнес модели, които ще се внедрят за производство на статистика от Big Data.

Перспективите за интегриране на „големите данни“ в официалната статистика са свързани с разработването на пилотни проекти за придобиване на знания и опит, които включват дейности по анализа на източниците на „големи данни“, изследване на потенциала за партньорство с доставчиците на данни, апробиране на ИКТ и т.н. Изложението в следващата глава на дисертационния труд е посветено на въпросите, свързани с организацията, провеждането и резултатите от емпирично проучване на екип от НСИ в рамките на подобен проект.

Глава втора
ПРИЛОЖНИ АКПЕКТИ НА „ГОЛЕМИТЕ ДАННИ“ В ОФИЦИАЛНАТА СТАТИСТИКА
(Емпирично изследване „Извличане на информация от интернет за характеристики на предприятията (web scraping)“)

Изложението във втора глава на дисертационния труд е посветено на представянето на една възможност за практическо приложение на „големите данни“, която е реализирана в рамките на проведено от екип на НСИ емпирично изследване на тема „Извличане на информация от интернет за характеристики на предприятията (web scraping)“).

В **първи параграф** се разглежда концептуалната рамка на изследването. Обоснована е необходимостта от разработването на международни и национални проекти в областта на Big Data и ключовата роля, която играе Евростат като източник на финансови, идейни и технически ресурси, включително организатор на редица специализирани обучения, семинари и научни конференции по тази тематика.

Националните статистически институти създават разнообразни масиви от статистически данни за използване на информационните и комуникационните технологии, които се използват за наблюдение на напредъка на страните към информационното общество. Бързото развитие на съвременните ИКТ поставя необходимостта от разработване на показатели, които да са повече релевантни и навременни от тези, изчислявани на база традиционните изследвания.

Направена е кратка характеристика на проведено през 2013 г. от националния статистически офис на Италия (ISTAT) тестване на техники за извличане на информация от интернет (web scraping) и извличане на съдържание от текст (text mining), за да направи опит за промяна на традиционното изследване „Използване на ИКТ в предприятията“ (Information and Communication Technologies in Enterprises)[21]. Натрупаният опит от успешното използване на „големите данни“ е изучаван и споделян с други държави с цел

[21] Information and communication technologies in enterprises. http://www.istat.it/en/archive/77760.

извличане на ценни познания и прилагане на добри практики по отношение на Big Data.

През ноември 2015 г. Националният статистически институт на Р България се включи като страна партньор в ESSnet проект „Рамково споразумение за сътрудничество Big Data План за действие", който е разделен на две отделни грантови споразумения (SGA-I и SGA-II) и се изпълнява последователно за периода 2016-2018 година. В рамките на това споразумение ще се постигат целите, заложени в т.нар. BDAR (Big Data Action Plan and Roadmap v. 1.0), който е част от портфолиото на ECC Vision 2020. Продължителността на SGA-I проекта е 18 месеца (февруари 2016 - юли 2017 г.), на обща стойност 1111111 евро и в него вземат участие консорциум от 20 национални статистически института и два органа на статистиката, като координатор на проекта е Статистическият институт на Нидерландия. Основната цел на проекта е да подготви Европейската статистическа система за интегриране на източници на „големи данни" в процеса на производство на официална статистика.

Международният опит показва, че **статистиката за използването на ИКТ от предприятията е естествен и логичен „кандидат" за пилотен проект** и реинженеринг на базата на интернет и подобни източници. Поради тази причина, в рамките на европейския проект, **през 2016 г. екип**[22] **от НСИ** провежда емпирично изследване на тема „Извличане на информация от интернет за характеристики на предприятията (web scraping)".

Основната цел на проведеното емпирично изследване е аналогична на проведеното от италианския статистически институт и е насочена към проучване на възможностите за прилагането на техниките „web scraping" и „text mining" и други подобни, както и да се оцени ефекта от използването им в процеса на събиране на данни и подобряване на качеството на информацията за предприятията от статистическия бизнес регистър на НСИ чрез достъп до техните уеб сайтове.

[22] Авторът на дисертацията е ръководител на екипа и участва активно като основен статистически експерт при организацията, провеждането и анализа на резултатите от изследването.

За сравнение са използват данните, получени чрез статистическото изследване „Използване на ИКТ от предприятията“, което се провежда регулярно и съгласно европейските стандарти и националното законодателство от Националния статистически институт. Наблюдението е годишно, извадково, като в обхвата му на случаен принцип се включват около 4900 предприятия. Генералната съвкупност обхваща всички предприятия от нефинансовия сектор с 10 и повече заети лица, включени в бизнес регистъра на НСИ. Именно тази съвкупност е базова за провеждане на изследването за извличане на „големи данни“ от интернет чрез техниките на „web scraping“. В списъка са включени 26836 предприятия, които са проучени за наличие на сайтове. Първоначално са открити 2006 броя URL адреси и 20649 броя е-мейл адреси.

Посочени са етапите на провеждане на проучването и е изяснено съдържанието на работния термин „Use-case“ или „сценарий“ (и още „случай на използване/употреба“). Терминът е от областта на софтуерното и системното инженерство и може да се разглежда като методология, използвана в системния анализ за идентифициране, изясняване и организиране на системните изисквания. „Сценарият“ се състои от набор от възможни последователности на взаимодействия между системите и потребителите в дадена среда, свързани с определена цел. Един „сценарий“ може да се разглежда като съвкупност от възможни действия, свързани с определена цел, като понякога „сценарият“ и целта се считат за синоними[23].

В обхвата на изследването са избрани следните **„use-cases“**:

- ✓ **Use-case 1.** Генериране на списък с фирмени URL адреси на предприятията за бизнес регистъра **(URLs retrieval)**.

[23] Cockburn, A. (2002). Use cases, ten years later. http://alistair.cockburn.us/Use+cases, +ten+years+later. Jacobson, I., Spence, I., Bittner, K. (2011). Use case 2.0: the guide to succeeding with use cases. Ivar Jacobson International, december. https://www.ivarjacobson.com/sites/default/files/field_iji_file/article/use-case_2_0_jan11.pdf; Jacobson I. (2004). Object-oriented software engineering: a use case driven approach. Addison Wesley Longman Publishing Co., Inc. Redwood City, CA, USA. ISBN:0201403471. Zielczynski, P. (2006). Traceability from use cases to test cases. IBM developerworks, 10 february. https://www.ibm.com/developerworks/ratioLarman, C.(2004). Applying UML and patterns. An introduction in object-oriented analysis and design and iterative development. Addison Wesley Professional, pp. 63–64. ISBN 0-13-148906-2 .nal/library/04/r-3217/; Larman, C.(2004). Applying UML and patterns. An introduction in object-oriented analysis and design and iterative development. Addison Wesley Professional, pp. 63–64. ISBN 0-13-148906-2 .

- ✓ **Use-case 2.** Електронна търговия в предприятията **(E-commerce)** – прогнозиране дали дадено предприятие предоставя възможности за електронна търговия на фирмения си уебсайт или не.
- ✓ **Use-case 3.** Присъствие на предприятията в социалните медии **(Social media presence)** – търсене и събиране на информация от фирмения уебсайт дали дадено предприятие съществува в различни социални медии.
- ✓ **Use-case 4.** Апробиране на софтуер (разработен от италианския статистически офис ISTAT) за генериране на списък с фирмени URL адреси на предприятията (**URLs retrieval**) и сравнение на получените резултати.

Извършено е детайлно описание на всеки „use-case“ и осъществяване на същинското извличане на данни от уебсайтове. За описание на всеки „use-case“ е разработен специален шаблон, който съдържа следните елементи: идентификационен номер; наименование; описание; участници; предварителни условия; очаквани резултати; честота на използване; сценарии; специални изисквания; въпроси. За практическото приложение на техниките „web scraping“ **екипът е разработил авторски софтуер**, чрез който е извършено изследването. За нуждите на изследването е направено *концептуално сравнение* между етапите на класическия производствен процес за провеждане на едно традиционно изследване и основните фази на бизнес процеса за получаване на информация от източници на Big Data.

Във **втори** параграф е представена технологичната среда за осъществяване на емпиричното изследване. За прилагане на техниките на „web scraping“ е разработена обща референтна логическа архитектура[24], съставена от четири блока, съответстващи на четирите основни етапа на работата по извличането на данни от уебсайтовете на предприятия, а именно: „Интернет достъп“ (Internet access); „Съхранение“ (Storage); „Подготовка на данни“ (Data preparation); „Анализ“ (Analysis). За всеки етап са описани логическите

[24] Web Scraping: Applications and Tools. (2015). European public sector information platform, Topics report № 10. https://www.europeandataportal.eu/sites/default/files/2015_web_scraping_applications_and_tools.pdf.

функционалности, които трябва да бъдат изпълнени от специфичните софтуерни продукти, разработени за нуждите на настоящото емпирично изследване.

Направена е обстойна характеристика на техниката „web scraping“ и са описани случаите, в които се използват различните видове – специфичен и генеричен „web scraping“. Изяснена е същността на подходите за извличане на уеб съдържание – подходи за машинно самообучение и детерминистични подходи. Въз основа на обработката на данните на НСИ със специално разработения за целта софтуер, е направено заключение, че ***конвенционалните ИТ инструменти са достатъчни за създаването на списък с URL адреси за уебсайтовете на няколко десетки хиляди предприятия.***

Изложението в **трети параграф** е посветено на практическата реализация на четирите пилотни „сценария“. Провеждането и анализът на резултатите от отделните „Use case“ са представени в аналогична последователност – цел, ресурсна и технологична осигуреност, постигнати резултати, правни ограничения.

За ***Use case 1: Генериране на списък с URL адреси на предприятията (URLs retrieval)*** резултатите включват тестови набор от ***9809 URL адреси*** на предприятия, проверени и установени като действителни, които са съхранени в базата данни в текстов и html формат. Скриптът за резултати и статистики дава информация за корпоративните URL адреси в реално време.

За ***Use case 2: Електронна търговия в предприятията (E-commerce)*** е приложена техниката „web scraping“ върху заглавните уеб страници на официалните уебсайтове на фирмите и е разработена прогноза за наличието на е-търговия на корпоративния сайт. По този начин са открити общо ***856 уеб страници за електронна търговия.*** Верификацията на резултатите е извършена с данните от изследването „Използване на ИКТ от предприятията“. След бенчмаркинг анализа между данните от традиционното изследване в областта на ИКТ и получената от настоящия пилотен проект информация, се получават следните резултати: ***от 26836 предприятия (обхвата на проекта), в извадката за 2016 г. на изследването попадат 4332 предприятия от тях. Намерени са 89 нови предприятия, извършващи в действителност***

електронна търговия, които са включени в обхвата на традиционното изследване, но дават отрицателни отговори при попълване на въпросника от анкетното проучване.

За ***Use Case 3: Присъствие на предприятията в социални медии (Social media presence)*** основният резултат е да се осигури информация за активността на българските фирми в социалните медии чрез извличане на уеб страницата на предприятието и последващо търсене на връзки към профилите в социалните медии. След сравнителен анализ между данните от традиционното изследване в областта на ИКТ и получената от настоящия пилотен проект информация, се получават следните резултати: ***идентифицирани са 382 нови предприятия, които присъстват в социалните медии и попадат в обхвата на изследването „Използване на ИКТ от предприятията“, но дават отрицателни отговори при попълване на въпросника на изследването.*** Предстои вземане на решение дали данните за социалната медийна активност на предприятията могат да се използват за актуализиране на статистическия бизнес регистър.

При провеждането на ***Use Case 4: Генериране на списък с фирмени URL адреси на предприятият чрез прилагане на италианския софтуер*** задачите са аналогични с тези на Use case 1 – генериране на списък от URL адреси на предприятията с тестване на софтуера на ISTAT за извличане на данни и инвентаризация на URL в български условия от базовата съвкупност. НСИ използва предложения софтуер с отворен код от ISTAT: Java Run time environment за програмите URLSearcher, RootJuice, URLScorer и URLMatchTableGenerator и Apache Solr платформа за съхранение. Решението за тестването на италианския софтуер върху българската съвкупност от предприятия беше взето с цел да се споделят добри практики и опит между европейските страни, участващи в европейския проект. Резултатът показва, че ***софтуерът прогнозира правилните URL адреси на 67% от общия брой предприятия***, което обуславя сравнително ниския процент на съвпадение на резултатите, в следствие на прилагането на българския и италианския софтуер. Заключението е, че разработеният авторски софтуер е осигурил по-добри

резултати при прилагането на „web scraping“ за извличане на уеб адреси на предприятията.

Практическата реализация на проекта поставя редица въпроси пред специалистите от НСИ, свързани с необходимостта от конструиране на сложните данни, извлечени в голям мащаб, преди по-нататъшен анализ. Не по-малко важен е проблемът с извършването на „web scraping“ от онлайн машини в мрежата и прехвърлянето на извлечените данни от офлайн мрежи в онлайн хранилища. НСИ се нуждае от сходни системи за съхранение на данни, които могат да управляват целия жизнен цикъл на данните – съхранение, проектиране и поддържане на големи бази.

З А К Л Ю Ч Е Н И Е

Big Data – конкуренти, необходимост, допълнение, временно явление или заместители на официалната статистика?! Въпрос, на който е трудно да се даде еднозначен отговор. Очевидно времето, развитието на информационните технологии и промяната в човешкото мислене на най-високо ниво ще дадат точния отговор. Към момента е ясно, че Big Data могат да се доставят по-бързо, на ниска цена и в голям обем, но те не са в състояние да заменят официалната статистика, а по-скоро могат да бъдат допълнение към нея. Въпреки това и успоредно с това, интересът към Big Data нараства, тъй като много фактори и причини предизвикват сериозни пукнатини между теорията и практиката при осъществяване на статистическите изследвания.

Един от начините да се получи отговор със задоволителна степен на достоверност е, да се анализират възможните елементи и допирни точки на „съжителството“ между „големите данни“ и официалната статистика, което е неизбежно в условията на глобализиращия се свят. В крайна сметка целта е да се намерят най-добрите източници на данни, като особено важни са реалистичността и логиката в информационните потоци.

Извършеният обзор и анализ на съвременните разбирания за Big Data дават основание да се направи изводът, че все още не съществува стройна теория, на базата на която да се правят обосновани изводи и заключения относно тяхната представителност и статистическа значимост. Фактът, че „големите данни” са големи, не е достатъчен аргумент. До началото на 20 век единственият начин за получаване на статистическа информация е бил чрез изчерпателно събиране на данни от населението. По-късно, теорията за извадковите изследвания променя тази парадигма и улеснява значително процеса на събиране на данните. Днес, научната общност и официалната статистика са в ситуация, при която много голяма част от Big Data се генерират като вторични продукти от различни процеси или дори от продукта от тези процеси. Едновременно с това провеждането на класически статистически изследвания става все по-скъпо и

много от резултатите не са достоверни, поради липса на отговор и множество други неизвадкови грешки.

Възможното решение е традиционната парадигма на статистическото изследване да се съчетае с новата парадигма за използване на различни източници Big Data, за да се постигнат максимални резултати и да се получи ценна, допълнителна информация за различни социално-икономически явления в обществото. От съществено значение е да се разработят „нови" методи, чрез които да може да се реализира пълният потенциал на Big Data. В процеса на оценяване административните данни се използват като рамки за извадките с цел подобряване на точността на оценките и в комбинация с традиционните изследвания, за да се сведе до минимум натовареността на респондентите. Социалните медийни платформи могат да бъдат използвани за получаване на бърза информация за това как хората мислят за различни явления, процеси и понятия и да се изследва зависимостта с показатели, разработвани от официалната статистика (равнището на безработица, инфлация, бедност и др.).

Въпреки теоретичните и практическите предимства на „големите данни", предпочитаната стратегия е да се използва комбинация от нови и традиционни източници на данни в подкрепа на научните изследвания, анализите и вземането на обосновани решения. Едновременно с това, данните от традиционните изследвания могат да се използват за провеждане на по-задълбочени проучвания на тенденциите, промените в тенденциите или аномалиите, които се откриват в първичните данни за мониторинг. Непрекъснато се усъвършенстват начините, по които Big Data могат да бъдат използвани за подобряване на статистическите изследвания с цел увеличаване на навременността на данните, подобряване на качеството на метаданните и намаляване на текущите разходи за тяхното събиране.

Традиционният подход за проектиране на статистическите изследвания е да се дефинира желания резултат, да се изберат подходящи източници на данни и да се оптимизира процесът. Експериментите с Big Data могат да се провеждат и в обратен ред на етапите в този подход: намиране на интересен източник на Big Data, събиране на относимата информация към изучаваното явление,

свързване на тази информация с вече налична информация от други източници (дори и само чрез създаване на корелационни модели).

Необходимо е да се създаде подходящата институционална среда, в която да се извършва тестване и експериментални проучвания на източници на Big Data. Това може да се постигне чрез изграждане на ИТ инфраструктура за работа с големи масиви от данни, подходящо управление на човешките ресурси, стратегическа подкрепа на Big Data инициативи, готовност за неконвенционални решения и други. Необходима е нестандартна нагласа на мисленето, при която новите източници на данни не се разглеждат само от позицията на тяхната „представителност“.

Следвайки представените идеи относно качеството на „големите данни“ и основавайки се на опита на автора при практическата реализация на Big Data – проекта в българската статистическа практика, би могло да се говори за формиране на нов подход към качеството. Официалната статистика е с високо качество, произведена в съответствие с най-високите професионални стандарти и отговаряща на потребителските нужди. Ключовите елементи на качеството: относимост, точност и надеждност, актуалност и навременност, съгласуваност и съпоставимост ще продължат да бъдат от първостепенно значение и в ерата на Big Data. Но тяхното съдържание ще се развива в съответствие с ролята на НСИ и професионалните стандарти. В действителност, подходът за гарантиране на качеството на „големите данни“ и тяхното регулярно използване в официалната статистическа практика ще доведе до промяна на парадигмата на конвенционалното статистическо изследване.

Проведеното емпирично изследване „Извличане на информация от интернет за характеристики на предприятията (web scraping)“ е първо в Big Data практиката на Националния статистически институт. Приложените методи и техники за набиране, структуриране, обработка и анализ на тази информация предопределя неговата уникалност. Практическата полезност намира израз в няколко направления:

- Направен е критичен преглед на съществуващия в момента бизнес регистър за предприятията в страната, което е основа за неговата

актуализация по отношение на основни характеристики. По този начин са попълнени „бели полета“ и пропуски в съществуващата официална статистическа информация за регистрите.

- Подходът на изследване може да се разглежда като фундамент за разширяване обхвата на анализа с включване на показатели, които са пропуснати от регулярното изследване по една или друга причина (най-често с цел намаляване натовареността на респондентите).
- Изследването ясно очертава законовата рамка, при която може да се използват данните единствено и само за статистически цели.
- Постигнатите резултати са основание да се постави началото на ревизия и дописване на нови елементи към сега съществуващия Общ модел на статистическия производствен процес в НСИ.
- Резултатите от изследването подчертават необходимостта и от създаване на нови и гъвкави ИТ средства за анализ.
- Подходът на изследване може да се приеме за добра практика в национален и международен аспект.

В заключение може да се каже, че разглеждайки Big Data с тяхната реална стойност и значимост и връзките им с други явления, смисълът на тяхното за използване като един от възможните източници в официалната статистика изглежда по-скоро логичен, отколкото необичаен. Националните статистически институти трябва да имат фундаментални знания и да разширяват опита си по отношение на използването на Big Data в ежедневната статистическа практика и извън нея. Прилагането на принципа „количество над качество“, възприет от потребителите на Big Data, не трябва да се пренебрегва. Дори когато източниците на Big Data не се използват за получаване на нови статистически продукти, те биха могли да се разглеждат като ефективно средство за намаляване на натовареността на респондентите, при условие, че методологичните предизвикателства могат да бъдат разрешени. Използването на Big Data за съставяне на ранни показатели за важни статистики, като например данни за цените или бизнес цикъла е достатъчно сериозна опция. Прилагането на Big Data за краткосрочни прогнози също не е за пренебрегване.

В крайна сметка полетата на теорията и практиката са обединени от една цел – получаване на реални и навременни изводи от публично и лесно достъпни данни. Развитието на обществото и информационните технологии разкриват нови потребителски очаквания, обекти и феномени на интерес, за които официалната статистика не е в състояние да предостави данни. „Големите данни“ са съществена част от това развитие. В този смисъл може да се добави, че чрез използване принципите на нанотехнологиите Big Data ще могат да се метрират и това ще даде облика на двадесет и първи век.

IV. СПРАВКА ЗА ОСНОВНИТЕ НАУЧНИ ПРИНОСИ В ДИСЕРТАЦИОННИЯ ТРУД

1. Анализирана е съвременната информационна и технологична среда, която обуславя възникването и развитието на концепцията за „големите данни". В резултат на проучване на еволюцията в теоретичните постановки за тяхното естество е изведена дефиниция за същността и характеристиките на Big Data в контекста на статистическия познавателен хоризонт.

2. Обоснована е необходимостта от усъвършенстване на класификациите на източниците на „големи данни" с оглед осъществяване на мониторинг по отношение на тяхната автентичност, качество и достоверност. Въз основа на сравнителен анализ на съществуващите класификации в международната практика е разработена обобщена класификация и е направена характеристика на основните източници на „големи данни" за нуждите на статистическите изследвания.

3. Анализирана е промяната на парадигмата за класическото статистическо изследване чрез сравнителен анализ на етапите на статистическия производствен процес, базиран на данни от официалната статистика и от източници на Big Data. Аргументиран е изводът, че „големите данни" са съвършено нови източници с характеристики, различни от тези на традиционните, като промяната в характера на данните, тяхната наличност, обработка и разпространение е фундаментална.

4. Представени и теоретично обосновани са подходите за оценка на качеството на Big Data. Обсъдени са предложения и възможни решения за повишаване качеството на „големите данни" при използването им в официалната статистическа практика въз основа на натрупания практически опит в областта на Big Data и са посочени предизвикателствата пред националните статистически институти на европейските държави в този контекст.

5. Приложените методи и техники за набиране, структуриране, обработка и анализ на данните от проведеното емпирично изследване „Извличане на информация от интернет за характеристики на предприятията

(web scraping)“ определят неговата уникалност в Big Data практиката на НСИ. Резултатите от изследването са база за дефиниране на основни предизвикателства и възможни решения, за да може пълният потенциал на тези данни да се реализира в официалната статистическа практика.

V. ДЕКЛАРАЦИЯ ЗА ОРИГИНАЛНОСТ И ДОСТОВЕРНОСТ

Във връзка с провеждането на процедура за придобиване на образователната и научна степен „доктор“ по научната специалност „Статистика и демография“ декларирам:

1. Резултатите и приносите в дисертационния труд на тема „Проблеми при статистическия анализ на „големи данни“ (Big Data)“ са оригинални и не са заимствани от изследвания и публикации, в които авторът няма участия.

2. Представената от автора информация във вид на копия на документи, лично съставени справки и др. съответства на обективната истина.

3. Резултатите, които са описани и/или публикувани от други автори, са надлежно и подробно цитирани в библиографията.

Докторант:
/Галя Статева/

02.10.2017 г.
Свищов

VI. СПИСЪК НА ПУБЛИКАЦИИТЕ ПО ТЕМАТА НА ДИСЕРТАЦИОННИЯ ТРУД

А. Студии:

1. Статева, Г. (2016). Статистическите изследвания и „големите данни“: допълващи се източници или конкуренти. В: *Годишен алманах Научни изследвания на докторанти*, книга 10, с.169-190. Свищов: АИ Ценов.

Б. Статии:

1. Богданов, Б. & Статева, Г. (2016). Въздействието на големите данни (Big Data) върху официалната статистика: възможност или провокация. *Статистика*, 3, с. 9-32.
2. Статева, Г. (2014). „Големите данни“ – възможност, предизвикателство или заплаха пред официалната статистика. *Статистика*, 3, с. 71-87.
3. Статева, Г. (2016). Информация за ESSnet проект „BIG DATA“. *Статистика*, 1, с. 137-146.
4. Статева, Г. (2016). Оценка на качеството на „големите данни“ в официалната статистическа практика. В: *Годишен алманах Научни изследвания на докторанти*, книга 11, с. 448-458. Свищов: АИ Ценов.

В. Доклади:

1. Статева, Г. (2016). Информационните технологии при статистическия анализ на Data – проблеми и решения. В: *Предизвикателствата пред информационните технологии в контекста на „Хоризонт 2020“*, сборник с доклади от Юбилейна научна конференция, 7-8 октомври 2016 г., Свищов (с. 457-463). Свищов: АИ Ценов.
2. Статева, Г. (2014). Ролята на големите данни в официалната статистика. В: *Актуални проблеми на науката, образованието и реализацията в областта на приложната статистика и информатика*, сборник доклади от Национална научна конференция, 7 ноември 2014 г., София, УНСС (с. 13-22). София: ИК – УНСС.