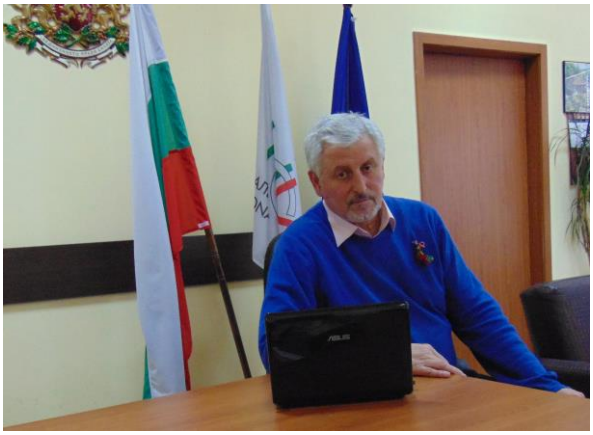
БЪДЕЩЕТО НА ИЗСЛЕДВАНИЯТА ... И ИЗСЛЕДВАНИЯТА НА БЪДЕЩЕТО: ВЪЗМОЖНИ ПРИЛОЖЕНИЯ НА ГОЛЕМИТЕ ДАННИ (BIG DATA) ПРИ ПРОИЗВОДСТВОТО НА СТАТИСТИЧЕСКА ИНФОРМАЦИЯ[1]

Богдан Богданов[*], *Галя Статева*[**]

„По-добре неясно прави, отколкото прецизно грешни.”
Амартия Сен



Въведение

Забележителни постановки в списание „Significance“: „Големите данни и големият бизнес: статистиците ще се присъединят ли?” („Big data and big business: Should statisticians join in?”) отразява значението, което се отдава на големите данни и морето от идеи, предизвикателства и опасания, съпровождащи този феномен. Наричат Big Data: следващата граница за иновации, конкуренция и продуктивност. Тази ситуация позволява на десетки хиляди анализатори (само в

* Д-р, заместник-председател на НСИ; e-mail: bbogdanov@nsi.bg.
** Държавен експерт в отдел „Обща методология и анализ на статистическите изследвания”, НСИ; e-mail: gstateva@nsi.bg.
[1] Участие на авторите при написването на статията: д-р Б. Богданов - част II и заключение; Г. Статева - въведение и част I.

САЩ са между 140 хил. и 190 хил. и броят им непрекъснато расте) да проявяват своите умения, капацитет и възможности, използвайки големите данни за разработването на задълбочени социално-икономически теории, пресъздаващи реални картини на настоящето и бъдещото развитие на обществото и света като цяло. Тяхната изследователска и научна работа на свой ред води до материализирането на теориите и възникването на нови. Това е процес, при който динамиката на еволюцията се ускорява непрекъснато. Експертите конкретизират този процес, дефинирайки, че Big Data не са универсален изход. Истинското противоборство и решение на познавателни задачи е между статистиката и компютърните науки. Това означава, че работата с източници на Big Data фокусира вниманието върху обема и скоростта на данните, новите методи за обработка, организирането на неструктурирани данни в структуриран формат и все по-нарастващият процес на автоматизация. Статистиката поставя ударение върху качеството на данните, характеристиките на извадките, валидирането на агрегациите и обобщенията, **баланса между хората и машините.** В крайна сметка тези две полета на науката и теорията са обединени от една цел: получаването на реални и навременни изводи от публично и лесно достъпни данни. В някои от стратегическите документи на Евростат отчетливо се казва, че развитието на обществото и информационните технологии разкриват нови потребителски очаквания и нови 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 на принципите на нанотехнологиите Big Data ще могат да се метрират и това ще даде облика на XXI век.

В наши дни монополът на официалните статистически институции като единствен източник на качествени данни за обществото и икономиката е приключил. Данните са навсякъде около нас, събират се все по-лесно и са евтини. Това ги прави лесно достъпен ресурс, който в допълнение е и в изобилие.

Осъзнаването на историческата възможност да се разбере, подкрепи и защити обществото чрез дигиталната информация, която то създава, трябва да бъде превърната в позитивна енергия за развитие. Не трябва да се забравя, че това обстоятелство се подсилва от факта, че на всеки две години компютърните чипове намаляват двойно размера си, а процесорите удвояват своята сложност и комплексност. С това може да се обясни експоненциалният напредък и прогресът в технологиите, който ще надхвърли прогнозите.

„Както нашият свят се променя, така и ние трябва да се променим с него.”[2] (ESS Vision 2020). Това означва още, че с непрекъснатото увеличаване на дигитализацията на процесите и явленията в света лавинообразно нараства информацията, която достига до нас. Умението да се „плува” в този океан от информация до голяма степен ще определя и статута в обществото както на отделния човек, така и на отделната институция и страна в глобалния свят. В този смисъл може да се изтъкне, че Big Data са огледало на официалната статистика. В някои случаи това огледало показва повече от официалната статистика, понякога изисква по-ясен образ, с повече детайли и нюанси. Например официалната статистика дава месечни данни за индекса на потребителските цени, а чрез използване на някои източници на Big Data той може да се изчислява дневно с поразителна точност.

Друг пример е създаването и функционирането на електронно правителство, което създава огромен терен за акумулиране на големи данни. Например в Естония действащото електронно правителство позволява ползването на 1 500 административни услуги, предоставяни от над 500 институции[3]. Сферата в тази област непрекъснато се разширява. Успоредно с акумулирането на тези данни възникват и въпросите за тяхното анализиране и използване като инструменти на доброто административно управление. Ключът към създаване на условията това да се случи е умелото и професионално използване на динамично развиващите се информационни технологии.

Един от най-големите и лесно достъпен източник на Big Data за официалната статистика са т.нар. „електронни следи“, които потребителите генерират всяка секунда, използвайки разнообразни уебуслуги. Много от тези услуги генерират данни в реално време или с много малки забавяния. Много човешки дейности, измервани от официалната статистика, са тясно свързани с поведението на хората онлайн и тази информация от интернет активността предлага потенциал за производство на статистически прогнози за социално-икономически показатели с цел подобряване навременността на статистиката. Например, когато потребителите напускат или губят работата си, те започват да търсят информация за нови работни места онлайн в различни специализирани уебсайтове, стават активни в социалните мрежи като Facebook или Twitter. Данните за уебактивността на потребителите е потенциално достъпна много бързо поради факта, че уебуслугите са изцяло електронни с много високи нива на автоматизация, свързани директно с ИТ системи. Тези данни се съхраняват автоматично в бази данни, подпомагащи уебуслугите или в лог файловете на уебсървъри. Част от събраната по този начин информация е публична (например Twitter) или се предоставя в агрегиран

[2] „As our world is changing, we have to change with it.”
[3] Петя Минкова „Губим 70 млрд. от тъпи и мързеливи чиновници”, в. „168 часа“, 16 - 22 септември 2016 година.

вид от самите доставчици на уебуслуги (например Google). Въпреки това, ако официалната статистика възнамерява да използва данните, получени от активността на хората в интернет за производство на експресни оценки за значими социално-икономически показатели, то не трябва да го прави по начин, какъвто други частни институции биха могли да го правят. Силата на официалната статистика е да използва своите предимства, изградени през годините, като единствена притежаваща гаранция за високо качество на своите данни. Преди всичко обаче, за да използва Big Data като мощен източник на данни, официалната статитика трябва да се справи с някои предизвикателства.

Големите данни са съвършено нови източници на данни за официалната статистика с характеристики, различни от тези на традиционните източници на данни. Процесът на добавяне към традиционните количествени измервания на качествени характеристики на индивидите и предприятията чрез големите данни потвърждава твърдението на много изследователи, че стойността на дадено явление може да се открие във всеки вид данни. Това включва данни от мрежата (например социалните мрежи и мобилните телефонни комуникации), текст (например Twitter), снимки, звук и видеоизображения.

Тези нови източници на данни поставят конкретни **предизвикателства** пред официалната статистика. На първо място, собствениците на данни са извън юрисдикцията на статистическите органи (например Google и Facebook). На второ място, количеството данни, което може да бъде получено от националните статистически институти от собствениците на данни е много по-голямо отколкото количеството на данни, събирани чрез традиционните статистически методи. Предизвикателствата обаче имат две основни последици: „шумът“ в информацията от интернет пространството се увеличава значително и в повечето случаи данните, които представляват интерес за статистическите органи, имат търговска стойност за доставчика на данни, дори са в основата на неговия бизнес модел (например Google и Facebook).

В противовес на това обаче новите източници на данни предлагат и няколко **възможности** за официалната статистика. Голяма част от Big Data се състоят от екстремално големи масиви от данни, които могат да бъдат използвани от националните статистически институти за производство на много по-детайлна информация (вкл. на регионално ниво за малки групи от населението), отколкото това е възможно с традиционните статистически методи. Допълнителна възможност е използването на данни, които вече са достъпни на потенциално много по-ниска цена в сравнение с цената за провеждане на едно традиционно статистическо изследване. Най-важното предимство на големите

данни си остава възможността за достъп до данни скоро или почти веднага, след като е настъпило събитието, за което се отнасят данните. Това се случва, защото обикновено източниците на Big Data произхождат от автоматизирани системи и следователно разлика във времето за събиране на данни практически не съществува.

I. Статистически коментари върху общата теория за Big Data и административните източници на информация[4]

През последните 10 - 15 години използването на административни източници се увеличи значително, така че използването на външни допълнителни източници за производство на официална статистика не е нов феномен за националните статистически институти. Освен това големите данни като алтернативен източник може да формират нови задачи и отговорности за статистическите институти. В допълнение, официалната статистика може да играе ролята на гарант за качеството на статистиката, произведена от източници на големи данни.

1. Някои примери относно събиране, съхранение и управление на неясни[5] (неструктурирани) данни

Много често Big Data се наричат „неясни данни“ (Fuzzy Data), които трябва да се преработят, за да отговарят на класическите стандарти за статистическа информация. Това е необходимата предпоставка за последваща обработка и анализ, което изисква въвеждането на модели и процедури за обновяване, „изглаждане“ и верификация на данните. Очевидно е, че в процеса на работа от съществено значение е подготовката и умението на експертите да анализират и обработват тези данни. Това означава данните да се въвеждат в система за управление, където нейните елементи позволяват решаването на познавателни задачи по отношение на тяхното качеството. Неясните данни се свързват с тяхната природа, което означава, че се получават като резултат от толеранса на измервателните инструменти или са оценки на респондентите. Например това могат да бъдат данни при измерване на околната среда или качеството на живота, където измеренията не могат да бъдат адекватно отразени еднопосочно или само с набор от числа. Обикновено измерванията в такива случаи предполагат въвеждането на определен интервал с числови или вербални значения (колкото

[4] Коментарите са направени, като се има предвид големият брой научноизследователска литература по темата за Big Data. Авторите предлагат своята гледна точка и мнение, като са използвани и авторизирани преводи. Част от тази литература е посочена в края на статията.

[5] Fuzzy Data.

се може по-малък), където се намира истинската стойност за измерваното явление или процес. В крайна сметка използването на Fuzzy Data предполага оценка на разликата за това, което се твърди, и това, което е правдоподобно. Малката разлика в това отношение дефинира и точността на данните. Може да се добави още, че това са данни, които могат да се представят чрез параметри, които са в съответстваща функционална зависимост. Това означава намиране на данни, чрез които може да се осъществи апроксимация на Fuzzy Data (да служат за основа за тяхното създаване).

Използването на новите източници на данни за статистически цели някои изследователи наричат Мрежа 2 (Web2). Определят се няколко основни направления за използването на данните за целите на официалната статистика, както следва:

- За **верифициране** на характеристиките на статистическите единици, попадащи в обхвата, и обекти на извадкови и/или изчерпателни наблюдения;
- За тематично **допълване** на статистическите изследвания;
- За възникнала необходимост от информационно осигуряване на изцяло **бели полета** в общественото пространство;
- За информационно осигуряване на явления и процеси, възникнали при **форсмажорни** обстоятелства от случаен или неслучаен характер;
- За информационно осигуряване на **нововъзникнали** явления и процеси, които постепенно заемат трайно място в развитието на икономиката и обществото като цяло.

От техническа гледна точка използването на нови източници на данни води към създаването на нова информационно-технологична структура и значими софтуерни промени. Това, от своя страна, води до съществени инвестиции за производителите и ползвателите на информация, за придобиване на знания и опит при анализа на данните. Разгледана в този аспект, постановката градира бъдещите цели за развитието на статистиката, които могат да се обобщят и етикират с наименованието „отвъд БВП (beyond GDP)”. Може да се добави, че това е една печеливша и добра платформа за интегриране на официалната и неофициалната статистика. Нейното изграждане започва с картографиране на добрите практики, създаване на критична маса чрез целеви кампании, конкуренция при визуализирането на данните, организиране на университетски програми, създаване на условия за обмен на информация, създаване и разпространение на данни. В целия процес следва да се открояват фрагментите на уникалност, съчетани с нови модели в статистиката. Това

предопределя възможността да се погледне на света по друг начин. Тогава макроикономическите показатели ще имат по-добрите оценки за растеж на обществено благосъстояние или не.

Разглеждат се въпроси за инвестиране в неосезаеми активи за разбиране на нарастването в икономиката. Идентифицират се три категории активи: компютризиране на информация; иновативна собственост; икономическа компетентност. Първата категория активи съвпада с компютърния софтуер. Втората предполага изграждането на научни знания и изобретения. Третата е свързана със стратегическо планиране, професионални обучения, изследвания, инвестиции в запазване и разширение на търговския дял и разпространение на търговската марка. В допълнение, при конструирането на инвеститорски серии проучването трябва да реши кой е подходящият дефлатор за превръщане на обема на измерванията и равнището (степента) на обезценяване при капитализиране на тези активи. Данните показват, че неосезаемите активи достигат до 14% от БВП в икономически развитите европейски страни и САЩ, като перспективите са към нарастване. Експертите считат, че този фактор ще има решаващо значение за нарастване на БВП в обозримото бъдеще. Изследванията показват, че неосезаемите активи са важни за развитието на иновациите и адаптирането на новите технологии. И отново се поставят въпросите за връзката, влиянието и параметризирането на тяхното влияние върху благосъстоянието на обществото.

2. Оценка на качеството на Big Data за целите на официалната статистика

2.1. Измерване на качеството на мултиизточници за производство на статистическа информация

Административните данни се използват от официалната статистика основно за: заместване на недостоверни данни; допълване на изцяло липсваща информация; допълване на данните от определено статистическо изследване. С появата на Big Data нарастват източниците на данни за статистически цели. В определени аспекти между административните данни и Big Data има сцепление, корелация и взаимопроникване. Задачата на статистиците е да намерят методите и подходите за комбиниране и използване по възможно най-добрия начин на възникналите възможности. Тези цели включват и опциите за подобряване на качеството на изходните данни. Нещо повече: **качеството на статистическите данни в среда от мултиресурси на информация** се превръща в един от най-съществените дискусионни въпроси в теорията и практиката на статистическия производствен процес. Кодексът на европейската статистическа практика

идентифицира пет основни принципа за качеството на статистическите продукти, а именно: **относимост; точност и надеждност; актуалност и навременност; съгласуваност и съпоставимост; достъпност и яснота.** Опорните точки на тези принципи са наличието на метаданни и доброто управление на статистическия производствен процес. Очевидно е, че всеки принцип визира специфики, които изискват конкретни и целенасочени подходи при тяхната реализация. В този смисъл принципите може да се разглеждат като фрагменти на една цялостна архитектура, определяща качеството и значимостта на статистическата дейност за обществото, управлението и държавата. Целият процес на обмяна на информация между институции в национален и международен аспект предполага преодоляването на поредица от проблеми и въвеждане на иновативно мислене за постигане на целите в условията на недостиг от ресурси, което означава качеството и производствената стойност на информационния продукт да изпреварват максимално обема на разходите (надхвърлящи себестойността му) за неговото получаване. Това може да се постигне само чрез използването на интелект, знания и ефективно мислене, с което започва темата за човешкия капитал, неговото влияние и отпечатък върху процесите. Очертават се три основни аспекта за проверка на качеството, при които статистиците имат решаваща роля за вземане на решение, както следва: наличните административни данни и Big Data **подходящи ли са за поставените цели** на статистическия производствен процес; описаните и определени в количествено измерение трансформации, при което редовете от данни ще могат **да се превърнат в статистически разпределения,** подходящи ли са за последваща обработка; създаденият статистически информационен продукт **разбираем ли е за потребителите.** Известни са и поредица от методи, които могат да се използват за отговор на тези въпроси. И към момента широко се използват техниките, подходите и методите за редактиране, импутация и калибриране на оценките от статистическите изследвания. Успоредно с това се наблюдава и все по-широкото прилагане на методите за съчетаване на данни от различни изследвания, при което се разширява признаковото пространство за единиците на наблюдение. Очевидно е, че чрез административните данни и Big Data е възможно тотално съкращаване на пътя към производството на обществено значима информация с необходимото качество на ниска цена и в съкратени срокове от време. За постигането на тези цели статистиката трябва да разработи ръководства, стандарти и средства за използването на административни данни и Big Data в различни аспекти на обществения пространство, които обхващат институциите, бизнеса и различни видове организационни структури. Като пример в това

отношение може да се посочи необходимият път, който трябва да се извърви при организацията на приближаващото се Преброяване 2021.

2.2. Профилиране на Big Data за оценка на тяхната селективност

Онтологията на Big Data и глобалната икономика трудно могат да се датират точно и да се определят, но забележимо нарастващият обем от информация започва с динамичното разрастване на информационните технологии в началото на 90-те години на миналия век. До голяма степен легитимността на големите данни се дължи на създалата се възможност за тяхното използване при производството на статистическия информационен продукт. Тази възможност се превръща в необходимост, тъй като се осъзнава потенциалът на една информационна среда, създадена от икономическата и социалната дейност на човешкото общество. Успоредно с това се появяват проблемите, които очертават спецификите на непрекъснато създаваната информация. Отрицателните характеристики, очертаващи особеностите на тази информация, са следните: първо - **отсъствието на релевантност** по отношение на изследователските задачи; второ - повечето Big Data, които са достъпни, се композират от събития и обикновено **осигуряват много малко информация или тя отсъства напълно за единиците, които генерират данни;** трето - ако информацията е достъпна за създателя на данни, може да се окаже, че **достъпът не е много лесен** поради специфики на личността или компанията; четвърто - **нарушена е представителността на изследваната съвкупност от Big Data** поради невъзможност да бъдат включени всички единици от целевата съвкупност, които изследователите искат да изучават.

От изброените характеристики от съществено значение е темата за представителност на изследваната съвкупност, източникът на която може да бъде Big Data. Като класически пример в тази постановка може да се вземат изследванията на социално-икономическите индикатори чрез социалните медии. От направени проучвания е установено, че само около 70% от населението са активни в тях. Проучването на тази съвкупност в съответствие с теорията на извадковите изследвания **не може да се приеме за представително,** въпреки че степента на покритие е висока. На практика отсъства вероятностният подбор за формиране на извадка. Освен това не може да се очаква, че всички лица ще вземат участие в изследването поради желанието да се спазват правилата за конфиденциалност на лични данни. Успоредно с това обаче тази ситуация се приближава до един съществен проблем на извадковите изследвания изобщо, а именно наличието на неотговорили в

процеса на формирането на случайната извадка. Очевидно е, че без корекция за селективност на единиците за наблюдение **оценките от такова изследване ще бъдат изместени.**

Общият метод за оценка на селективността в извадковите изследвания е чрез **сравнение на разпределенията на релевантни основни характеристики в източниците на данни с техните известни разпределения в целевата съвкупност.** По принцип този подход може да се използва и за Big Data. Съществува идеална ситуация, при която изучаваните единици са свързани с регистъра на населението, съдържащ основни техни характеристики (като пол; възраст, образование; местоживеене; семеен статус и т.н.). Емпиричният опит показва, че голяма част от единиците притежават детерминирана връзка с регистъра. Това позволява интересуваща ни информация да бъде извлечена от източниците на Big Data по известни признаци на изучаваните единици. Този подход се нарича „профилиране". По друг начин казано, този подход позволява - например на базата на корелиращи признаци една неструктурирана съвкупност от социалните медии да може да се структурира, като се използва регистърът за обекти, като се поддържа и актуализира в съответствие с утвърдени статистически стандарти. По аналогичен начин може да се разглежда статистическият бизнес регистър, като съответно за предприятията основните характеристики са: брой на заетите; отрасъл; размер, оборот и т.н. Процесът на създаване на най-добрите профили чрез корелиране на характеристики от регистрите и събраните големи данни съществено може да подобри едно статистическо изследване по теми, съдържащи предварително добре дефинирани познавателни задачи.

Big Data са предизвикателството към официалната статистика, но успоредно с това могат да запълнят успешно съществуващата *пукнатина* между теорията и практиката на статистическите изследвания. През фокуса на Общия модел на статистическия производствен процес (GSBPM)[6] се наблюдават три основни фази на статистическия производствен процес, осъществен с Big Data, както следва: достъп и въвеждане - кореспондира със събиране на данни; обработка и съхранение - кореспондира с анализ на данните; изходни данни, свързани с разпространение на данните. Тази йерархическа структура включва три основни аспекта: източници на Big Data; метаданни за Big Data; и изходни данни. Трите аспекта кореспондират също със съществуващите административни данни, които вече успешно се използват за статистически цели. За всяка фаза и аспекти са характерни различни изисквания за качество на данните. Очевидно е, че при работа с Big Data се съкращават всички фази, изискващи провеждането на теренни статистически изследвания за събиране на

[6] Generic Statistical Business Process Model.

първични данни. Именно това обстоятелство по съвършено различен начин поставя изискванията за качество на данните.

Процесът за оценка на качеството на източниците на Big Data може да се представи като акредитация (стандартизиране, достоверност). Принципите на акредитацията могат да се представят със следните ключови думи: осигуряване на достъп до източника на данни и непрекъснатост във времето; гъвкавост при дефиниране на познавателни задачи; прагматичен подход; емпирична експертна оценка; оценка на качеството на единиците за изучаване; инкорпориране на комбинация от общи правила за работа, контрол и оценка.

Трябва да се отбележат някои важни предимства на Big Data: по-големият брой единици, които могат да се изследват в сравнение с принципно ограничените по обем извадки от традиционните изследвания; генерират се без човешка намеса, т.е. отсъства отпадането на единици за изследване; сравнително евтини са и се генерират в реално време. Всичко това ги прави атрактивен източник на данни за регулярното статистическо производство.

Източниците на големи данни, и по-специално източници на данни от уебдейностите представят някои предизвикателства по отношение на спазването на принципите на официалната статистика (например Кодекса на европейската статистическа практика). Като допълнителни външни източници те са извън контрола на Националния статистически институт (НСИ). При традиционните източници на данни НСИ има пълен контрол върху данните от изследванията или поне има влияние върху качеството на административните данни.

Липсата на контрол върху Big Data създава някои рискове. На първо място, съществува риск, че източникът на данни е една черна кутия. НСИ полага непрекъснати усилия да документира, колкото е възможно по-пълно статистическия производствен процес. Тази прозрачност е задължителна и необходима, за да се поддържа задоволително ниво на доверие в официалната статистика от страна на обществото и политиците. При източници на големи данни, чиито държатели са частни единици, обикновено не е възможно да се гарантира същото ниво на прозрачност както при официалната статистика. Понякога дори е възможно разкриването на обработката на данни на уебуслугите да постави доставчика на данни в неизгодно конкурентно положение на пазара.

На второ място, дори ако НСИ старателно проверява обработката на данни на уебуслугата, не може да се гарантира, че източникът на данни не е бил обект на манипулация. А понякога пълна и

коректна проверка на данните не е възможна (ако доставчикът на данни е извън юрисдикцията на статистическия орган) или изисква голям финансов ресурс.

На трето място, източникът на големи данни може да не е постоянен във времето и това да доведе до прекъсване на динамичните редове с голяма честота. Например след своя старт през 2006 г. Google Trends направи няколко ревизии в своите алгоритми, влияещи върху непрекъснатостта на динамичните серии на Google.

На четвърто място, има риск от липса на непрекъснатост на данните от Big Data източника, тъй като НСИ не е в състояние да гарантира, че източникът ще е достъпен толкова дълго, колкото е необходимо за производство на статистически продукти. Полезността на данните, идващи от конкретни уебуслуги - например от една интернет търсачка, зависи пряко от тяхната популярност, която се променя във времето. Наличието на източника може да се наруши също и от технологични промени, които отново не са под контрола на НСИ.

Някои от тези рискове могат да бъдат намалени чрез комбиниране на данни от няколко уеб-услуги, прилагайки различни модели за прогнозиране. Това би довело до намаляване на влиянието на отделните източници (които НСИ не контролира) в прогнозните стойности и по този начин се предоставя гаранция, че официалните „експресни“ оценки са достатъчно надеждни. Разнообразието от източници също ще позволи да се подобри евентуалната липса на приемственост на някои от източниците. Например в случай на „експресна“ оценка на нивото на заетост възможен източник може да бъде броят посещения на интернет страници, свързани със заетостта. Като обобщение може да се каже, че е необходимо създаване на процедури за акредитация/оценка на качеството и сертификация на източниците на Big Data за официалната статистика с цел гарантиране на прозрачността и повишаването на качеството на тези източници.

Правейки паралел между използването на административни данни в производството на официалната статистика и потенциалното използване на големи данни, може да се заключи, че:

- Данните са навсякъде около нас и макар да са произведени за нестатистически цели, те могат да се окажат важен допълнителен източник в официалното статистическо производство. Ние, официалните статистици, може да бъдем насърчавани и мотивирани да търсим данните и да ги използваме като инструмент за подобряване на традиционните статистически изследвания.
- В същото време ние като официални статистици трябва да сме здраво стъпили на земята и да бъдем селективни при интегрирането на нетрадиционни източници на данни в официалното статистическо производство. Отправна точка при вземането на решение биха могли да

бъдат отговорите на два важни въпроса: ще бъде ли достъпен даден източник на Big Data в бъдеще и има ли гаранции, че може да се произвежда официална статистика чрез него достатъчно дълго време; информацията, която се извлича от огромните налични масиви Big Data наистина ли е сигнал за някаква значима тенденция или е просто „шум“ и ако това е сигнал, измерва ли се чрез него социално-икономическият феномен, който искаме?

II. Добри практики при използването на Big Data и административните източници на информация за статистически цели

1. Оценка на гъстотата на населението в Европа чрез мрежа от мобилни оператори

Всеки ден милиони хора по света използват своите мобилни телефони. Те се ползват както за лични потребности, така и за развиване на бизнес начинания. Тези малки устройства се превръщат в жизненоважна част от живота на всеки човек. Чрез тях той може успешно и за много кратко време да управлява десетки процеси на ден от големи разстояния. Нещо немислимо преди 10 - 15 години. На практика тези устройства извършиха революция в общественото развитие. Динамиката на тяхното усъвършенстване е драстична и въвежда все нови и нови измерения в дейността на човека. Потоците от информация също се ускоряват и стават неотменна част от човешкия живот. Една част от тази информация се създава от самите собственици на мобилни телефони, друга се създава от многобройни сензори, устройства, анализатори, експерти, институции и организации в национален и международен аспект. Използването на тази информация от местните и социалните органи на властта също е важно, тъй като се създава добра представа за разпределението на населението по определени признаци и се разработват стратегии за позитивни влияния. Природните бедствия като наводнения и земетресения са случаите, когато тези данни ще дадат възможност за по-точно планиране на определени действия за минимизиране на проблемите.

Основното предизвикателство при изследване **на гъстотата на населението в Европа с помощта на мрежа - базирана на основата на мобилни телефони,** е създаването на методологическа рамка, включваща начина и организацията за получаване на данни чрез извадка от отделните мобилни оператори (Mobile Network Operator (MNO). Дългосрочната цел на това изучаване е да се поставят основите за нови, разширени и различни изследвания в мобилните мрежи. Тези цели са изпълними и важни за общественото развитие като цяло. Методологическото развитие на изследването може да се разгледа по-конкретно в следните основни аспекти:

- Използване на разширена мрежа от типологични данни: включване на цялата радио-мрежа, не само локациите на антени.
- Надхвърляне и повече от данните за подробните записи на повикванията (Call Detail Record (CDR): включване на други сценарии с различни комбинации - например получаване на данни от CDR и/или от регистъра на посетителите по местоположение (Visitor Location Register (VLR).
- Multi-MNO: разработване на проект (дизайн) на сливане и премахване на границите между различните оператори, извършващи услуги във всички страни.

Следва да се отбележи, че изследователите разработват инструментариума от гледна точка на неговото използване в световен аспект. Изследването се основава на логичния строеж на мобилната мрежа - всеки мобилен телефон е включен в различни многобройни радиоклетки, които са на различни разстояния от оператора. Главната част от работата се фокусира върху детайлите на записите от обажданията (CDR) и се основава на извадка от мрежата на мобилните оператори (MNO). Следва да се отбележи, че съществува разлика между мобилните станции и лицата, когато се оценява гъстотата на населението. На практика едно лице може да притежава повече от един мобилен апарат. Съществуват и лица без нито едно мобилно средство, макар и рядко да се наблюдават такива случаи. Тези предпоставки са основание за изместване на оценките за разпределението на населението по определени територии. Централен момент при оценяването на гъстотата на населението е дефиниране на понятието „гъстота”. В този смисъл се дефинират определени аспекти на това понятие, както следва: пространствена гъстота; вероятностна гъстота; времева гъстота.

Методологията на изследването се основава на два вида данни, а именно: мрежа от топологически данни относно географската локация и покритите области от радиоклетки; данни от мобилните станции. Основните източници на данни са CDR и VLR. Следващата важна стъпка е оценка и верифициране на корелацията между структурите на мобилните станции като локация и активност. Спазването на тази постановка е от съществено значение, тъй като нейното пренебрегване води до подценяване или надценяване на оценките за гъстотата на населението във времето и пространството. В крайна сметка проблемът се изразява в недостъпността на данни в отделните клетки (CDR и VLR). Това налага включването на презумпцията, че съществува фундаментален компромис между пространствената точност и риска от изместване. За общото идентифициране на клетките (Cell Global Identifier (CGI) може да се използва системата за общо

позициониране (GPS), за да се открие бързо локацията на мобилното средство. Този подход също е придружен с известни трудности и недостатъци.

Следователно може да се обобщи, че използването на големите данни с използването на мобилните телефони е един революционен подход за създаването на модерна статистика с иновативни средства независимо от рисковете и недостатъците. Очевидно е, че тези оценки следва да се верифицират с други източници на информация, включително и данни от официалната статистика (например данни от преброяванията и текущата демографска статистика, които могат да се използват за изчисляване на гъстотата на населението за периода, към който се отнасят). В този смисъл Big Data при получаване на данни чрез използването на мобилни телефони следва да се разглеждат, от една страна, като алтернативен източник на информация, но от друга, като източник на информация, който не изисква средства за неговото активиране, а също така, който може да се използва с по-голяма честота и когато е необходимо.

2. Използване на нова база на иновации и растеж: картографиране на общественото мнение, нагласи и поведение чрез източниците на Big Data

Участниците в социалните мрежи генерират непрекъснато данни за тяхната дневна активност и настроения. Във всяка точка на времето и пространството се създават хиляди индивидуални данни, които измерват пулса на обществото. Погледнато в този аспект това е информация, която отразява развитието на обществото. По този начин следват и по-нататък създават представата за развитието на света като цяло. Така например мобилните телефони се използват не само за индивидуални разговори, но също за постигане на съгласие по делови въпроси, трансфер на пари, търсене на работа, купуване на стоки и услуги, търсене на медицинска информация и т.н. По принцип Big Data изискват прилагането на мощен алгоритъм, който е способен да разкрива модели, трендове и корелации между данните от различни времеви хоризонти, а също и с използване на модерни средства за визуализация. Въпросът е: какъв е капацитетът на Big Data за анализа? Отговорът на този въпрос не е лесен, тъй като трябва да се разработи модел и да се разкрият насоките за развитие, очертавани от данните, получени чрез наблюдение и съпоставяне от различни по вид информационни източници. Тези източници, от своя страна, изискват дефиниране и обща рамка на информационния процес, което означава стандартизиране и управление. Това обстоятелство изисква най-малко създаването на своеобразен въпросник или лексикон, който е специфициран, като се отговори на най-съществените въпроси:

- Какво: определя се типът на информацията, съдържаща се в данните;
- Кой: идентифицира се авторът на данните;
- Как: идентифицира се източникът, чрез който данните са придобити;
- Колко: отбелязва се дали данните са количествени или качествени;
- Къде и кога: уточнява се географската локация, както и времето, през което са придобити данните;
- Накъде: очертават се трендовете, хипотезите и насоките за развитие.

В процеса на аналитичната работа се следва процедура, която гарантира качеството и достоверността на коментарите и изводите:

- Филтриране: запазва се отстояние от наблюдението и се отхвърлят нерелевантните парчета от информацията;
- Обобщаване: екстрахират се ключови думи и поредица от ключови изрази (аналитични фрагменти, конструиращи архитектурата на текста);
- Категоризиране: включване на данни и поредица от подходящи измерители, подсилващи текста и параметризиращи анализираните явления и процеси;
- Рамкиране: определя се мащабът на анализа, акцентите и границите.

Изкуството да се задават въпроси предопределя възможността да се търсят и намират правилните отговори и да се материализират правилните решения. Обратно, очевидно е, че при задаване на неправилни въпроси, умен отговор почти е невъзможно да се получи. При търсенето на правилния (умен) въпрос и отговор, решение и действие се следва определена логика: започва се с *извличане на информацията от мрежата (например: данни от социалните мрежи в информационното пространство); усвояване (асимилиране) на данните, превръщайки ги от неструктурирани в структурирани; съхраняване на данните за реалния отрязък от време.*

3. Big Data: изследователи в Google прогнозират безработицата във Финландия

Прогнозите за безработицата се налагат, тъй като публикуването на данните от официалната статистика закъсняват, при което параметризирането на явлението се отдалечава повече или по-малко от момента на неговото проявление. Тези прогнози се оказват от съществено значение в моменти на икономическа криза, тъй като дават възможност за бързи решения за редуциране на безработицата. На практика публикуването на данните се осъществява в реално време, което

определя тренда за развитието на процеса. С други думи, интернет придобива важна роля по отношение на пазара на труда. Прилаганите модели и тяхното тестване демонстрират полезността на големите данни.

Изследователите във Финландия използват интернет[7], за да получат данни за търсенето и предлагането на работна ръка на пазара на труда. По същество това са времеви редове от обяснителни данни за т.нар. „бизнес сантиментален индекс“ (конфиденциален индикатор на заетостта) и данни от онлайн търсенето на предложения за работа в Google. За целите на изследването се прилага моделът ARIMAX. На тази основа се използват месечни данни от януари 2004 до 2012 г., за да се направи прогноза за 2013 година. Анализът на времевите редове за безработните се осъществява чрез добре известния метод Box-Jenkins. Резултатите от изследването показват, че бъдещите изследвания на пазара на труда могат и трябва да се осъществяват чрез обработката на Big Data в тази област. Този извод се налага не само като иновативен подход, но също и като реалистичен изход от бюджетните рестрикции на статистическите изследвания в условията на икономическа криза.

4. Научените уроци от уебуслугата „Google Trends“

През 2006 г. компанията Google стартира уебуслугата „Google Trends“, предоставяща данни за това колко често някои специфични термини са търсени в търсачката на компанията за даден период от време. Първоначално Google Trends се е използвал за идентифициране на онези термини, които създават някаква тенденция, т.е. термини, за които постоянно се увеличава броят на наблюдаваните търсения. Точно тази висока навременност на Google Trends подхрани значителен брой научни изследвания, посветени на използването на Big Data източника, за да се прогнозират социално-икономически показатели с цел получаване на бързи резултати, изпреварващи значително публикуването на официалните статистически данни. Самата компания Google през 2009 г. публикува в своя изследователски блог един от първите опити да се прогнозират социално-икономически показатели, базирани на данни от Google Trends. Използват се данни от търсенето за производство на краткосрочни прогнози за няколко показателя: продажби на автомобили, продажби на дребно, продажби на жилища и брой посетители. Като резултат от изследването се налага изводът, че за прости авторегресионни времеви модели въвеждането на данни от търсенето като предиктори увеличават тяхната прецизност по отношение на краткосрочните прогнози. Тъй като

[7] В Испания също се изследват възможности за оценка на безработицата, като се използва интернет.

данните от Google Trends са с много висока степен на актуалност - на практика няколко дни след референтния период, то тези модели биха могли да се използват за прогнози почти в реално време. Други проучвания - например за грипните епидемии в световен мащаб, безработицата в определени райони или частното потребление, също са били обект на прогнози чрез данни от търсенето в Google Trends. Опитът с приложението Google Flu Trends, предсказващо грипните епидемии, е един от „научените“ уроци за внимателното използване на данни от интернет търсачка за производство на експресни оценки. През 2008 г. стартира Google Flu Trends приложението, което използва обобщени Google данни за търсене, за да оцени грипната активност в САЩ с по-висока актуалност и навременност отколкото официалният показател, изчисляван от Центровете за контрол и превенция на заболяванията (CDC). През периода 2009 - 2013 г. прогнозните резултати от Google Flu Trends са добри. Въпреки това през 2009 г. прогнозните резултати спрямо официални данни на CDC се изразяват в подценяване на честотата на грипни заболявания. Това събитие се приписва на промени в поведението на хората при търсене в интернет мрежата, което от своя страна, довежда до преразглеждане на Google Flu Trends алгоритмите. През 2013 г. по време на пика на грипния сезон оценките на Google Flu Trends са почти двойни като стойност на по-късно публикуваните реални данни от CDC. Възможната причина за тази значителна разлика между експресни оценки и реални данни, която се изтъква тогава, е широкото медийно отразяване на тежкия грипен сезон през 2013 година. Това генерира до известна степен негативна реакция срещу използването на големите данни и появата на твърдения сред научната общност, че все още не може да се произведе надежден статистически продукт чрез Google Flu Trends. Част от трудностите в процеса на прогнозиране на грипни епидемии чрез Google Flu Trends се обясняват с честите промени на алгоритмите за търсене, въведени от инженерите на Google, които имат влияние върху обратните резултати към потребителите и върху начина, по който потребителите изпълняват многократни търсения. Тази нестабилност на изходните данни променя валидността на прогнозния модел и би могло да се изисква неговото динамично калибриране.

Прозрачността е един от основните принципи на официалната статистика, който е необходим за правилното тълкуване на официалните статистически данни от потребители и изследователи, а в примера по-горе думите за търсене, които се използват за изготвяне на експресни оценки, дори не са известни. Вземайки предвид примери като този, става ясно, че големите данни все още не могат да заменят всички традиционни методи за събиране на данни. Ключът за извличане на „добавена“

стойност от големите данни за официалната статистика е интеграцията им в статистическия производствен процес чрез различни източници.

Други подобни източници за прогнозиране на социално-икономически показатели са посещенията в социалната мрежа Twitter и страницата на Уикипедия. Например броят на посещенията в Уикипедия е използван като адекватен измерител за прогнозиране на грипоподобни заболявания в САЩ и в сравнение с Google Flu Trends (през периода 2012 - 2013 г.) дава по-точни резултати за прогнозиране на пика на грипната епидемия.

Друг пример е използването на Twitter за прогнозиране на официалната статистика. В едно проучване са оценени международни и вътрешни миграционни модели от позиционирани географски данни за около 500 000 потребители на Twitter. Заключението е, че разработените методи могат да се прилагат за прогнозиране на повратните точки в миграционните тенденции, като по този начин ще се подобри разбирането на взаимовръзката между вътрешната и международната миграция.

5. Сканиране на данни и използването на интернет като източник на данни за статистиката на цените

В редица европейски страни се осъществява сканиране и използване на съвременни технологии за извличане на информация за оборотите, вида и количествата на продадените стоки и услуги. Сканираните данни се генерират от касовите устройства в магазините и представляват информация за оборота на магазина - количествата, продадени стоки по GTIN (Global Trade Item Number), познат в миналото като EAN номер или по друг начин казано - като баркод за определен период. От тези данни статистическите служби извличат единична цена за продукт. Към настоящия момент шест европейски статистически служби **използват сканирани данни за конструирането на индекси на инфлацията:** Норвегия - от 1995 г., Нидерландия - от 2002 г., Швейцария - от 2008 г., Швеция - от 2012 г., Белгия и Дания - от 2016 година. Процентът продажби на супермаркетите, за които статистическите офиси получават информация, е следният:

- 60% в Дания, като скоро се очаква да достигне 80%;
- 75% в Белгия и Швейцария;
- 90% в Нидерландия и Швеция.

Web scraping е техника, при която автоматично се събира информация от цялата информационна мрежа (World Wide Web), като за целта се използват специални устройства („паяци“,

интернет роботи, „гъсенични трактори“ и т.н.). Опити за извличане на информация по този начин, която да се използва като допълнителен източник при производството на статистика, се правят в редица страни, включително и в България в НСИ.

Целта на тези подходи е получаване на информация за индекса на потребителските цени (и не само). Този иновативен подход постепенно ще изпрати в миналото съществуващите статистически изследвания, при които чрез анкетьорски екипи по места се регистрират цените на стоки и услуги от магазинната мрежа. Предимствата на този подход са следните:

- **Съкращаване на разходите** по изработването на информационния продукт;
- **Съкращаване на сроковете** за предоставяне на данни на потребителите;
- **Съкращаване на човешката намеса** чрез въвеждане на новите информационни технологии и в процеса на производство на информация, като по този начин обективността, реалистичността и достоверността на информацията рязко нараства.

Очевидно е, че тази информация ще помогне на бизнеса да **разработи по-реалистична и гъвкава стратегия** в бизнес плановете за развитие. Когато информацията се съчетае с данни за доходите и разходите на домакинствата, ще може **да се оцени *поносимостта* на цените на стоките и услугите върху техните бюджети** при равни други условия**.** По този начин бизнесът ще има много по-ясна представа за търсенето и предлагането на пазара - един от основните закони на пазарната икономика.

6. Опит на Евростат при изчисляване на експресни индикатори

Експресните оценки на хармонизирания индекс на потребителските цени (ХИПЦ) за еврозоната по основни компоненти е статистически продукт, който се произвежда месечно и е един от най-значимите показатели, произвеждани в Евростат. В края на всеки месец се публикуват оценки на инфлацията през този месец. Експресните оценки на инфлацията са важен показател за широката общественост, финансовите пазари, както и за Европейската централна банка с цел формулиране на адекватна парична политика в еврозоната. Следователно, необходимо е гарантиране на високо качество на данните по отношение на точността, навременността и непрекъснатостта на публикуваните данни.

Експресните оценки на ХИПЦ за еврозоната са комбинация от предварителни данни, изпратени от някои държави членки с прогнозни данни за останалите страни. Предварителните данни се основават на същите събрани цени, от които се получават окончателните ХИПЦ индекси, така че макар и не

валидирани, се оказват много точни и са за предпочитане пред каквито и да са прогнозни модели. За съжаление, не всички страни могат да предоставят предварителни данни навреме: за тези страни е необходимо да се прогнозират липсващите данни. Различни основни компоненти на инфлацията имат много специфично стохастично поведение и са много нестабилни и трудно предсказуеми. Поради тази причина всеки компонент на инфлацията се третира отделно и всякакви допълнителни данни, които могат да подобрят прогнозата, са от важно значение. Например като допълнителен източник за цените на енергията при изчисляване на експресните оценки се използват данните от седмичния бюлетин за горивата, публикувани от Генералната дирекция по енергетика на Европейската комисия (DG ENER) - административен източник на данни. Бюлетинът съдържа референтни цени на енергийни продукти, които имат много по-висока корелационна връзка с цените на енергията отколкото цените, които плаща средният потребител. Макар че този бюлетин за горивата е създаден да обслужва изцяло нестатистически цели, в днешно време се прилага успешно като допълнителен източник на информация при производството на експресни оценки за ХИПЦ.

Използването на административни данни в случая е рационално поради няколко причини: Евростат не изразходва допълнителни финансови ресурси за събиране и оработка на данните - взема ги директно от административния източник; DG ENER регулярно публикува данни, т.е. има непрекъснатост на административния източник и освен това чрез интернет мрежата са лесно достъпни за всеки, който иска да ги използва.

При изчисляването на експресните оценки на ХИПЦ за еврозоната не се използват източници на Big Data. Въпреки това необходимостта от допълнителни административни данни (както беше посочено по-горе), за да се преодолее проблемът с липсващи предварителни данни за някои страни, може да служи като пример за една възможна употреба на Big Data в регулярното производство на официалната статистика.

7. Моделиране на извадкови данни от „умни“ електрически измервателни уреди

Умни измервателни уреди са електронните средства, чрез които се правят записи и се съхранява информация за ток, газ, вода на чести интервали от време. Те се включват по определен начин към битовата инфраструктура в жилището. Тези данни се изпращат онлайн в бази данни и се използват за мониторинг, документиране на процеса и дефиниране на нови цели. През последните години в много страни се наблюдава тенденция за използване на умните устройства при измерване на потреблението на електричество и газ. Разбира се, данните от тези устройства са особено

атрактивни за официалните статистически организации, защото ще осигурят детайлна информация с голяма честота за потреблението в домакинствата. В статистическия офис на Обединеното кралство такива данни се използват, за да се моделира дневното потребление на домакинствата и да се направят оценки за степента на консумация по вид и време. На практика по този начин се осъществяват милиони наблюдения, без пряката намеса на човека, като се осигурява качествена информация в реално време.

8. Дефиниране на обичайна среда с мобилно позиционирани данни

Туризмът е дефиниран като активност на хората да пътуват и остават на места извън тяхната обичайна среда за не повече от една последователна година за отдих, бизнес или други цели, което не е свързано с активно занимание срещу заплащане на посетеното място. В този процес Big Data са източникът, подходящ за изследователите по-лесно и автоматично да обвържат данните за ежедневното движение на хората. В Естония се използват записите на мобилните оператори, интернет и услугите за данни, за да се оцени активността на местния туризъм. Сравненията между данните на официалната статистика за туристическите пътувания и данните от мобилните оператори показват силно изразено сцепление между тях. Сравнителните анализи са съществен момент при верификацията на тенденциите, но проблемът за качеството на данните и при двата източника винаги трябва да бъде обект на допълнително внимание и методологически коментар. Очевидно е, че когато се използват повече от един източник на данни, по принцип по-релефно се оценява тяхната достоверност и точност, т.е. нарастват възможностите за редуциране на изместването на оценките.

Развитието на глобалната икономика и разширяването на ЕС създават реални предпоставки за увеличаване на възможностите и ускоряване на движението на населението в различни направления. Развитието на информационните технологии и техните средства се превръщат в ежедневна професионална и битова потребност на хората по света. Промените са динамични, процесите в различни сфери на икономиката и обществото също. Възможността за менажиране на тези процеси е възможно чрез активното използване на данните от средствата на информационните технологии. **Синхронизацията между динамиката на промените и адекватните реакции при тяхното управление** е сериозното предизвикателство на XXI век. Това се явява една от първостепенните задачи на ръководителите от Европейската статистическа система за развитието и стратегията за интегриране на Big Data в официалната статистика.

Създават се научноизследователски центрове за развитие на теорията и практиката за достъп и използване на Big Data за целите на официалната статистика. Такива центрове вече са създадени в Германия, Франция и други страни. Изследователите знаят, че този източник на данни напълно ще промени света на статистиката по отношение на начините за събиране, обработка и анализ на информацията. Практически Big Data стават съществена част от развитието на бизнес средата в общественото пространство.

9. Използване на Big Data като източник на данни за изследване на домакинските бюджети

В Норвегия започва използването на Big Data за изучаване на домакинските бюджети. Установено е, че традиционното изследване на домакинските бюджети, което се осъществява чрез попълване на дневници, може да бъде сполучливо заменено от други източници на данни. Тази идея се материализира на основата на комбинация от дигитални данни за осъществените сделки и различните видове въпросници. Целта е да се екстрахира консумативен модел от база данни от магазинната мрежа и свързването на резултатите с извадково изследване. По този начин се противодейства на изместването на оценките от изследването, тъй като неотговорилите влияят негативно[8]. В този аспект се разработва **дизайн на мултимодел, който включва различни източници на данни и формира уеббазиран дневник.** В дизайна ключовият елемент са метаданните, дефиниращи различните източници на данни. Забележим източник на данни в този дизайн са трансакциите от дебитните и кредитните карти, които са подбрани от разплащателната система на магазините и са електронно трансферирани към дневниците на респондентите, където се отбелязват дневните разходи. Съществен момент при използването на този модел е обстоятелството, че само 4% от разходите на домакинствата се осъществяват с пари в наличност (кеш). По същия начин предстои получаване по електронен път на отчетите за изтеглените и внесените парични средства. Стремежът е бюджетът на домакинствата да се разработва, като максимално се използват съвременните информационни технологии и тяхното приложение в бита на обикновените домакинства. При тези обстоятелства отпадат редица технологични процеси, свързани с интервюиране на респондентите. На преден план излизат реални дейности, които са документирани факти от бита на домакинствата и тяхната достоверност не подлежи на съмнение. Успоредно с това забележимо се намаляват разходите за получаване на необходимата информация. Важен проблем

[8] Неотговорилите в изследване на домакинските бюджети в Нидерландия е около 50%.

при изследването се откроява отново нежеланието на респондентите да участват в наблюдението. Забележимо е, че предимно хора от младите възрасти в повечето случаи са съгласни да предоставят свободен достъп до личните си финансови трансакции[9].

Дотук бяха посочени няколко примера за прогнозиране на разнообразни социално-икономически показатели, като повечето от тях не са правени от официалните статистически служби (национални статистически институти, Евростат и други международни официални статистически служби). Тогава възниква един основателен въпрос: защо е необходимо официалната статистика да прави сама неща, които другите могат да правят? В тази статия авторите не се опитват да отговорят на поставения въпрос. Това, което твърдим, е, че ако официалната статистика изчислява експресни оценки за някои социално-икономически показатели, като прилага прогнозни модели, базирани на уебданни, тогава защо същите тези официални статистически организации да не „консумират“ своите специфични предимства пред другите фирми и организации на частния пазар?

Както неколкократно вече беше подчертано, най-очевидното предимство на официалните статистически служби е, че те произвеждат официалните статистически показатели, които имат „печат“ за качество съгласно Кодекса на европейската статистическа практика. Друго важно предимство е богатият емпиричен опит в провеждане на изследвания, включително наличието на добре работещи традиционни статистически експерти и утвърдени системи за събиране на данни. Заради всичко това официалната статистика трябва да интегрира производството на експресните оценки в своите регулярни статистически производствени системи, което означава използване на възможно по-детайлна информация за показателите, отколкото това, което се публикува. Традиционните изследвания може да се адаптират така, че за тяхното обогатяване да се използват ефективно данни от уебактивността на населението или да се коригира изместването на оценките, произлизащи от големите източници на данни.

Заключение

Наричат Big Data: следващата граница за иновации, конкуренция и продуктивност. Очаква се бизнесът и свързаните с него информационни технологии да нарастват с 1.3% всяка година от 2010 до 2020 година. Основно свързаните с този процес професии са на учени (изследователи) и статистици, но не се знае дали в бъдеще ще съществува разлика между тях. Те нарастват всяко десетилетие с 15%. Big Data се определят като серия от данни отвъд (прехвърляща) способността на

[9] От лятото на 2014 г. в Норвегия се провежда наблюдение на извадка от 1 082 респонденти.

типичните средства (устройства) да събират, извличат, управляват и анализират. Безпристрастната оценка за ползата и използването на големите данни се губи в хиперпространството. В този смисъл може да се приеме изречението: „*Надеждата е, че ако вие измервате икономическия пулс в реално време, вие ще бъдете способни да отговорите на аномалиите по-бързо*[10]*"*. Тази постановка може да се допълни и с още едно обстоятелство, свързано с развитието на човешката мисъл и умения. Те трябва да бъдат в синхрон и да изпреварват възможностите на информационните технологии. Те трябва да бъдат креативни към разработването на подходящия софтуер и хардуер, за да посрещнат лавината на големите данни.

Вместо послеслов

Написаното в тази статия трябва да се разглежда като ***анонс*** към една бъдеща политика на НСИ основно в две направления:

1. Постепенно **пренасочване на финансови и човешки ресурси** от традиционните статистически изследвания към работа с Big Data. На практика това означава:

- **Изграждане на екип от експерти с високо ниво на знания и умения** в различни области, което означава още и създаване на модел за различен начин на мислене и мотивация за работа от съществуващите в момента.
- **Промяна в начина на изследване и анализ на явленията и процесите** в заобикалящата ни действителност. Това означава модернизация на съществуващия общ модел на статистическия производствен процес чрез въвеждане на иновативни подходи за получаване, обработка, визуализация и анализ на данните.
- **Създаване на център** за развитие на кохерентността между теорията и практиката на Big Data и извадковите статистически изследвания.

2. Повишаване на качеството на реформата в институцията като професионален мениджмънт. На практика това означава:

[10] *„The hope is that as you take the economic pulse in real time, you will be able to respond to anomalies more quickly."* Hal Varian, Google Chief Economist (Professor Emeritus, University of California, Berkeley).

- Повишаване на професионалното **качество на структурните единици** при наскоро направената реформа[11] в НСИ.
- Създаване на **гъвкави и всестранни връзки** със собствениците на Big Data и институциите, отговорни за създаването и съхранението на административни данни.
- Актуализиране на **връзките с потребителите** на статистическа информация, което означава постепенно заличаване на *белите информационни полета*, към които има потребителски интерес.

Тази политика е абсолютно необходима поради четири основни причини, както следва:

- **Хроничен недостиг на ресурси,** което се отразява негативно върху представителността, точността и достоверността на регулярните статистически изследвания.
- Необходимост от **информация в съкратени срокове от време,** която трябва да бъде в синхрон с динамичните промени във всички сфери на икономиката и обществото в национален и международен аспект.
- **Синхронизиране на структурните реформи в статистическата дейност** с иновациите в информационните технологии. Отсъствието на синхронизация може да доведе до колапс на системата, тъй като съставна част на всяка реформа е промяна в структурата, т.е. замяна на старите елементи с качествено нови.
- През последните две - три години **почти във всички европейски страни започна изграждането на специални методологически центрове за изучаване и приложение на Big Data** като източник за информация за целите на официалната статистика. Това е достатъчна предпоставка да се търси добрата практика и да се приеме и използва опитът на по-напредналите страни в това отношение. По този начин ще се изградят реални условия НСИ *да не бъде в ролята на догонваща институция*, а да бъде **равноправен партньор** в процеса за развитие на информационните технологии в обозримо бъдеще.

[11] Има се предвид въвеждането на тристепенна структура на управление на НСИ през 2015 г., както следва: Централно управление; шест териториални статистически бюра; двадесет и осем статистически отдела.

ИЗПОЛЗВАНА ЛИТЕРАТУРА:

Arrington, M. Google Trends Launches (2006), http://techcrunch.com/2006/05/10/google-trends-launches/

Big Data Strategy Document, 2016, EUROPEAN COMMISSION, EUROSTAT

Big Data for Development: Challenges & Opportunities, May 2012, Global Pluse.

Big Data and big business: Should statistcians join in? Contrroversyq August 2013.

Big Data: A Perspective from the BLS, 1 January 2013 2,982 views One Comment.

Brilliant, L. Detecting influenza epidemics using search engine query data, Nature Vol. 457 N. 7232 (2009), 1012--1014, http://www.nature.com/nature/journal/v457/n7232/pdf/nature07634.pdf

Butler, D. When Google got flu wrong., Nature Vol. 494 N. 7436 (2013), 155, http://www.nature.com/news/when-google-got-flu-wrong-1.12413

Choi, H. and H.R. Varian, Predicting the present with google trends, Economic Record Vol. 88 N. s1 (2012), 2-9, http://onlinelibrary.wiley.com/doi/10.1111/j.1475-4932.2012.00809.x/pdf

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

Cook, S. and C. Conrad and A.L. Fowlkes and M.H. Mohebbi, Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic, PloS one Vol. 6 N. 8 (2011), e23610 http://www.plosone.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0023610&representation=PDF

Daas, P.J.H. and M.J.H. Puts, Social media sentiment and consumer confidence, ECB Statistics Paper Series (2014), http://www.pietdaas.nl/beta/pubs/pubs/Daas_Puts_Sociale_media_cons_conf_Stat_Neth.pdf

Florescu, D. and M. Karlberg and F. Reis and P.R. Del Castillo and M. Skaliotis and A. Wirthmann, Will 'big data'transform official statistics? (2014) http://www.q2014.at/fileadmin/user_upload/ESTAT-Q2014-BigDataOS-v1a.pdf

Harford, T. Big Data: are We Making a Big Mistake, Financial Times Magazine (2014), http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz2xlNF6IjV

Lazer, D. and R. Kennedy and G. King and A. Vespignani, The Parable of Google Flu: Traps in Big Data Analysis, Science Vol. 343 N. 41712 (2014), http://dash.harvard.edu/bitstream/handle/1/12016836/The%20Parable%20of%20Google%20Flu%20%28WP-Final%29.pdf

Reis, Ferreira, Perduca (2014). The use of web activity evidence to increase the timeliness of official statistics indicators_IAOS_conference_paper

Ricciato, F., Pete Widhalm, Massimo Craglia and Francesco Pantisano, Extraction of population density distribution from network-based mobile phone data, July 22,2015. New Techniques and Technologies for Statistics 2015, Reliable Evidence for a Society in Transition, Brussels 9 – 13 March 2015.

Toth, I.J. and M. Hajdu, Google as a tool for nowcasting household consumption: estimations on Hungarian data Vol. 7 (2013), http://m.gvi.hu/data/research/ciret_2012_tij_hm_paper_120415.pdf