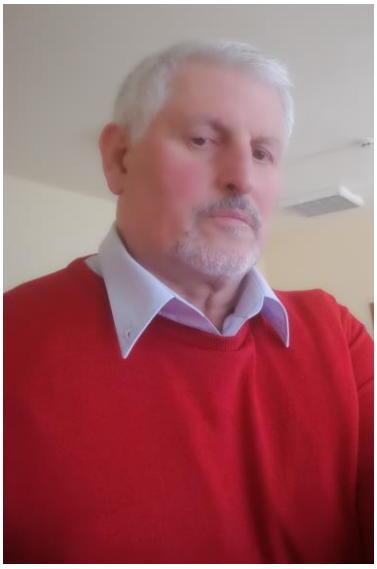
КОХЕРЕНТНОСТ И РАЗЛИЧИЯ МЕЖДУ ГОЛЕМИТЕ ДАННИ (BIG DATA) И ПРЕДСТАВИТЕЛНИТЕ СТАТИСТИЧЕСКИ ИЗСЛЕДВАНИЯ

Богдан Богданов, Галя Статева***

„В промяната намираме предназначението си.“
Хераклит




Въведение

Big Data са конкуренти, необходимост, допълнение, временно явление или заместители на официалната статистика?! Въпроси, на които 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 очевиден факт е, че процесите на глобализация и технологизация във всички сфери на обществото оказват огромен натиск по отношение на управлението в национален и международен аспект. За успешното управление е необходима добра информация, което означава, че тя трябва да притежава следните характеристики: синтезирана,

* Д-р, заместник-председател на Националния статистически институт; e-mail: BBogdanov@nsi.bg.

** Д-р, държавен експерт в отдел „Обща методология и анализ на статистическите изследвания“, дирекция „Методологично-учебен център“, Национален статистически институт; e-mail: GStateva@nsi.bg.

кратка, точна, акцентирана, подсказваща най-доброто решение на проблемите и очертаваща (прогнозираща) хоризонтите за развитие на процесите и явленията.

В този смисъл дебатът може да започне и от друга гледна точка: анализът на събитията да бъде отвъд източниците, измерването и напрежението, необходими при създаването на информация, а също така отвъд политиката. Основният аргумент за това обстоятелство е, че Big Data е друга, различна от това, което познаваме, екосистема. Успоредно с това през двадесет и първи век тази екосистема е исторически феномен на човешкото развитие. Към момента е ясно, че Big Data могат да се доставят по-бързо, на ниска цена и в голям обем, но те все още не са в състояние да заменят напълно официалната статистика, а по-скоро могат да бъдат допълнение към нея . В контекста на тази ситуация интересът към Big Data нараства и се търсят пътища за покриване на съществуващите бели полета в официалната статистика, тъй като много фактори и причини предизвикват сериозни пукнатини между теорията и практиката при осъществяване на статистическите изследвания. Как се очертава бъдещото съжителство между официалната статистика и Big Data? Единственият начин да се получи отговор със задоволителна степен на достоверност е, когато се анализират възможните елементи и допирни точки на съжителството, което е неизбежно в условията на глобализиращия се свят. В крайна сметка целта е да се намерят най-добрите източници на данни, но това съвсем не означава, че могат да се очакват чудеса и че тези данни по магически начин и механично ще доведат до вземането на най-добрите икономически и политически решения, премахващи бедността, безработицата, замърсяването на околната среда и т.н. Търсенето и намирането на реалистичност и логика в информационните потоци е особено важно, но за това са необходими инструменти и подготвени експерти. Същевременно очарованието на Big Data може да се разглежда и като политически ангажимент, тъй като съществуват ясно изразени връзки с възможността за изграждането на електронно управление, което ще бъде значително по-прозрачно и достъпно за обществото.

1. Основните принципи на официалната статистика и проблеми при реализацията на представителните изследвания

Официалната статистика е **необходим елемент** на информационната система, обслужваща правителството, икономиката и обществото. Нейната значимост се основава на фундаментални научни теории, доказани в практиката. Запазването на **доверието** на държавните органи и цялото общество към дейността на официалната статистика е основополагащ принцип.

В публикацията на ООН „Основни принципи на официалната статистика“[1] ясно е посочено, че официалната статистика следва следните водещи принципи:

- Улесняване на **коректната интерпретация** на статистическите данни чрез представяне на изчерпателна и научнообоснована информация за метаданните.
- Коментира и **поправя погрешната интерпретация** и неправилната употреба на статистическите данни.
- Избира различни източници на данни независимо дали са от статистически изследвания, или от административни записи, **гарантирайки тяхната достоверност.**
- Индивидуалните данни са **конфиденциални** и се използват единствено и само за статистически цели.
- Законите, регулациите и измерванията, чрез които статистическата система действа, са **публични.**
- **Координацията** между институциите и статистическата система е съществен момент в процеса на създаването на статистическа информация.
- Използването на **международни концепции, стандарти, методи и класификации** гарантира ефективността и съдържателността на статистическата система на всички нива.
- **Двустранната и многостранната координация** допринасят за подобряване на системата на официалната статистика във всяка страна.

Принципите и задачите на официалната статистика трябва да се допълнят, обновят и актуализират в съответствие с промените в общественото развитие в национален и световен аспект. Това означава, че с разширяване на полето на познание следва да се генерират идеи за ревизия и надграждане, преразпределение на отговорностите, институционалната колегиалност, стратегическите направления и отчитане на спецификите на новите явления и процеси. Успоредно с това на дневен ред е една особено етична категория - ***доверието на обществото към данните на официалната статистика.*** Това е **категория, която е трудно измерима и доловима,** но има огромно влияние върху работата на статистиците. Принципно погледнато, съмненията и подозренията по отношение на достоверността на статистическите данни започват, когато са налице **сериозни *пукнатини* между теоретичните постановки и практическото изпълнение.** Правилото е, че колкото по-големи са пукнатините, толкова по-очевидно е лошото качество на получените

[1] Source: United Nations (2014). Fundamental Principles of Official Statistics, UN General Assembly Resolution 68/261, United Nations.

резултати от дадено статистическо изследване. Това означава, че изследването може да е подготвено теоретично много добре, но неговата реализация на терена да бъде тотално провалена по редица причини (лошо обучение или недостатъчен брой анкетьори, недостиг на транспорт за достигане на труднодостъпни респонденти, отказ на респондентите от участие в статистическите изследвания и т.н.). Възможно е и друго: неясно написани инструкции за изследването; грешни постановки на методологията; неудачен дизайн на извадката и в крайна сметка лошо формирана извадка; грешки при разработване на софтуера за обработка на първичната информация и т.н. Не на последно място от голямо значение е и финансирането на дадено статистическо изследване, за да се постигне желаната и необходима стохастична точност на оценките и качество на информацията. Това означава, че броят на изследваните случаи (обем на дадена извадка) трябва да съответства на предварително поставените цели. В своята книга „Мисленето“ Канеман отбелязва, че „Малкият брой случаи определят крайните резултати, както високи, така и ниски“. Оттук следва, че „Изследователите, които избират прекалено малки извадки, се излагат на произвола на късмета“ (Канеман, с. 146). Това може да се приеме и като същността на „закона за малките числа“.

Основните и важни аспекти на проблемите от практическата реализация на дадено статистическо изследване могат да се представят, както следва:

Първо. Както беше посочено, при планирането на дадено статистическо изследване въпросът за неговото финансиране е стратегически важен. Накратко: от средствата за неговото осъществяване зависи точността на оценките за основните показатели от изследването. Известно е, че при осъществяване на дадено статистическо изследване с малък брой респонденти оценките ще бъдат с ниска степен на точност (високи стойности на стохастични грешки), което ще се отрази негативно върху качеството на последващата аналитична работа. Изводите от такива изследвания са рискови и с ниска степен на надеждност, тъй като интервалът на доверителност за всяка отделна оценка на интересуващ ни показател ще бъде в доста широки граници „от - до“, т.е. с ниска стохастична точност на оценките.

Второ. В хода на изследванията по една или друга причина отпадат респонденти. На практика се случва така, че предварително планираната извадка не запазва своята цялост при работа на терена и се стига до т.нар. деформиране на извадката. При това обстоятелство се поставят редица въпроси, които предопределят надеждността на крайните резултати. Те могат да се формулират и така: отпадналите респонденти в количествен и качествен аспект, от които възможността за получаване на информация е безвъзвратно загубена, могат систематично да повлияят върху стойностите на

крайните оценки в една или друга посока - прекалено подценяване или надценяване на феномена, който се изследва.

Трето. Работата на анкетьорските екипи по места е от изключителна важност за получаването на достоверна информация от респондентите. Недообхватът на информация води до повишаване на нестохастичната грешка на оценката за показателите от изследването. Причините могат да бъдат различни, но особено откроваща се сред тях е прекаленото натоварване на респондентите, което води до отказ или неотговаряне на определени въпроси. Това е проблем, който нараства през годините с увеличаване на въпросите в статистическите въпросници и формуляри. При такива обстоятелства качеството на статистическата информация се понижава.

Четвърто. Отговорността за реализирането на дадено статистическо изследване се разпределя от центъра към регионалните структури на държавната статистика. Статистическите данни и аналитичните справки са послания към обществото на национално и регионално равнище. Това означава необходимост от постигането на прагматична стойност на резултатите, което ще се отрази при вземането на решения от регионалното управление на страната в съчетание с държавните интереси. Малките извадки по регионите на страната обаче не могат да осигурят такава точност на оценките както на национално ниво. Това означава, че местните органи за управление не разполагат с необходимата информация, за да осъществяват целенасочено и действено влияние върху процесите и явленията, развиващи се в рамките на региона.

Пето. Експертите са категорични, че основната и важна роля на официалната статистика е да произвежда и предоставя знания за обществото. На тази основа се оценяват и ефектите от икономическите и политическите решения. Втората важна функция е свързана с осигуряване на достатъчно свободно пространство за публични дебати между институциите за естеството на статистическите измервания. Невинаги обаче информацията на официалната статистика е достатъчна като обхват, точност и разнообразие за процесите и явленията в общественото пространство. Това обстоятелство се дължи на факта, че обществената индустрия непрекъснато произвежда допълнителна информация за фундаментални трансформации в обществото. По този начин се създава и ново знание. Протичат два процеса, еднакво необходими за развитието на обществото като цяло. Въпросът е: как те да се срещнат и свържат? Необходимостта това да се случи, се дължи на обстоятелството, че извън официалната статистика се създава ново знание, което влияе на общественото развитие, което включва, естествено, и правещите политика, т.нар. полисмейкъри. По този начин обхватът на необходимото знание за развитие на обществото нараства

непрекъснато и официалната статистика не може да компенсира това, ако не надгражда теоретически и практически дейността си по производството на информация, съчетавайки я с информационните потоци на Big Data. Информацията, създавана от официалната статистика, заема все по-малко място в света на информацията изобщо. Това обстоятелство тотално променя потребителските нагласи и очаквания за получаването на навременна и надеждна статистическа информация.

Следва да се отбележи, че посочените проблеми на официалната статистика произтичат от една строга и респектираща рамка, където условията не могат да се пренебрегнат, а трябва да се спазват. Тази рамка в голяма степен ограничава и в много малка степен допуска експериментиране и комбиниране на съществуващите информационни потоци. Със сигурност тези ограничения напълно отпадат, когато изследователите анализират Big Data. Така например отпадат проблемите, свързани с: отказите на респондентите за участие в изследванията; умората и натоварването на респондентите; необходимостта от импутиране на данни; проблемите с малките извадки; нелеките задачи за организационна работа с анкетьорски екипи; недостига в обхвата на информацията за наблюдаваното явление и процес и т.н. Това съвсем не означава, че проблемите отпадат изцяло. На дневен ред си остава проблемът с качеството и достоверността на данните от различни източници на Big Data като например социалните мрежи. Наред с това нарастват изискванията към самите изследователи като знания, рутина, опит и широка обща култура. Възможностите и предизвикателствата на работата с Big Data са много големи. Мащабите за съчетаване, комбиниране и селектиране на информация на практика нямат ограничения. Тази ситуация предопределя необходимостта от овладяване на информационната сила, съхраняваща се в Big Data.

2. Big Data в контекста на класическата теория на извадковите изследвания

Всяко число, резултат от статистическо изследване, е материализация на поредица понятия от теорията на вероятностите, статистиката и математиката. Основните понятия, с които си служи официалната статистика, са: „закон за големите числа“; „генерална съвкупност“; „представителна извадка“. За генезиса и същността на тези ключови понятия са направени редица фундаментални изследвания. Добре ще бъде да се припомнят някои аспекти от тях в контекста на Big Data. По този начин ще се открояят различията и ще се търсят пътищата за съчетание. В този аспект може да се започне с определението на закона за големите числа, който гласи: „Свойствата на много закономерности от обективния свят да се формират отчетливо само в масовите процеси, само при

достатъчно голям брой елементи на съвкупности, се нарича закон за големите числа“ (Пасхавер, 1974, с. 18).

На практика това означава, че всеки отделен елемент съдържа в себе си част от обективните закономерности, която се проявява само при изучаването на съвкупност от елементи. Обратно, „Закономерности, проявяващи се в единичното, във всеки отделен елемент, се наричат динамични закономерности“ (Пасхавер, 1974, с. 4).

Първият математически израз на закона за големите числа е теоремата на **Бернули (1654 - 1705).** Теоремата на Бернули е публикувана през 1713 г. в труд, озаглавен „Изкуството да се правят догадки“, и се дефинира така: „Ако вероятността за настъпване на някакво събитие „А“ в последователни и независими опити е неизменна, постоянна и равна на „р“, а относителният дял (релативната честота) за неговата поява е m/n, то вероятността „Р“ на абсолютната разлика |m/n - p| да бъде по-малка от произволно избрано положително число ε, ще клони към 1 при увеличаване на броя на изпитанията“ (Пасхавер, 1974, с. 34).

От практическа гледна точка това означава, че разликата |m/n - p| показва доколко честотата (делът на случаите за настъпване на събитието в една извадкова съвкупност) се отличава от вероятността (относителния дял на случаите за настъпване на събитието в генералната съвкупност) или в каква степен **възможността се отличава (различава) от действителността.**

По същество идеята на Бернули се отнася до значително прост модел, когато събитието се появява или не и когато вероятността за всяко събитие е постоянна. Развитието на идеите на Бернули са продължени от **Поасон (1781 - 1840),** който въвежда понятието ***закон за големите числа.*** Той доказва, че теоремата на Бернули е вярна и в случаите, когато вероятността „р“ се мени в хода на изпитанията независимо от резултатите на предходните изпитания. Доказателството се основава върху поредица от опити, при които се използват няколко урни с различен състав на бели и черни топки. По този начин се доказва, че ако вероятностите за настъпване на събитието „А“ (например появяването на топка с определен цвят) са $p_1, p_2, p_3, \ldots p_n$, а относителната честота на събитието „А“ е m/n в направените изпитания, то вероятността „Р“ на абсолютната разлика

$$\left| \frac{m}{n} - \frac{p_1 + p_2 + p_3 + \ldots + p_n}{k} \right|$$

да бъде по-малка от произволно избрано положително число ε, ще клони към 1 при увеличаване на броя на опитите:

$$\lim_{n\to\infty} P\left(\left| m/n - \overline{p} \right| < \varepsilon\right) = 1$$, където

$$\overline{p} = \frac{p_1 + p_2 + p_3 + ... + p_n}{k}.$$

Поасон публикува своите идеи през 1837 г. в труда си „Изследвания за вероятностните съждения“, като означава:

$$\overline{p} = \frac{p_1 + p_2 + p_3 + ... + p_n}{k}$$, където „к“ е пример с броя на урните.

Теоремите на Бернули и Поасон са свързани с алтернативната изменчивост на разглежданите събития, но в действителност признаците, по които се изучава дадено събитие, имат повече от две значения. Това обстоятелство математически е изразено с теоремата на **Чебишев (1821 - 1894),** от която произтича, че ако средната на генералната съвкупност от изучавани случаи по определен признак е μ, а средната на извадковата съвкупност - $\overline{x}$, то вероятността на абсолютната разлика $|\overline{x} - \mu|$ да бъде по-малка от произволно избрано число ε, ще клони към 1 при увеличаване на обема на извадката n (Пасхавер, 1974, с. 40 - 42).

$$\lim_{n\to\infty} P\left(\left| \overline{x} - \mu \right| < \varepsilon\right) = 1.$$

С продължаване на математическия анализ на закона за големите числа се доказва, че не само средната величина, но и нейното разпределение, както и разпределението на отклоненията от средната, се формират в масовите процеси под влияние на причини в самото явление (Пасхавер, пак там).

Закономерността за приближаването на разпределението на извадковите средни към нормалното разпределение с увеличаване на обема на извадките е обобщена в теоремата на **Ляпунов (1857 - 1918)** и се нарича **централна пределна теорема.** При нейното дефиниране математическият израз определя общите и необходими условия, в съответствие с които разпределението на извадковите средни се приближава към нормалното при достатъчно голям обем на извадките (даден предел), независимо от това как се разпределят вероятностите на индивидуалните величини (единиците по значенията на признака), от които са формирани извадковите средни.

Като се отчита обстоятелството, че извадковата средна величина е случайна променлива величина, теоремата на Ляпунов може да се формулира така: случайна променлива величина,

състояща се от голям брой взаимно независими слагаеми, сред които няма нито едно рязко отличаващо се със своите колебания, има нормално разпределение.

Практически Ляпунов доказва, че при достатъчно голям обем на извадката и ограничен размер на дисперсията от генералната съвкупност вероятността разликата между извадковата средна и средната на генералната съвкупност да бъде в пределите на максималната грешка, е равна на плътността на вероятностите при нормалното разпределение:

$$P(|\overline{X} - \mu| < z\frac{\sigma}{\sqrt{n}}) = \frac{1}{\sqrt{2\pi}}\int_{-z}^{z} e^{-\frac{z^2}{2}} dz .$$

По този начин от теоремите на Бернули, Поасон и Чебишев, разглеждащи и характеризиращи закона за големите числа по отношение на честотите и средните, като естествено продължение теоремата на Ляпунов представя същия закон по отношение на закономерностите на разпределение на случайната променлива.

Следват разработките на Марков (1856 - 1922) и Бернщайн (1880 - 1968), които доказват, че законът за големите числа се проявява при определени условия и при взаимозависими единични събития. В частност Марков разпространява закона за нормално разпределение за зависими величини и при условие, че връзката между тях намалява в размера на тяхното отделяне една от друга и че вариацията на тези величини е ограничена в определени предели.

Big Data, разгледани в контекста на тези постановки, като начало могат да се представят с кратко описание на тяхната същност. Погледнато в този аспект, следва да се отбележи, първо, че отсъства съгласие по отношение на пълнотата на дефиницията за Big Data, но това се приема за по-малък проблем отколкото цялостния фокус върху тези данни, които понякога се представят като „голяма база“ от данни, или трите V-та (обем, скорост и разнообразие) (volume, velocity, and variety). Нещо повече, дефиницията се обогатява непрекъснато в резултат на разширяване и динамизиране на работата с Big Data. Вече се говори за преход от трите V-та към трите C-та (общност, капацитет и среда) (community, capacities, crumbs) (Emmanuel Letouzé Johannes Jütting, p. 11).

Средата идентифицира Big Data като нов вид инертно генерирани индивидуални и комбинирани поредици от следи, резултат от човешката дейност, осъществявана от дигитални средства. Тези средства непрекъснато се усъвършенстват и притежават способността да *нарисуват невероятно реалистични картини* на социално-икономическия живот, отличаващи се с много високо ниво на качество.

Капацитетът определя Big Data не само като набор (обем) от данни, но и като цел, смисъл, стремеж и качество. Този фрагмент от дефиницията за големите данни е имплицитно включен в цялостния производствен процес на информационния продукт.

Big Data съдържат в себе си също **общност,** което се определя от обединението на групи хора с еднакви интереси. Те съставляват елементите, частиците, фрагментите, които формират цялото информационно пространство. Това пространство може да се разглежда като жив организъм, който непрекъснато увеличава своите мащаби, обхващайки всички сфери на жизнения цикъл.

На тази основа Big Data определят екосистемата, съдържаща в себе си повече от данни, средства, методи и действащи лица. Познанието за екосистемата е необходимо, но нейното влияние върху официалната статистика може да се разбере напълно, ако официалната статистика също се познава добре като теория и практика. Експертите подчертават, че когато се говори за официална статистика, думата „официална“ не е синоним на думата „реална“. Тази кратка констатация може да се приеме като начало на генезиса за познавателния смисъл на взаимозависимостта между Big Data и официалната статистика. Започва се с това, че данните на официалната статистика се оприличават на сенки в пещера, отразяващи реалния живот, което всъщност може да се приеме и като грубо приближаване на действителността. Като пример в това отношение може да се посочи представителното извадково наблюдение на домакинствата, осъществявано от всички държави - членки на ЕС - „Статистика на доходите и условията на живот (SILC)“, където съвкупността на бедните под относителната линия на бедност се формира, като се приема, че те получават до 60% от медианния еквивалентен доход. Това обаче не означава, че техният бит задължително отразява реалния живот на бедните. Използването на абсолютния и субективния метод за изчисляване на оценките на бедността показва други равнища и линии на бедност. Оттук следва и изводът, че официалната статистика оценява по-скоро съвкупност от населението със сравнително ниски доходи, което не означава и не е задължително тяхното класифициране като бедни. Освен това изследователите са наясно, че съобразно икономическото развитие на дадена страна е необходимо да се използват специфични подходи при изследване на феномена „бедност“ и разработване на „линия на бедност“, съответстващи на провежданите социални политики като цяло[2]. С други думи, става въпрос за подоходно разпределение на населението (независимо от степента на икономическо развитие на отделната страна), а не за действителни оценки за бедността при равни други условия.

[2] Вж. Geranda Notten and Chris de Neubourg (2011). Monitoring absolute and relative poverty „not enough“ is not the same as „much less“. The review of income and wealth, series 57, number 2, June 2011.

Под равни други условия се има предвид отсъствието на оценки за други материални активи, получавани от членовете на домакинствата доходи от дейности в областта на неформалната икономика (например работа в домашното стопанство, частно хотелиерство и т.н.), както и обстоятелството, че начинът на живот на отделното домакинство може да бъде напълно осъзнат избор. Могат да се дадат и редица други примери за дейности и области, етикетирани като „скрита икономика“ и „неформална икономика“.

Тези случаи, разгледани в светлината на Big Data, могат да придобият видимо по-реални измерения, тъй като данните (информацията) в социалните мрежи се създават за лица и/или обекти, които не се интересуват от рамките на официалната статистика. Нещо повече, тези данни са фактически и транслирането им в системата на информационните технологии се осъществява автоматично, т.е. без човешка намеса. Това обстоятелство предотвратява допускането на случайни или умишлено направени грешки в информационните масиви. Така например получаването на данни за оборота и цените от търговските вериги чрез средствата на информационните технологии, а не чрез посредническата роля на анкетьори води до намаляване на време, средства, ресурси и не на последно място свежда риска от грешки до минимум. Това означава, че тези примери, разгледани в контекста на Big Data, превръщат понятията „закон за големите числа“, „генерална съвкупност“ и „представителна извадка“ в евфемизъм[3].

3. Възможности за оценка на качеството на големите данни

Независимо от източника на данни следва да се формулират пет основни принципа за качество, на които те се подчиняват: полезност; достоверност; наличност; достъпност; устойчивост. Когато се анализират числа, получени от статистически изследвания, трябва да се знае, че зад тях се намират редица постулати на статистическата теория, които започват със закона за големите числа и централната пределна теорема. С други думи, тези познания трябва да бъдат на *десктопа* в паметта на анализаторите, за да бъде успешна тяхната мисия при реализацията на дадено изследване.

При анализа на числата от Big Data ситуацията се усложнява, тъй като познанията на анализаторите трябва да надхвърлят натрупаните знания и опит от теорията и практиката на статистическите изследвания, официално осъществявани от държавната статистика. Това означава

[3] Понятието „евфемизъм“ се тълкува по различни начини в съответствие с различните области на познанието. Съотнесено към постановките в изложението, означава, че разглежданите понятия в контекста на Big Data са загубили буквалното си значение.

още, че анализаторите трябва да притежават изострена сензитивност по отношение на естеството на числата. Тази чувствителност може да се изразява, образно казано, в интуиция и усещане за *екологичната чистота* на информационните потоци (по отношение на съдържащите се елементи и фрагменти на дезинформация).

Принципите и методите при анализ на качеството с използване на конвенционални методи могат да бъдат, както следва:

- Верифициране чрез **информация от други изследвания,** които са провеждани от официалната статистика в минали или по-близки периоди от време. Тази информация може да бъде косвена или пряка, но винаги в определена логическа връзка с информацията, получена чрез Big Data. Това е подход, който се прилага при верифициране на данните от статистическите изследвания в официалната статистика. За тези цели могат да се използват също и т.нар. core variables, т.е. ключови променливи, които се срещат в масивите от данни от различни източници.

- Създаване и **анализиране на динамични редове** от данни, получени чрез информацията от Big Data. Търси се тенденция, която следва логиката на изследвания процес и/или явление. Определят се връзките, взаимозависимостите и кохерентността с други явления и процеси, които имат подобни белези, природа и вид. Използват се закономерностите за наличието на причинно-следствената зависимост между явленията и процесите в природата и обществото, предопределящи неговото развитие. По принцип един динамичен ред на показател, разработван по данни от официалната статистика, може да служи като опорна точка за неговото продължение чрез използване на Big Data.

- Търсене на **сходство с разпределенията,** известни и използвани от официалната статистика. Например: нормално, логнормално, експоненциално, многомодално или други разпределения, познати и прилагани в статистическата практика.

- Сравняване с ключови данни от **административни източници** на данни. Мнението на експертите е, че административните данни могат да бъдат разгледани като отличаващи се от Big Data, тъй като са организирани като страничен продукт в голям мащаб от административни системи и обикновено генерират цели, които се различават от официалната статистика (Big Data: Potential, Challenges, and Statistical Implications, с. 19). Редица северноевропейски страни (Дания, Норвегия, Финландия, Швеция) използват различни регистри за производство на статистически данни по време на преброяванията (пак там, с. 19). Отбелязва се важната роля на държавната статистика за създаването на дизайн за генериране и получаване на административни данни.

- Използване на **допълнителни източници на данни от различно естество** за допълнително верифициране на естеството на изследваните явления и процеси.

- **Крайният статистически продукт на практика са характеристиките на статистически разпределения[4].** Те могат да имат формата на абсолютни и относителни величини, коефициенти, индекси, съотношения и т.н. Значимостта и дълбочината на техния анализ зависят от опита и знанията на изследователския екип. Основно правило е изводите и заключенията да отразяват не само изучаваното явление, но и неговата връзка (корелация) с други явления и процеси, които имат най-съществено влияние за неговия генезис, естество и развитие. В този смисъл трябва да се търси изследователската рамка за обхват. Противоречивите тенденции за изследвани явления и процеси са знак, че трябва да се направи верификация на изследваните съвкупности от първични данни.

- **Стохастичната точност на оценките** може да се изчисли условно. За целта могат да се използват подходите за определяне на точност, като се приеме, че генералната съвкупност съдържа много голям брой единици, което прави възможно отпадането на поправката за крайна генерална съвкупност. Очевидно е, че представителността и точността на оценките ще бъдат единствено и само за изследвания период от време. На преден план и преди всичко е важно да се гарантира достоверността на данните. Ако се използва аналогия със статистическата терминология, става дума за т.нар. нестохастични грешки. Техният произход е свързан с наличието на умишлено или случайно представени данни в източниците на Big Data.

Целта е чрез използване на стандартите на официалната статистика една неструктурирана съвкупност от данни да се превърне в структурирана. Тази концепция се разглежда от експертите на държавните статистики в много страни. В техните аналитични материали се дава поредица от примери и предложения за използване на големите данни за целите на официалната статистика (Monica Scannapieco, Antonino Virgillito, Diego Zardetto, Placing Big Data in Official Statistics: A Big Challenge?, p. 10). Всички те са наясно, че има важни за обществото и държавата явления и процеси, в които може да се проникне чрез Big Data и където официалната статистика е безсилна.

[4] Вж. „Big Data conversion techniques including their main features and characteristics“, 2017 edition. Statistical working papers, Eurostat (p. 16 - 18).

4. Бъдещето на извадковите статистически изследвания чрез използване на Big Data

В условията на глобализация традиционните статистически изследвания са поставени под заплаха, тъй като **намалява възможността да представят реални оценки за редица явления и процеси в обществото.** Причините за това обстоятелство могат да се търсят основно в две направления. Първото е свързано с динамиката на икономическите и социалните промени, които дават своят отпечатък върху ценностната система, традиции, обичаи, начин на живот и мироглед. Негативните аспекти на промените пряко влияят върху достоверността на резултатите от изследванията. Например при изследване на бюджетите на домакинствата това са: отчуждението; нихилизмът; страхът от чуждо посегателство; недоверието към институциите и между хората; нежеланието за разкриване на информация, която се счита, че засяга личния живот. Друг пример е за предприятията в страната: желанието да се укриват данъци върху печалбата; стремежът да се прикрият дейности, които не отговарят напълно на приетите стандарти и закони в страната; опити да се наложи нелоялна конкуренция и т.н.

Второто направление е свързано с ускорените темпове на трудовото ежедневие, при което времето за анализ на получените числа намалява. Това означава още, че настъпват тотални промени в мащабите на икономическите и социалните нагласи, интереси и цели в трудоспособните слоеве от населението. В своята съвкупност те определят цялостното поведение на потенциалния респондент, при което **има опасност лоялността, доверието и отговорността към държавните институции да ерозират** във времето и пространството, ако не се намерят начини за противодействие.

Съществува една интересна мисъл на Гьоте: **„Няма нищо по-рядко от разума, тръгнал по нов път“.** В контекста на тази мисъл може да се каже, че използването на големите данни за статистически цели е **предизвикателство** не само **за Европейската статистическа система,** но и за всички изследователи, които искат да извлекат аналитично познание за социално-икономическите явления чрез новите източници на данни. Налице е вече един обективен феномен и следва да се разработят методологични процедури и подходящи ИТ средства за реализация, които са предмет на редица международни и европейски проекти към момента. Счита се, че съчетаването на големите данни с данните от официалната статистика ще доведе до решаващи промени, разнообразие, обогатяване и детайлизиране на информационните потоци в редица области на икономиката и социалния живот. Много от проектите са с хоризонт до 2021 година.

Какво се очаква? Очаква се чрез реализацията на част от тези проекти да се постигне ефективно и практическо внедряване на Big Data в статистическия бизнес процес, по-голяма степен

на структуриране и превръщане на работата на държавните статистически институти в ЕСС в **цялостен процес на производство** и разпространение на информационния продукт като резултат от използването на тези нови източници на данни. Чрез активното участие на НСИ в подобни проекти се създават реални условия **българската статистика да бъде надежден и равноправен партньор** в Европейската статистическа система.

Какво ще бъде бъдещето за Big Data? Отговорът на този въпрос ще започне с това, на което сме свидетели сега:

- **Глобализация** на икономиката;
- Ускорено развитие на **информационните технологии;**
- Ускоряваща се **динамика на пазара на труда** и пазарите на суровини и материални блага;
- **Нарастваща конкуренция** във всички сфери на обществото за по-добро качество и по-разнообразни стоки и услуги;
- Непрекъснато **нарастваща цена на човешкия капитал.** Най-добрите фирми искат най-добрите експерти. Битката между тях в това отношение е безпрецедентна!;
- Непрекъснат **глад за знания и информация** и подчертано доказана достоверна информация с „печат“ за качество.

След 2030 г. официалната статистика няма да бъде такава, каквато я познаваме сега. Вече релефно се очертава тенденцията статистически данни да се произвеждат чрез **Big Data, което включва: разнообразна информация от различни източници: задължително и своевременно прилагане на най-модерните технически техники и средства; коренно различни софтуерни решения и динамично променяща се информационна и комуникационна инфраструктура, която непрекъснато се видоизменя; и най-важното - качествено различен като мислене и по-квалифициран човешки капитал от гледна точка на знания и умения.**

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. Основните професии, свързани с този процес, са на учени (изследователи) и статистици, но не се знае дали в бъдеще ще съществува разлика между тях. Те се увеличават всяко десетилетие (декада) с 15%. Big Data се определят като серия от данни отвъд (прехвърляща) способността на типичните средства (устройства) да събират, извличат, управляват и анализират. Безпристрастната оценка на ползата от големите данни се губи в хиперпространството.

Заключение

Българската статистиката след Втората световна война се развива със сравнително бавни темпове като теория и практика. Известно ускорение се наблюдава през 60-те години с навлизането на големите електронни машини с дискови устройства, в които данните се въвеждаха чрез перфокарти. Този етап продължи приблизително 30 години.

След 1990 г. в продължение на 20 години статистиката направи забележим скок в своето развитие под влияние на редица фактори - промяна в държавното устройство; преминаване към принципите и законите на пазарната икономика; въвеждане на нови информационни технологии за производство на информация; присъединяването ни като член на Европейската статистическа система, наличие на значително по-образован човешки капитал.

През **следващите 10 години е логично да се очаква нов скок в развитието на официалната статистика.** Един от основните фактори за това ще бъде наличието и използването на големите данни за производство на официална статистика.

Наблюдава се нещо, което вече е принципно доказано в социалните теории по отношение на епистемологията (наука за познанието) за развитието на човешкото общество: 40 години от миналото сега се равняват на 20 години, а в бъдеще ще се равняват на 10 години. Налице е процес на непрекъснато ускоряващо се развитие на човечеството (за съжаление, съпътствано и от редица негативни явления).

В момента, метафорично погледнато, **Big Data е все още бялата врана** за официалната статистика. Причината е, че все още не се осъзнава обстоятелството за **необратим процес,** който изисква задълбочено изучаване, промяна на мисленето и създаването на необходимата интелигентност за натрупване на нови знания в теоретичен и практически аспект. Статистиците, информатиците, икономистите имат едно много важно предимство: познаването на статистическите методи и аспекти за обработка и анализ на информацията. Това предимство е и шанс за по-бързо опознаване, разбиране и използване на новата информационна енергия, съдържаща се в големите данни. Считаме, че вече е настъпило времето, когато ще трябва да се направи ревизия на нашите знания по отношение на методите на официалната статистика такива, каквито ги познаваме до сега. Това означава, че: традиционните **наблюдения** постепенно ще се реформират като етап, включващ много подетапи за работа с големи данни; **групировките** ще се основават главно на клъстеризацията на обектите по определени признаци, които няма да бъдат никак малко във времето и пространството, тъй като отразяват многообразието на света, в който живеем (тези обекти ще бъдат в

области, които са недостъпни, труднодостъпни или скъпоструващи за официалната статистика); **използваните методи за обработка** на информацията ще изискват верификация чрез повече от един източник на информация; **анализите,** които се правят сега от експертите, ще изискват нови познания, на значително по-високи равнища от сега съществуващите. Всичко това ще се основава и развива със съвършено нови информационни средства и среда. Това е процес с ускорение и е ясно формулиран в стратегическия документ на Евростат „Визия 2020“: **„As our world is changing, we have to change with it“[5] (ESS Vision 2020).**

[5] „Както светът се променя, така и ние трябва да се променим с него.“

ЦИТИРАНА ЛИТЕРАТУРА:

Даниел Канеман (2012)**.** Мисленето, изд. „Изток-Запад“.

Пасхавер, И. (1974). Закон больших чисел и статистические закономерности. Москва, Статистика, с. 18.

Cornelia L. Hammer, Diane C. Kostroch, Gabriel Quirós, and STA Internal Group. Big Data: Potential, Challenges, and Statistical Implications, September 2017 /SDN/17/06.

Daas, P. J. H., M. J. Puts, B. Buelens and van den Hurk P. A. M. Big Data and Official Statistics.

Letouzé, Emmanuel Johannes Jütting. Official statistics, Big Data and human development, June 2015, In partnership with Paris21.

New Techniques and Technologies for Statistics 2015. Reliable Evidence for a Society in Transition, Brussels 9 - 13 March 2015.

Notten, Geranda and Chris de Neubourg (2011). Monitoring absolute and relative poverty „not enough“ is not the same as „much less“, The review of income and wealth, series 57, number 2, June 2011.

Pohl, Jeans and Kym Pohl, Big Data Opportunities and Ghallenges, InterSymp-2013, 29 July 2013 RESU104IS13.

Scannapieco Monica, Antonino Virgillito, Diego Zardetto (2017). Placing Big Data in Official Statistics: A Big Challenge?

Struijs Peter, Barteld Braaksma and Piet JH Daas. Official statistics and Big Data, Peter Struijs, Barteld Braaksma and Piet JH Daas Big Data & Society 2014 1: DOI: 10.1177/2053951714538417.