ОЦЕНКА НА КАЧЕСТВОТО НА „ГОЛЕМИТЕ ДАННИ“ В ОФИЦИАЛНАТА СТАТИСТИЧЕСКА ПРАКТИКА

Галя Живкова Статева[1]
Стопанска академия „Д. А. Ценов” – Свищов
Катедра „Математика и статистика”

Резюме: Настоящата статия обосновава теоретично подходите за оценка на качеството на т.нар. „големи данни“ (Big Data), свързани с тяхното приложение в официалната статистическа практика като източник на данни. Представени са предложения и възможни решения за повишаване качеството на „големите данни“ при използването им в емпиричните статистически изследвания. Едновременно с това е направен опит за изследване на ефекта от повишаване на тяхното качество в контекста на разширяване обхвата на прилаганата класическа статистическа методология. Посочени са предизвикателствата пред националните статистически служби при въвеждането и използването на „големите данни“ в реалната статистическа среда.

Ключови думи: „големи данни“, качество, статистическа информация, официална статистическа практика.

JEL: C10, C82, C90.

QUALITY ASSESSMENT OF “BIG DATA” IN THE OFFICIAL STATISTICAL PRACTICE

Galya Zhivkova Stateva
D. A. Tsenov Academy of Economics – Svishtov
The Mathematics and Statistics Department

Abstract: The present article theoretically depicts the approaches for assessment of the quality of the so-called “big data” when applied in the official statistical practice as a data source. Some proposals and possible solutions for improvement of the quality of “big data” in their use in empirical statistical surveys are presented. At the same time, we have attempted to research the effect of raising the quality of big data in the context of widening the scope of the applied classical statistical methodology. The paper also outlines the challenges the national statistical offices face in the implementation and the use of “big data” in a real statistical environment.

Keywords: Big Data, quality, statistical information, official statistical practice.

JEL: C10, C82, C90.

[1] *Държавен експерт в отдел „Обща методология и анализ на статистическите изследвания“ на Националния статистически институт.*

Въведение

Използването на „големите данни“ (Big Data) в официалната статистика като възможен източник на информация е свързано с широко дискутираните методологични въпроси за тяхното събиране, обработка и анализ, както и с качеството на получените резултати. Особен интерес предизвиква „селективността“ на Big Data, когато директно се прилагат вече установените статистически методи в ежедневната практика. Прилагането на класическата теория за репрезентативни изучавания за получаване на оценки на базата на Big Data може да се окаже неефективно, особено когато данните не могат да бъдат свързани с предварително дефинирана и известна съвкупност на изучаваните единици. И тогава възниква логичният въпрос, до каква степен и по какъв начин традиционният подход за извършване на статистическите изследвания може да се прилага и за „големите данни“. Според Стратегическата група на ООН за модернизиране на статистическото производство и услуги, „големите данни“ (Big Data)[2] могат да бъдат определени като "източници на данни, обикновено с огромен обем, висока скорост и голямо разнообразие, които изискват разходно-ефективни и иновативни форми на обработка с цел извършване на задълбочен анализ и вземане на адекватни решения за различни социално-икономически явления".

Обект на настоящата статия са „големите данни“ като източник на информация в официалната статистическа практика.

Предмет на изследването са качеството на „големите данни“ и ефектът от тяхното използване в официалната статистика.

Целта на изследването е, теоретично да се обосноват възможните подходи за подобряване на качеството в процеса на производство на официални статистически данни от източници на Big Data, както и опит за обследване на ефекта от тяхното използване по отношение на качеството на статистическата информация и прилаганите методи.

Като правило качеството на данните и методологията са тясно свързани по такъв начин, че постигането на задоволително качество зависи от избраните статистически методи. Значителна част от установения статистически инструментариум е изградена върху теорията на извадковите изследвания, чиято основа се състои в дефиниране на целева съвкупност от единици и променливи, от която се излъчват извадки; събиране и обработване на данни, оценки на получените резултати за изучаваното явление и др., като едновременно с това се съблюдава качеството на данните и се оптимизират разходите за тяхното производство. Тези обстоя-

[2] В статията терминът “Big Data” се използва едновременно на английски език и възможно най-адекватния български еквивалент „големи данни”, тъй като към момента не е намерен по-точен превод на понятието. Английският термин придобива все по-голяма популярност в различни български източници.

телства водят до *три основни групи въпроси***,** които ще бъдат разглеждани в настоящата статия: какви са ограниченията на установената рамка за качество и методология, използвани при традиционните статистически изследвания и административните данни, по отношение на прилагането им върху Big Data; като се вземат предвид тези ограничения, какви са алтернативите за справяне с произтичащите от това предизвикателства; какви са възможните избори, пред които са изправени националните статистически институти в бързо развиващата се Big Data заобикаляща ни среда. Тъй като тези въпроси са твърде „големи“ и са необходими задълбочени отговори, в същинското изложение на статията са описани основни насоки за по-нататъшен размисъл по въпросите за качеството на Big Data.

За изпълнението на поставената цел се поставят следните **изследователски задачи**:

Първо, да се изяснят и теоретично да се обосноват възможните подходи за оценка на качеството на „големите данни“.

Второ, въз основа на натрупания практически опит в областта на Big Data да се обсъдят предложения и възможни решения за повишаване качеството на големите данни при използването им в официалната статистическа практика.

Трето, да се посочат предизвикателствата пред националните статистически институти на европейските държави в контекста на променящата се Big Data информационна среда.

Изследователската теза се основава на убеждението, че подобряването на качеството на информацията в процеса на производство на официални статистически данни от източници на Big Data ще доведе до получаване на бърза и навременна информация, с висока съдържателна стойност от гледна точка на потребителите, в т.ч. държавното управление, стопанската практика, социалната сфера, научните изследвания.

1. Подходи за оценка на качеството на „големите данни“ (Big Data)

A. ***Разширяване обхвата на използваните статистически методи:*** Обикновено, годишните статистическите програми на статистическите институти са базирани на входна информация от статистически изследвания и/или данни от административни източници. Много от вече утвърдените методи са предназначени за класическо статистическо изследване, но в действителност повечето изследвания използват за рамка на съвкупността данни, получени от административни източници. В днешно време дори за целите на преброяването на лицата и домакинствата се използват административни данни. Административните данни сами по себе си също са основен източник на данни за производство на статистически изходи. В последните няколко години се забелязва тенденция, при която

статистическите изследвания все повече се използват за допълване и обогатяване на информацията от административни източници, а не обратното. Това е логично следствие от широко преследваните цели за намаляване тежестта на респондентите за постигане на максимална ефективност. В допълнение към така наречената „коминна“ организация на статистиката, при която, всеки от статистическите процеси се изпълнява независимо един от друг, съществува още и „интегративна“ статистика на базата на множество източници. Типичен пример за такава статистика са националните сметки където се комбинират различни източници на данни от различни статистически области и изучавани променливи. Интеграцията на източниците се извършва на базата на специално разработени модели и методологични техники, тъй като е необходимо да се обединяват разнородни понятия и съвкупности. Интересното в случая е, че статистическите изходи от националните сметки, обикновено не включват оценки на съвкупностите от предприятия. Това може да се дължи на факта, че националните сметки включват доста експертни допускания и моделиране отколкото оценка на съвкупността.

Характеризирането на горепосочените статистически подходи и методи е непълно, тъй като в теорията и практиката съществуват редица методи, насочени към специфични видове статистика, например модели за оценка на резултати, моделиране на динамични редове, регресионни модели и други. Особен интерес в този смисъл представляват методите извън традиционната извадкова теория, които се прилагат за работа с Big Data.

Б. Ограничителни условия – установени подходи за качеството. По отношение на ограничителните условия за качеството в официалната статистика могат да бъдат разграничени две нива. *Първо*, качеството на статистическите данни е пряк резултат от прилаганите методи и тяхната параметризация. Методите и параметризацията се подбират въз основа на техния ефект върху качеството, като се акцентира върху точността. *Второ*, разработени са рамки за качество в Европейската статистическа система, които се прилагат за дефиниране на критерии за качеството и оценка на изпълнението на статистическите програми. Например в базовите документи за управление на качеството *Кодекс на европейската статистическа практика (CoP)* и *Рамка за оценка на качеството (QAF)* съществуват указания за изискуемото ниво на качеството за различните статистически области, включително за националните сметки. В *Кодекса* са дефинирани пет аспекта на качеството на статистическата продукция: относимост, точност и надеждност, актуалност и навременност, съгласуваност и съпоставимост, достъпност и яснота, като първите четири са обект на тази статия.

В. ***Предизвикателства при оценка на качеството на Big Data – примери:*** Към настоящия момент използваемостта на Big Data като надежден източник за производство на официални статистически данни е в тестови етап и е обект на пилотни проекти на европейско и национално

ниво. Все още няма почти никаква официална статистика, произведена и публикувана въз основа на източници на „големи данни“ (с изключение на част от сухопътната транспортна статистика в Холандия), и е твърде рано да се предвидят и изброят всички предизвикателства по отношение качеството на тези данни. За илюстрация могат да се посочат няколко примера, свързани с различни видове източници на Big Data, като резултат от европейската практика, които изискват иновативни решения и формират интересни методологични казуси.

Първият пример се отнася до използването на информация от около 20 000 пътни сензори, отчитащи броя на преминаващите превозни средства по различните класове пътища, и тя е достъпна всяка минута. Този източник има потенциал да се използва в статистиката за седмични индекси на обема на транспортния трафик, включително спецификации за натоварен трафик на регионално ниво. В процеса на събиране и обработка на данните възникват следните проблеми:

- в различните зони разпределението на пътните сензори е неравномерно, има пропуски и при най-краткия времеви период не всички данни са достъпни за всички сензори;
- връзката между съвкупността на превозни средства на национално ниво и данните от сензорите е неизвестна на микониво. Освен това, индивидуалните превозни средства да бъдат проследени във времето;
- метаданните на пътните сензори имат лошо качество.

Вторият пример е свързан с използването на публични съобщения от социалните медии, например от Twitter или Facebook. Този източник има потенциал да се използва за изчисляване на седмични индекси на настроение/нагласи на хората, включително индекс на потребителското доверие. В процеса на събиране и обработка на данните възникват следните проблеми:

- съвкупността на хората, публикуващи съобщения не е известна, нито връзката й с генералната съвкупност на населението;
- възможно е да се разработи система за свързване на видовете настроения към съответните текстови съобщения, но не е очевидно как да се тълкуват измерените по този начин настроения.

Третият пример касае използването на данни от мобилни телефони за местоположението. Този източник има потенциала да бъде използван за ежедневна статистика на населението, мобилността на населението, транспортна статистика и статистика на туризма. В процеса на събиране и обработка на данните възникват следните проблеми:

- наличните данни зависят от гъстотата и устойчивостта на мрежата на доставчика на мобилни телефонни услуги;

- дори данни за собственика на мобилните устройства да са налични, данни от телефоните могат да не се използват, тъй като са изключени или се използват от други лица, различни от собственика.

На базата на идентифицираните проблеми в трите емпирични примера, може да се обобщят някои от основните видове предизвикателства, свързани с качеството и методологията на данните от източници на „големи данни“, а именно:

✓ Възможно е да ***липсва част или цялата информация*** за съвкупността от населението, която генерира Big Data данни (записи, текстови съобщения, видеоизображения и други подобни). Това може да възникне както на микониво (свързване на микроданни е невъзможно), така и на макониво (липса на информация за селективност на данните).

✓ Средствата за измерване могат да показват ***небалансирано или неструктурирано физическо разпределение***, да има **липсващи данни** поради разнообразни причини или да са налице ***проблеми с обхвата на данните*** (свръхобхват или недообхват), включително в динамика.

✓ ***Значението или относимостта на данните*** може да бъде трудно да се оцени, т.е. каква информация всъщност е предадена чрез дадено текстово съобщение, мимолетно настроение, интернет търсачка или просто снимка.

Въпреки че критериите за качество са разработени за целите на традиционните статистически изследвания, при определени допускания, те могат да се адаптират за приложение и при източниците на „големи данни“.

2. Възможни решения за подобряване качеството на данните от Big Data източници

Теоретично погледнато Big Data може да се използва като самостоятелен източник на данни за производство на статистически данни или да се използва като допълнителен източник към традиционните статистически изследвания в комбинация с източници на административни данни. Когато липсва установена методология за работа с „големите данни“, предизвикателствата могат да бъдат решени по алтернативни начини. В научната литература се предлагат различни възможни решения на базата на резултати от тестови пилотни проекти и зависещи от начина и предназначението на използване на Big Data.

В случаите, когато липсва информация за съвкупностите, има начини да се извлече минимална информация за съвкупността, генерираща Big Data. Например за съобщения от социалните медии е възможно да се оценят основни променливи на базата на корелацията между думите на текстовото съобщение и демографски характеристики на автора на съоб-

щението като: пол, възраст, образование и социален статус. На следващ етап придобитата информация за основните променливи позволява прилагането на традиционно установените статистически методи.

Друго възможно решение е свързването на Big Data на мезо- или макроравнище към други масиви от информация, позволяващи прилагането на моделиращи техники. Например, дори ако съвкупността на притежателите на мобилни телефони не е известна, взаимозависимостта на тази съвкупност с административните регистри на населението може да бъде изучавана на агрегирано ниво. Данните за мобилността на населението, измерено чрез мобилните телефони, могат да бъдат верифицирани и свързани със съществуващите статистически данни за трафика на транспортни средства, с данните от статистика на туризма и така нататък.

В други случаи, отнасящи се до проблеми със средствата за измерване или обхвата, е възможно адаптиране на установените методи към някои от конкретните казуси, но има ситуации, в които е необходимо да се прилагат нетрадиционни за статистиката методи като например вероятностно моделиране. Познанията и експертният опит на работещите експерти в националните сметки също могат да доведат до възможни решения на проблемите с обхвата.

В други случаи, свързани със значението и относимостта на „големите данни“, едно от най-разпространеното решение е изучаване на взаимозависимостта между източниците на Big Data и други източници на данни, като между тях трябва да има задължително силна корелационна връзка. Показателите, изчислени на базата на Big Data, могат да бъдат калибрирани или адаптирани към друг набор от данни. Например индексът на настроенията на базата на данни от социалните медии чрез прилагане на специални техники за извличане на данни (data mining) може да бъде адаптиран към вече съществуващия индекс на потребителското доверие, изчислен чрез традиционно статистическо изследване. В допълнение могат да бъдат изградени корелации с други известни явления, ако Big Data позволяват успешно прогнозиране.

Най-радикалният подход би бил да се произвеждат статистически продукти изцяло на данни от източници на Big Data, като анализът и тълкуването на тези резултати са отговорност само на потребителите на информация. На пръв поглед това може да изглежда като недостатъчно разумно предложение за националните статистически институти, но търсенето на нов вид информация, която няма очевидна интерпретация, е достатъчен мотив за това. Например търсенето на Twitter данни за настроенията и нагласите на хората по важни социално-политически въпроси е безпогрешно и ако такава информация се представи като "бета-версия", то може да се предизвика ценна обратна връзка от потребителите. Някои интернет „гиганти“ силно насърчават този вид емпиричен подход.

Един от иновативните подходи е да се извърши отделно изследване с цел събиране на характеристики за неизвестни съвкупности, напри-

мер съвкупността на Facebook потребителите. След като се получат резултати от това проучване, традиционните методи за оценка могат да бъдат лесно приложими. Подобни проучвания успешно се прилагат за измерване качеството на административните източници на данни.

Ако Big Data се използва за бързи и предварителни данни, едно приемливо, макар и нежелателно допустимо решение е намаляване на изискванията и критериите за достигане на определено ниво на качеството на данните.

3. Националните статистически институти: правилните избори в контекста на променящата се среда

При разглеждане на възможностите за използване на източници на Big Data в официалната статистика и свързаните с това потенциални предизвикателства, е необходимо да се обмислят добре рационалните решения, които трябва да предприемат националните статистически служби и да се анализират всички аргументи „за“ и „против“ използването на Big Data в реалния статистически бизнес процес.

Едно от основните предимства на официалните статистически институции пред другите производители на данни е „печатът“ за качество на данните и общественото доверие, което не трябва да се рискува при никакви обстоятелства. В този смисъл предоставяне на информация за съмнителни корелации между явленията, на базата на Big Data, които не са добре обяснени и анализирани, не е задача на НСИ. Предпазливото отношение на официалните статистици е обяснимо, защото те се подчиняват на най-високите професионални стандарти в областта на статистиката. Нещо повече, информацията, получена от Интернет като един от основните източници на „големи данни“, невинаги е убедителна и с добро качество.

Едновременно с това не трябва да забравяме за някои неметодологични въпроси, свързани с неприкосновеността на личните данни, липсата на специализирани умения сред експертите на НСИ или сериозните ИТ предизвикателства по отношение изграждане на подходяща инфраструктура.

Не е ли обаче наивно да се смята, че използването на Big Data за производство на официалната статистика е наистина с негативен ефект? Средата, в която официалната статистика се произвежда, се променя непрекъснато. Традиционно производството на официални статистически данни е монопол на държавата. Но докато в миналото единствено статистическите служби са предоставяли данни за икономически и социални явления на широката общественост, то в днешно време неофициалните статистически данни са широко достъпни и евтини. Разбира се, тяхното качество и обективност може да се оспори, но пък се произвеждат и разпространяват значително по-бързо отколкото официалните статистически данни. По този начин неофициалните статистически данни постепенно се

превръщат в реален конкурент на пазара на данни. Налице е реална опасност, че това застрашава позицията на националните статистически институти и по-специално тяхното финансиране.

Вярно е, че официалната статистика все още изпълнява важна роля в обществото пространство заради своята безпристрастност и високите професионални стандарти, които прилага. Фундаменталната причина, заради която статистическите служби ще продължат да играят водеща роля и в ерата на Big Data, освен производството на информация с добро качество в съответствие с потребителските нужди, е новата им роля като валидатор на данни, произведени от други институции. По този начин доверието в националните статистически служби може да се използва като актив, който непрекъснато да нараства.

Заключение

Националните статистически институти трябва да имат фундаментални знания и да разширяват опита си по отношение на използването на Big Data в ежедневната статистическа практика и извън нея. Прилагането на принципа „количество над качество“, възприет от потребителите на Big Data, не трябва да се пренебрегва. Дори когато източниците на Big Data не се използват за получаване на нови статистически продукти, биха могли да се разглеждат като ефективно средство за намаляване на натовареността на респондентите, при условие че методологичните предизвикателства могат да бъдат разрешени. Използването на Big Data за съставяне на ранни показатели за важни статистики, като например данни за цените или бизнес цикъла, е достатъчно сериозна опция. Прилагането на Big Data за краткосрочни прогнози също не е за пренебрегване.

Традиционният подход за проектиране на статистическите изследвания е да се дефинира желаният резултат, да се изберат подходящи източници на данни и да се оптимизира процесът. Експериментите с Big Data могат да се провеждат и в обратен ред на етапите в този подход: намиране на интересен източник на Big Data, събиране на относимата информация към изучаваното явление, свързване на тази информация с вече налична информация от други източници (дори и само чрез създаване на корелационни модели).

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

В заключение може да се каже, че разглеждайки Big Data с тяхната реална стойност и значимост, и връзките им с други явления, смисълът на тяхното за използване като един от възможните източници в официалната статистика може да бъде не толкова необичаен, колкото изглежда. Когато административни данни се използват от НСИ, подходът е същият, без да се използват като източник за измерване на предварително дефинирани статистически понятия. Например потребителите могат да се интересуват от причините за появата на престъпност, но повечето статистически служби произвеждат статистика на база на административните записи за докладваните престъпления в полицейските регистри. В действителност, когато цената на данните от всеки външен източник по номинална стойност е относително ниска, се препоръчва неговото използване, независимо дали това са структурирани, административни данни или Big Data.

От всичко казано дотук става ясно, че в днешния динамичен свят съществуват основателни причини за преразглеждане на институционалните граници на всяка национална статистическа служба. Една такава причина е наличието на свързаност между корелационната връзка-причина-следствие и прогнозиране. Това е област, в която икономисти и иконометрици имат достатъчно опит и могат да бъдат ефективни. Тогава възниква и логичният въпрос: дали все още е необходимо и трябва да се поддържа разграничението между официалните статистически институти и други икономически институции, които се занимават с прогнозиране на социално-икономическите явления?

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

Използвани източници:

Ангелова, П. (2013). *Статистика.* Свищов: АИ Ценов.

Мишев, Г.; Цветков, Ст. (1998). *Статистика за икономисти.* София: УИ „Стопанство“.

Петков, П. (2010). *Иконометрия.* Свищов: АИ Ценов.

Booleman, M.; al, at. (2014). Statistics and Big Data: Quality with uncontrolled inputs. *Paper prepared for the Q2014 conference.*

Buelens, B.; Daas, P.; Burger, J.; Puts, M.; Van der Brakel, J.;. (2014). Selectivity of Big Data. *Discussion paper, 11*.

Daas, Piet J.H.; al., at. (2014). Big Data as a Source of Statistical Information. *The Survey Statistician, 69*, pp. 22-31.

Daas, Piet J.H.; at al;. (2013). Big Data and Official Statistics. *Presented at the 2013 New Techniques and Technologies for Statistics conference (NTTS).* December 21, Brussels, Belgium.

Daas, Piet J.H.; Marko J.H. Puts;. (2014). Social Media Sentiment and Consumer Confidence. *European Central Bank Statistics Paper №5, Frankfurt, Germany, 5*.

De Jonge, E.; Van Pelt, M.; Roos, M.;. (2012). Time patterns, geospatial clustering and mobility statistics based on mobile phone network data. *Discussion paper Statistics Netherlands, 4*.

ESSC. (2011). *European Statistics Code of Practice.* Retrieved from http://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-32-11-955.

ESSC. (2012). *Quality Assessment Framework.* Retrieved from http://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646.

IBM. (2014). *Big Data roadmap assessment, carried out on behalf of Statistics Netherlands.*

Struijs, P.; Braaksma, B.; at al;. (2014). *Official Statistics and Big Data.* Big Data and Society.

UNECE. (2013). *The role of Big Data in the modernisation of statistical production. Project plan for 2014 as approved by the High-Level Group for the Modernisation of Statistical Production and Services.*

UNECE. (2013). *What does Big Data mean for Official Statistics?. Paper produced by a Task Team on request of the High-Level Group for the Modernisation of Statistical Production and Services.*

Varian, H. R. (2014). Big Data: new tricks for econometrics. *Journal of Economic Perspectives, 28-29*.