

Някои проблеми при осигуряването на условия за формиране на представителни извадки

Драгомира Белчева*

Резюме: Настоящата статия е посветена на изследване на проблемите, свързани с моделирането и осъществяването на представителни извадкови проучвания в България. Третираны са основно трудности, свързани с необходимостта от наличие на изчерпателен адресен списък на единиците от генералната съвкупност за формиране на класическите модели на представителни извадки. Предложени са някои модификации, направени върху тях, с оглед адаптиране към условията на провежданите изследвания. Изтъкнати са предимствата и слабите страни на предложените решения. Разгледан е оригинален модел на териториална извадка. Моделът използва случайна извадка от географски координати за излъчване на извадка от индивиди. Подходящ е за използване при проучвания на съвкупности от индивиди, живеещи в нерегулирани местности, и за които е невъзможно съставянето на адресен списък.

Ключови думи: извадка, представителна извадка, неактуален списък.

JEL: C49.

Въведение

В литературата е широко разглеждана темата за проектирането на извадкови проучвания и свързаните с това

проблеми. Известно е, че при провеждането на такива проучвания изследователите се сблъскват с редица трудности, свързани главно с осигуряване на необходимите условия и предпоставки за реализиране на определен модел на извадка. Когато става въпрос за проучвания в социалната сфера, трудностите нарастват. Планирането на извадка поставя за решаване множество въпроси – какъв тип да бъде тя, какъв модел да се използва, какъв обем да има и други. От една страна, за да могат да се пресмятат равнищата на стохастична грешка, е задължително да се излъчи представителна (или още вероятностна) извадка.

От друга страна, задължително условие за формирането на такава извадка, е наличието на пълен и изчерпателен адресен списък на единиците от генералната съвкупност – нещо, което в общия случай изисква много ресурси, а понякога е и неосъществимо. Следователно, като пръв сериозен проблем, стоящ пред изследователя, планиращ представителна извадка, може да се изведе съставянето на списък на единиците от генералната съвкупност. За да бъде подходящ за основа на извадката, един списък на единиците на генералната съвкупност трябва да притежава следните свойства: а) всяка единица от генералната съвкупност да е записана в него точно по веднъж и всички единици от генералната съвкупност да фигурират в него; б) в

* Драгомира Белчева е докторант в катедра "Статистика и иконометрия" на УНСС, e-mail: dragomira_belcheva@abv.bg

списъкът да няма единици, които не принадлежат на генералната съвкупност; в) срещу всеки запис да има точен адрес (в най-общ смисъл), т.е. еднозначен начин, по който съответната единица може да бъде открита. Съставянето на такъв списък е предпоставка, необходима за осигуряване на представителност на извадката, а същевременно е задача, която трябва да съчетае използването на минимално количество ресурси и набавянето на достатъчно количество актуална информация.

Както при изследванията, провеждани през миналия век, така и при съвременните, съставянето на регистър на единиците от генералната съвкупност се оказва сред основните и най-трудно преодолими предпоставки при реализирането на извадкови изследвания. Както тогава, така и сега, наличието на качествен регистър е по-скоро изключение, отколкото правило.

Вторият проблем, който донякъде се препокрива с първия, е свързан с актуалността на данните в съществуващите регистри. Този проблем се отнася най-вече до спецификата на проучванията сред хора. Дори и да е възможно да се състави регистър на определена група, информацията в него много бързо остарява и става неизползваема – някои хора си сменят адреса, заминават за дълго извън страната, едни умират, други по естествен начин влизат в генералната съвкупност, навършвайки определена възраст, а трети я напускат по същата причина и т.н.

Трети проблем се оказва и незадължителният характер на повечето социални проучвания. С изключение на преброяването, всички останали проучвания имат доброволен характер, което налага правенето на множество компромиси от страна на изследователите със строгите принципи на вероятностния подбор. Колкото и добре да е планирана извадката, крайният резултат е проучване, направено само сред отзовалите се, т.е. сред съгласилите се

да участват, което, според класическата статистическа литература, посветена на извадкови изследвания, го прави непредставително.

Настоящото изложение се концентрира главно върху проблемите, свързани с осигуряване на условията за представителност и по-точно на необходимостта от наличие на адресен регистър на единиците. Като основа за изясняването им е използван опитът на социологическия отдел на Институт "Отворено общество" – София (ИОО). Без да претендира за изчерпателност, авторът се концентрира върху него главно поради три причини. *Първо*, възникналите трудности и проблеми са характерни за повечето съвременни проучвания в социалната и икономическа област. *Второ*, като част от екипа на ИОО, авторът е добре запознат с методологията на провеждане на всички проучвания на ИОО през последните 5 години и с всички препятствия, появили се при осъществяването им. *Трето*, авторът е пряк участник в разработването на някои оригинални решения на възникналите проблеми.

В зависимост от спецификата на разглежданите генерални съвкупности възникват различни проблеми и съответно се търсят различни подходи за моделиране и реализиране на съответните извадки. Като цяло, използваните извадкови модели са класически и често употребявани в социално-икономическите изследвания. В случая интерес представляват някои техни модификации и оригинални решения, предложени за справяне с конкретно възникнали проблеми.

1. Съществуващ, но непълен и неактуален списък

Наличието на неактуален или непълен списък на единиците на генералната съвкупност се очертава като основен проблем не само при реализирането, но още в процеса на организацията при планирането на

ресурсите, необходими за провеждане на представително извадково проучване. То води до лъжливо усещане от страна на изследователя, че разполага с необходимото, за да направи представителна извадка, и го кара да отдели малко ресурси за подготвянето му, а също и да планира модел на извадка, който се оказва неизпълним. Достатъчно често, особено в България, се получава така, че се съставят регистри на дадени съвкупности, но те не се поддържат и са напълно негодни за използване за каквито и да било статистически цели. Всеки изследовател, занимаващ се със социални изследвания, е установявал например, че няма актуален и публично достъпен регистър на фирмите, нито на лечебните заведения, нито на лекарите, нито на неправителствените организации и още много други.

Като конкретен пример за проблеми, възникващи при използване на непълен и неактуален регистър, е разгледано изследването *"Еконобус: условия за правене на бизнес в България"*¹. За определяне на генералната съвкупност е използван регистърът БУЛСТАТ. Интерес представляват само фирмите със седалища в 27-те областни града в България и в Ботевград (като представител на Софийска област). Планирано е да се излъчи проста случайна извадка, включваща по 100 фирми за всеки град. Следователно вероятността определена фирма да попадне в извадката е различна в зависимост от града, в който се намира. При подобни ситуации оценките за цялата съвкупност от фирми в областните градове се получават като претеглени величини. Избран е прост случаен подбор – първо, защото дава най-прости методи за оценки на параметрите, второ, заради лекотата на приложението му и, трето, поради факта,

че първоначално се приема, че е изпълнено условието за наличие на изчерпателен адресен списък на единиците.

Стратифицирана извадка по големината на фирмите евентуално би могла да даде по-прецизни оценки на параметрите по някои икономически индикатори поради по-малката дисперсия в сравнение с оценките, получени чрез прост случаен подбор при един и същи обем на извадката, но това предполага наличие на надеждна информация за стратификационния признак – например, годишен оборот, брой служители, обем на активите и т.н. За съжаление търговският регистър БУЛСТАТ не разполага с такива актуални данни. Добре известно е, че стратифициране по признак, който не дава вярна информация за единиците, по-скоро би навредил на получените оценки отколкото да подобри тяхната точност.

Основните трудности, възникнали при провеждането на това проучване, са свързани главно с търговския регистър. Проблемът с него е, че той не се поддържа добре и не предоставя актуална информация. Трябва да се отбележи, че по време на провеждане на теренната работа в ход беше процедура по пререгистриране на фирмите в нова база от данни с цел представяне на по-добра и по-актуална информация за тях, която вероятно в бъдеще ще може да се използва за статистически цели. В момента на конкретното проучване обаче новият търговски регистър все още беше непълен. За целите на проучването се наложи прилагането на процедура по верификация на данните, изтеглени от БУЛСТАТ – от една страна, всяка фирма, попаднала в извадката, беше проверена в новата база от данни, а ако не беше намерена в нея, се използваха странични източници за установяване на съ-

¹Това е представително изследване на стопанските субекти в България, тип омнибус, проведено от ИОО съвместно с Институт за пазарна икономика и Риск Монитор през октомври – декември 2010 г. Целта на изследването е да представят нагласите на бизнеса за икономическото положение в страната, влиянието на световната финансова криза и последиците от нея, очакванията за бъдещето развитие на бизнес климата в България.

ществуването ѝ като например Интернет, проверка на адрес и т.н. В течение на верификацията отпаднаха около 50 % от фирмите, първоначално попаднали в извадката поради невъзможност за осъществяване на контакт с тях. Последното наложи генериране на нова извадка и повторна проверка, за да може накрая да се изготви списък от 2800 фирми плюс 30 % резерви, общо 3640.

Заблудата, че БУЛСТАТ е подходящ за формиране на представителна извадка се оказва ключов проблем и води до забавяне на проучването, до оскъпяване и необходимост от използване на непредвидени човешки и технически ресурси. По-общо предложеният подход може да се представи по следния начин. За справяне с проблема *"неактуален регистър"*, първо, трябва да се излъчи извадка с достатъчно голям обем от наличния регистър и данните от нея да бъдат верифицирани от други източници. Избраната процедура има множество недостатъци. *Първо*, тя не гарантира, че всички единици на генералната съвкупност съществуват в списъка. *Второ*, като следствие от първото, тя не гарантира, че всяка единица от генералната съвкупност има една и съща известна вероятност да бъде изтеглена. *Трето*, получената извадка представя по-скоро съвкупността от фирмите, регистрирани в двата регистра, или поне тези, за които може да се намерят данни в информационното пространство. Изброеното до тук в известна степен компрометира представителността на извадката. От друга страна, верификацията на данните от наличния регистър се явява единствена възможност за провеждане на проучването. Получените резултати трябва да се тълкуват внимателно, отчитайки всички несъвършенства на подхода. Подобни проблеми се появяват и при изследване на други специфични или "редки" съвкупности – винаги се достига до липса на единен регистър на единиците на дадена съвкупност.

2. Невъзможност за съставяне на списък на единиците от генералната съвкупност

Когато обект на изследване е цялото население на една държава или дадена негова подгрупа, е необходимо наличието на добре поддържан адресен списък на населението. Ако такъв списък трябва да се изготви само за целите на едно проучване, това би било изключително скъпо и нецелесъобразно. Описаното е известна и стандартна ситуация при повечето маркетингови и политически проучвания. Обикновено се прибегва към използване на списъци с гнезда от елементи на изследваната съвкупност, каквито съществуват или е по-лесно да бъдат изготвени. За целта например могат да се вземат списъци с адреси, списъци с жилищни блокове, болници, училища, университети и т.н. Използването им вече предполага формиране на по-сложни извадкови модели като гнездови и стратифицирани извадки или комбинации от двете.

Проучването *"Деца на София"*² е пример за невъзможност за съставяне на изчерпателен списък от елементарните единици. В началото замисълът беше да се изучи мнението на всички "деца на София", т.е. деца на възраст от 10 до 18 години по определени актуални теми. Следователно, генералната съвкупност включва както учащите деца, така и тези, които по някаква причина не са записани в училище, въпреки че попадат в разглежданата възрастова група. За целта беше необходим списък на тези деца. Такъв може да се намери през системата ГРАО, която за съжаление не се поддържа добре и не може да се използва за нужните цели.

Съществуват няколко възможности за преодоляването на този проблем. Единият подход беше да се направи извадка от домакинства и да се проведе "специална селекция" (*screening*) за тези, в които има

² Това е проучване, представително за учениците 3-12 клас, записани в столичните училища. Провеждено е от ИОО съвместно с МОМН и Столична община през месец декември 2008 г.

дете на определената възраст. Това е добър вариант, но изисква обемът на извадката да е толкова пъти по-голям, колкото е делът на домакинствата в България с деца на необходимата възраст. Проучване с толкова голям обем изисква огромно количество ресурси – не само финансови, но и човешки, и като правило се прилага много рядко и само в краен случай. Затова идеята беше отхвърлена. Търсеше се друга събирателна за децата институция извън домакинствата.

Очевидното решение на проблема всъщност са училищата. *Първо*, съставянето на адресен списък с училищата в София е сравнително лесна задача, като се има предвид териториалната ограниченост. *Второ*, те концентрират много деца на едно място и следователно дават възможност за намаляване на разходите за провеждане на едно интервю чрез използване на гнездови модел на извадка. Основният проблем с излъчването на представителна извадка, използваща за център училище, е, че автоматично от генералната съвкупност се изключват децата, които не са обхванати по някаква причина от образователната система. Едно решение е да се раздели съвкупността на две подсъвкупности – ученици и други, за всяка от които да се излъчи отделна извадка, след което двете оценки да се обединят.

Възможен модел за съчетаване на данни от различни извадки е подробно разгледан в доклада на Стойкова-Къналиева, А. ([1], 2009). Решението е от по-голямата и с известни параметри подсъвкупност – тази на учениците, да се излъчи представителна извадка по метода на гнездовия подбор, като за гнезда се използват учили-

щата. В случая той дава свободата да се формира извадка от наличен списък, без да се налага изготвянето на такъв на всички ученици³. Като цяло, моделът на гнездови подбор изглежда очевидно, просто и добро решение, когато вместо списък на единиците от генералната съвкупност има наличен списък на по-големи групи от единици – гнезда.

С втората, по-малка съвкупност, на неучащите деца, за която е трудно да се набави информация, може да се направи проучване с извадка със сравнително малък обем по друг модел, дори и той да не води до представителни данни. Например, да се излъчи извадка от домакинства и коментираната вече "специална селекция" или да се потърсят допълнителни източници на информация – Агенция за защита на детето, социални служби и т.н. Възможно е също необхванатите да се търсят сред обособени групи, в които се предполага, че има по-висока концентрация на неучащи. Или да се направи извадка на принципа на човешките вериги⁴, стартирайки с малко на брой известни случаи.

Изброените възможности обаче водят до непредставителна за разглежданата подсъвкупност извадка. Както е коментирано и в цитирания доклад ([1] Стойкова-Къналиева, А, 2009), предвид това, че непредставителната извадка представлява сравнително малка част от цялата извадка, данните от нея могат да се използват в съчетание с тези от представителната и не водят до особено влошаване на получените оценки. Разгледаният модел предлага добро решение в множество ситуации, но в проучването "Деца на София" беше отхвърлен по няколко причини.

³ Трябва да се отбележи, че предимството да не се съставя списък от единици, а от гнезда, има две страни. От една страна, това облекчава изследователите, но, от друга, води до влошаване на точността на оценките поради еднородност на единиците в отделните гнезда. Изтъкнатите организационно-технически предимства на гнездовата извадка обаче позволяват при почти еднакви разходи включването на повече единици в извадката, което като цяло води до по-точни оценки.

⁴ Вариант на такава извадка е известният модел на "снежна топка".

Първо, трудно е да се определи обемът на изследваната съвкупност, респективно и големината на необходимата извадка. *Второ*, в съвкупността има огромна дисперсия по множество признаци – възраст, пол, ниво на отпадане от училищната система, основен говорим език, ниво на умствени възможности и т.н. *Трето*, част от съвкупността се състои от деца с физически и/или психически затруднения и следователно провеждането на интервю с тях би било изключително трудно. *Четвърто*, като следствие от трето, инструментите за събиране на данни от изброените групи би трябвало да бъдат подходящо адаптирани за тях и това допълнително усложнява задачата. *Пето*, голяма част от поставените изследователски въпроси и хипотези касаят училищния живот и взаимоотношения и респективно не биха били подходящи за тези извън училищната система.

От друга страна, концентрирането само върху подсъвкупността от ученици и провеждането на изследване само в нея би улеснило многократно работата и би представило достатъчно добре нагласите и мненията на основната част от генералната съвкупност. В описания случай подходът за справяне с проблема "невъзможност за съставяне на списък" беше да се преформулират целта и задачите на изследването и да се промени обектът. В примера основната цел на изследването се преформулира от проучване на мнението на всички деца, попадащи в определената възрастова група, до проучване на мнението само на учениците от трети до дванадесети клас, записани в училища в град София.

3. Метод на "случаен обход" – предимства и недостатъци. Преминване от извадка на домакинства към извадка от индивиди

Предизвикателство пред социологическите проучвания на пълнолетното население е да се излъчи качествена и същевременно икономически ефективна извадка – проучването трябва да е лесно и в сравнително кратки срокове осъществимо, защото резултатите бързо остаряват. Класическо решение, използвано от все повече маркетингови и социологически агенции, е излъчването на квотна по основните демографски характеристики извадка. Тя има редица предимства – не се нуждае от списък на генералната съвкупност, евтина е за изпълнение и, най-вече, изискваща минимално време за реализация. Основният ѝ недостатък, е, че не е представителна и получените оценки важат само за фактически наблюдаваната част от населението. Представителната извадка, от своя страна, е по-сложна и скъпа за изпълнение, но дава възможност за генерализиране на получените резултати за цялото население. И тук най-често използваният подход е да се вземат различни адресни регистри – например избирателни списъци, данни от ГРАО, от пощите и т.н. Основният недостатък на всички адресни регистри, особено в България, е, че като цяло са с остаряла или неточна информация, нещо, което многократно беше споменато в изложението. Въпреки законовите регламенти, не са единични случаите, в които хората са регистрирани на едно място, а в действителност живеят на друго. Следователно, част от единиците (тези, които не живеят на постоянния си адрес) имат нулева вероятност да попаднат в извадката. Изключването им води до повишаване на нестохастичния компонент на грешката, свързан с наличието на необхванати⁵. Класическото решение

⁵ Източниците на нестохастичен компонент на грешката са подробно разисквани в Съикова, И., Стойкова-Къналиева, А., Съикова, С., 2002 г., стр. 44.

В случая е вместо извадка от индивиди, да се формира такава от адреси и като единици от извадката да се разглеждат живеещите на съответните адреси.

За конкретност са разгледани трудностите при планиране на национално представителни изследвания на населението от ИОО⁶. Използваният модел е на двустепенна гнездова извадка, стратифицирана по два признака – административна област и тип населено място. Генералната съвкупност се състои от пълнолетното население в страната. Избран е стратифициран модел, *първо*, за да могат резултатите да се интерпретират не само на национално ниво, но и в отделните подгрупи – региони за планиране (NUTS 3), области, тип населено място. *Второ*, поради предположение за хомогенност на единиците по някои икономически и социални показатели, се очаква по-малко разсейване на единиците, а оттам, при равни други условия, и намаление на стохастичната грешка в отделните страти.

От друга страна набавянето на информация за разпределението на населението по изброените признаци е лесно изпълнимо и не води до повишаване на общата цена⁷. Гнездовият модел е избран *на първо място*, защото значително намалява разходите за изготвяне на списък на единиците от съвкупността, като стеснява задачата до създаване на списък от групи адреси. *Второ*, той дава възможност за включване на повече единици в извадката поради факта, че значително намалява транспортните и времевите разходи. Въпреки че като правило увеличава стохастичната грешка, гнездовият модел в случая влияе положително, като позволява увеличаване броя на изследваните единици в отделните страти и, като цяло, подобрява точността на оценките.

Във всяка страта извадката се провежда на три стъпки. На първа стъпка се избират населените места с вероятност, пропорционална на големината им. На втора стъпка се избират гнезда, определени от клоновете на Български пощи, които обхващат група от адреси в близост едни до други. Причините за използването на списък с адреси от пощите са основно две. Първо, информацията в тях е сравнително актуална и, второ, базата е сравнително лесно достъпна. Основният и недостатък е, че пощенските адреси обхващат няколко жилищни адреси, например група от блокове, части от квартали или улици и т.н.

Базата се използва за определяне на гнездата и за генериране на един случаен адрес. След това се прилага класическата процедура на "случаен обход – започва се от първия адрес, прескачат се пет и се посещава всеки шести адрес на домакинство или още – "Врата на дом". Ако на съответното място има жилищен блок, се прави обход на всеки шести апартамент, а след като се изчерпят всички апартаменти в блока, се продължава по същата схема, докато се направят необходимият брой интервюта. Ако има къщи, се процедира по аналогичен начин като се обхожда всяка шеста къща. Описаната процедура се изпълнява само в рамките на предварително очертаните от пощенския код граници. Интервюира се лицето от домакинството, което е на 18 или повече години и има най-скоро предстоящ рожден ден. Ако откаже или отсъства за дълъг период, се преминава на следващия адрес. Не се допуска замяна на едно лице от домакинството с друго при никакви условия.

Основен проблем при такъв тип проучвания е даването на равен шанс на всички единици от генералната съвкупност да попаднат в извадката. Целта е да се постиг-

⁶ Става въпрос за многоцелеви проучвания от тип "омнибус", провеждани от ИОО регулярно през няколко месеца.

⁷ Данни могат да бъдат взети например от регулярните преброявания на населението, провеждани от НСИ.

не баланс между ресурсоемкост и точност на резултатите. Описаната процедура предизвиква множество въпроси. Първо, трябва да се изясни как се постъпва, ако на един адрес, в едно жилище, живеят повече от едно домакинство. Второ, когато домакинството вече е определено, как се подбира лицето или лицата, които да бъдат интервюирани.

Класически решения на тези проблеми дава Лесли Киш ([8] Kish, L., 1949). Той разглежда две възможности за избор. *Първо*, когато няма нужда да се прави избор, ако предварително е определено кой член от домакинството да участва, например главата, най-възрастният и т.н., или ако домакинството е единицата, представляваща интерес за анализа. *Второ*, когато повече от един член на домакинствата принадлежи на генералната съвкупност (например всички пълнолетни лица), трябва да се реши как да се избере респондент измежду всички членове. Киш ([8] Kish, L., 1949) предлага да се избира само един член на домакинството по определена схема.⁸ При този подход на членовете на "големи" домакинствата се дава по-малък шанс да бъдат избрани в сравнение с членовете на домакинство с по-малко хора⁹.

От друга страна, добавянето в извадката на нов член на домакинството е по-малко ценно за точността на резултатите от добавянето на човек от друго домакинство. Това е така, защото членовете на едно домакинство могат да си повлияят, обсъждайки въпросите от интервю-

то, а дори и въпросите да са за нагласи, корелацията между отговорите на такива индивиди е доста висока. При първото от разглежданите проучвания беше направен опит за модификация на схемата на Киш, при която във всяко домакинство се избира един, двама или трима респонденти, в зависимост от броя на възрастните членове¹⁰. Идеята беше да се намали ефектът на недостатъчно представяне на членове на многочленни домакинства. Резултатите от този опит са доста противоречиви.

По данни от последното преброяване¹¹ в България средният брой членове на домакинство е 2,4, а около 79,4 % от домакинствата са едно-, дву- или тричленни. Следователно, ако предположим, че домакинствата в извадката са разпределени почти като тези в генералната съвкупност, вероятността да е избрано четири- или повече членно домакинство е около 20 %. Като се вземе предвид, че обект на интерес в случая са само възрастните членове, което ще увеличи още повече дялът на домакинствата с до трима възрастни, то случаите, в които се налага избор на повече от един член, стават съвсем малко.

При анализа на данните от проучването се забелязва един важен проблем, свързан с човешкия фактор. Използвайки възможността да направят повече анкети с по-малко усилия, анкетаторите провеждат по две интервюта дори и в домакинства, в които предписанието е да се направи само едно и, следователно, дялът на домакинствата, в които е

⁸ Схемата, предложена от Киш, предполага съставяне на таблица с описание на домакинството за всяко такова, попаднало в извадката. Таблицата съдържа възраст, пол и връзка с главата. Членовете на домакинството се номерират с последователни числа, като се започне от мъжете в ред по намаляваща възраст, следвани от жените в същия ред. Има предварително изготвени 8 таблици, посочващи номера на онзи член на домакинството, който да бъде интервюиран, в зависимост от големината му. На всеки въпросник е прикачена точно една такава таблица за подбор и анкетаторът взема от нея номера на човека, когото да интервюира.

⁹ Твърдението важи за случаите, в които домакинствата се изтеглят с равна вероятност.

¹⁰ Схемата е разработена от г-р Алексей Пампоров и е следната: 2 - 3 лица над 18 г. – анкетира се едно лице от домакинство; 4 - 6 лица над 18 г. – анкетират се две лица от домакинство; 7 и повече лица над 18 г. – анкетират се три лица от домакинство.

¹¹ <http://www.nsi.bg/census2011/pagebg2.php?p2=175&sp2=192&SSPP2=194>

проведено по повече от едно интервю, става по-голям от предвидения. Поради тази причина моделът не е повторен при следващите изследвания. Вместо него се прилага предписанието на Киш да се избира само по един човек от домакинство. Но неговите схеми за подбор са заместени със значително по-прост метод, който също осигурява приблизително равен шанс на всеки от домакинството да бъде включен в извадката – "най-близък предстоящ рожден ден" ([12] Salmon, C. & Nichols, J., 1983, [5] Gaziano, C., 2005). Изборът е еднозначен и дава на всеки индивид от домакинството един и същи шанс да попадне в извадката.

Има и аргументи, оспорващи случайността на метода ([5] Gaziano, C., 2005). Ако изследваните характеристики на единиците от генералната съвкупност са свързани по някакъв начин с рождената дата, това може да окаже въздействие върху получената информация, като се вземе предвид и времето на провеждане на теренната работа. Все пак методът на "най-близкия предстоящ рожден ден" има повече предимства отколкото недостатъци, особено при социологически проучвания по актуални социални или политически въпроси. Най-вече той е бърз, лесен и разбираем начин за подбор на респондент и избягва необходимостта от съставяне на таблици с членовете на домакинството. Като се пресметнат разпределенията на попадналите в извадката от конкретното изследване индивиди по пол, възраст и образование, те с голяма точност се доближават до разпределението на цялото население по тези признаци (също по данни от преброяването), което, от своя страна, дава гаранция за работоспособност на модела.

4. Когато случайният обход не може да бъде проведен

Както беше показано, класическото решение за формиране на извадка от домакинства е методът на случаен обход. Той дава добри резултати при сравнително неголеми разходи за формиране на извадката. Има и случаи, в които класическото решение е неприложимо. Например, когато генералната съвкупност е съставена от домакинства, живеещи в зони извън регулация, в квартали или местности, които са обособени почти във формата на лабиринт. Дотук бяха представени леки модификации на класически и очевидни методи за преодоляване на трудностите при изготвяне на представителни извадки. Сега ще бъде представен един оригинален метод, който може да бъде използван в множество ситуации, при които предходно описаните процедури са неприложими.

За изясняване на проблеми от подобен характер е разгледано изследването *"Превенция на ранните бракове сред жителите на сегрегирани квартали"*¹². Причината за разглеждането на този пример е в особеностите на всички социологически проучвания, свързани с дадени рядко срещани или специфично обособени подсъвкупности на населението. В литературата има разгледани редица модели на извадки на единици, притежаващи определена рядка характеристика ([6] Kalton G., Anderson, D., 1986, [13] Sudman S., Kalton G., 1986). В България особено актуален е въпросът с изучаване на проблемите в сегрегирани квартали – т.нар. "гета". Да се направи представителна извадка на домакинства, населяващи сегрегирани квартали, е предизвикателна и твърде трудна задача.

¹² Проектът "Превенция на ранните бракове сред жителите на сегрегирани квартали" е реализиран от ИОО по поръчка на Центъра за межкултурен диалог и толерантност (ЦМЕТ "Амалие"). Основната цел на изследването е да се изучат нагласите на населението, живеещо в сегрегирани квартали по отношение на нуждата от образование, ранното встъпване в брак или съжителство, произтичащото от това спиране на момичета от училище, ранните бременности и последвалите проблеми.

За излъчването на представителна извадка, трябва да се изготви списък с всички обособени квартали и с информация за големината им. За тази цел е необходимо предварително да се дефинира какво означава сегрегиран квартал и колко голяма трябва да е общността, за да се нарече така. Това, от своя страна, води до необходимост да се даде определение на това колко компактно трябва да живее дадена общност, за да може да се нарече местообитанието ѝ квартал. Освен това трябва да е ясно каква е разликата между "обикновен" и "сегрегиран" квартал и т.н. Изброеното до тук е само илюстрация на множеството задачи, които следва да бъдат решени преди излъчването на каквато и да е вероятностна извадка.

В разглеждания пример е излъчена гвустепенна гнездова извадка. На първа степен, от експертна база от данни, съдържаща всички обособени квартали в страната с население повече от 150 човека (в ролята на гнезда), се прави случайна извадка на 60 такива, претеглена спрямо населението им. Втората стъпка от алгоритъма предполага определянето на броя единици от всяко гнездо, които да са в извадката. За излъчването на представителна извадка от домакинства трябва да се познава броят на единиците, принадлежащи на всяко от избраните гнезда. И точно в осигуряването на това условие се явява основният проблем на реализиране на модела.

Характерните особености на сегрегираните квартали пораждаат редица трудности. Например, в тези квартали понякога има части, които са извън регулация, няма обособени улици или каквато и да било инфраструктура, част от жилищата са незаконни и следователно обитателите им нямат адресна регистрация и не са включени в никакви списъци. Друга особеност е, че на един адрес може

да са регистрирани няколко домакинства, нямаша нищо общо помежду си и обитаващи различни жилища. Основният недостатък при прилагане на процедурата по "случаен обход" в случая е, че би дал на анкетьорите твърде голяма свобода на избор и възможност от извадката да бъдат изключени част от единиците, например най-бедните и неузледни части на квартала, които са непривлекателно място за теренна работа.

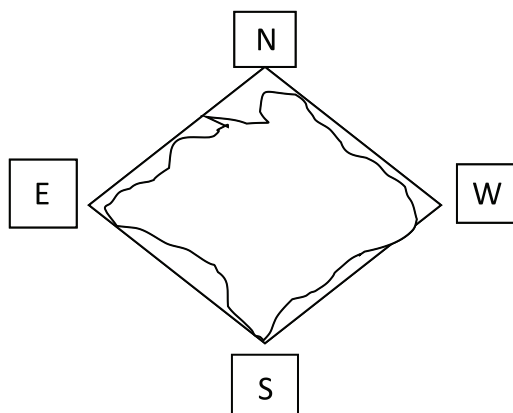
Поради изброените причини е необходимо да се търси друг подход за излъчване на извадка. Киш ([7] Kish, L., 1965) дава алгоритъм за изготвяне на териториална извадка, предполагаща разделяне на изследваните райони на блокове и описание на всички домакинства, попаднали в извадката от блокове. В конкретния случай дори и този подход е трудно приложим, тъй като описването на домакинствата в сегрегираните квартали, дори и само в определени техни части, е невъзможна задача поради изброените вече специфики. Даже и внимателно да обхождат терена, няма начин, по който анкетьорите да опишат домакинствата, така че на следващата стъпка някой друг да може да ги идентифицира за интервю.

Разработен е и оригинален модел на географска извадка (по-точно извадка от географски координати), който преминава през следните етапи:

1. Определят се 4 крайни точки на квартала – най-северна, най-южна, най-западна и най-източна, и техните географски координати:

$E = (X_E, Y_E)$, $W = (X_W, Y_W)$, $N = (X_N, Y_N)$, $S = (X_S, Y_S)$.

Полученият четириъгълник включва и местностите извън регулация, незаконните постройки и временните жилища. Така се получава четириъгълник, който е описан за многоъгълника (или най-общо фигурата), представляващ границите на квартала.



Фигура 1.

2. Определят се интервалите между най-голямата и най-малката географска ширина и дължина, дефиниращи границата на квартала. Следователно всяка точка $X_i \in (XW;XE)$ и $Y_i \in (YW;YE)$ с географска ширина между крайната западна и крайната източна точка се намира в границите на квартала. Аналогичното е вярно и за точките с географска дължина между най-северната и най-южната точка.

3. Чрез проста случайна извадка се излъчват точки $P_i = (X_i; Y_i)$ – с географски координати, намиращи се в дефинираните на предходната стъпка граници. Списъкът съдържа поне два пъти повече точки, отколкото интервюта трябва да се направят, за да се осигурят и резервни адреси.

4. По този начин се прави списък от координати, лежащи в четириъгълника ESWN, определен от краищата на квартала. Точките P_i , които лежат в него, но не са част от квартала, се изключват от списъка. На останалите точки съответстват географски координати, принадлежащи на квартала.

5. За всяка точка P_i се определя жилището (къща, блок, фургон, барака и т.н.), което се намира най-близо до нея и там трябва да бъде направено интервю. Важно е да се отбележи, че при оп-

ределяне на жилището се има предвид входната врата, т.е. става въпрос за най-близката "Входна врата на дом". По изложената схема всяка точка P_i има една и съща вероятност да бъде избрана. Това е вероятността от даден затворен интервал върху реалната права да бъде изтеглено произволно число. При определяне на домакинството, което да попадне в извадката, трябва да се има предвид, че дори дадена точка P_i да попадне в двора на една къща, тя (точката) може да се окаже по-близо до входната врата на съседна къща. Или, иначе казано, попадането на точката в нечий двор не означава задължително, че тази къща е от извадката. Това уточнение е от ключова важност, защото в противен случай има по-голяма вероятност в извадката да попаднат къщите, имащи по-голяма хоризонтална площ. Друго важно уточнение е, че всеки дом може да бъде избран само по веднъж (т.е. извадката е без връщане) и, следователно, ако се окаже, че дадена точка е най-близка до вече избрана врата на дом, то тя се отстранява от извадката.

6. Ако в разглеждания квартал има жилищни блокове, трябва да се приложи схемата за комбиниране на две различни по своя характер извадки. Като се има предвид приблизителният брой на жителите на квартала, се пресмята частта на живеещите в блокове. Извадката се стратифицира пропорционално – на живеещи в блокове и останалите. Изготвянето на груба оценка на броя на жителите на блокове обикновено не е трудна задача, тъй като в повечето случаи блоковете не са много и следователно съставянето на списък и описание на брой входи, брой етажи и съответно брой апартаменти не изисква особени усилия. При веднъж съставен такъв списък, оценката на приблизителния брой домакинства, обитаващи блокове, става бързо и лесно. Важно е тази стъпка

да не се пропуска, защото negliжирането ѝ би довело до опорочаване на извадката. Ако с жилищните блокове се процедурира по начина, описан в разгледания модел, домакинствата от по-високите сгради биха получили значително по-малка вероятност да бъдат изтеглени от тези в къщите или по-ниските кооперации.

Основното предимство на предложения модел е, че дава много добра схема за подбор на домакинства, без да изисква какъвто и да било предварителен списък и осигурява условия за формиране на представителна извадка. Като недостатък може да се изтъкне сложността на изпълнението му. Важно е да се отбележи, че ресурсите, необходими за изготвянето му, са по-скоро от техническо естество. При наличие на добра географска карта в електронен формат, процесът не е твърде времеемък. При първото тестване на модела екипът на ИОО не разполагаше с подходяща карта и се наложи замерването на координатите, както и вторичното описване на жилищата да се реализират в самите квартали от хора, т.е. налага се своеобразно изготвяне на географски карти. Това изисква голямо количество човешки ресурс, както и закупуване на специализирана техника (GPS системи). При следващи приложения на вече разработения теоретичен модел се използват електронни географски карти, което многократно намалява разходите и времето за изпълнение.

Друго важно уточнение е, че предложеният модел е най-подходящ за райони, в които жилищата са разположени в хоризонтала – т.е. еднофамилни къщи, каравани, временни постройки, бараки, бунгала и т.н. За квартали, в които преобладава вертикалното разположение на жилищата, т.е. стандартните блокове, използването му не е препоръчително. Много по-лесен вариант би бил да се опишат жилищните блокове и директно да се направи извадка от тях в качеството им на гнезда

или да се приложи вече описаната схема на "случаен обход". Без да е универсален, географският модел предлага добро решение в случаите на нерегулирани и трудно описуеми жилищни райони. Той е подходящ за приложение не само в сегрегирани квартали, но и в селски райони, в които съставянето на списък с адреси е също в голяма степен непосилна задача.

* * *

Въз основа на разгледаните примери от реални социологически проучвания може да се заключи, че проблемът със съставяне на адресен списък на единиците от генералната съвкупност е основен при реализирането на повечето проучвания. В зависимост от характера на генералната съвкупност има различни причини, поради които е трудно или невъзможно съставянето на такъв списък. Оказва се, че в България голяма част от обектите, които представляват интерес за съвкупностно изследване, не се водят на отчетност и е трудно да се направят дори приблизителни оценки за разпределението, за броя им или за други техни характеристики. Това прави прилагането на класическите модели за представителни извадки доста трудно.

Съставянето на адресен списък на единиците на генералната съвкупност би следвало да бъде грижа от първостепенно значение за изследователя и за него да се предвиждат достатъчно времеви, технически и финансови ресурси. Класическите решения, прилагани в повечето социологически проучвания, като "специална селекция", "случаен обход", избор на един член от селектирано домакинство, предлагат добри решения, когато единиците не могат да бъдат изчерпателно описани. Важно е обаче да се имат предвид недостатъците от прилагането им, когато се правят оценки на параметрите на генералната съвкупност от извадки, получени чрез тези способы.

Прилагането на модела за разделяне на съвкупността на подсъвкупности и излъчване на различен тип извадки от всяка, без съмнение, предлага добра възможност за изучаване на явления, които не биха могли да бъдат проучени по друг начин. Без да бъде универсално решение, предложеният модел на географска извадка дава възможност за излъчване на представителна извадка, когато повечето други методи са неприложими. В много случаи при провеждане на социологически проучвания, изследователите се сблъскват с невъзможност за излъчване на представителна извадка поради липса на достатъчно данни за единиците на изследваната съвкупност.

Въпреки, че не са обект на интерес в настоящето изложение, непредставителните модели на извадкови проучвания може да дават добро решение, когато е напълно невъзможно да се състави списък на генералната съвкупност. Дори и да не може да се оцени размерът на допуснатата грешка, по данни от непредставителни извадкови проучвания може да се направят добри изводи в зависимост от конкретната тематика. Много често се налага да се прави компромис между качеството на получените резултати и ресурсите, необходими за провеждане на изследването. За да бъдат смислено изразходвани средствата за конкретно проучване и за да бъдат достоверни получените резултати, е от изключително значение данните от него да бъдат тълкувани, вземайки предвид всички условности и недостатъци на излъчената извадка.

Литература

1. Стойкова-Къналиева, А., 2009. За възможностите да се съчетават данни от различни по своя характер статистически изследвания.

2. Стойкова, И., А. Стойкова-Къналиева, Стойкова, С., 2002. Статистическо изследване на зависимости, С., Университетско издателство "Стопанство".
3. Стойкова И., Б. Чакалов, 1986. Методология и методика на емпиричното социологическо изследване, София.
4. Цонев В., Основи на репрезентативните изучавания.
5. Cochran W.G., "Sampling Techniques", John Wiley & Sons, 1977.
6. Gaziano, C., "Comparative Analysis of Within-Household Respondent Selection Techniques", *The Public Opinion Quarterly*, Vol. 69, No. 1 (Spring, 2005), pp. 124-157.
7. Kalton G., Anderson D., "Sampling rare populations", *Journal of the Royal Statistical Society. Series A (General)*, Vol. 149, No. 1 (1986), pp.65-82.
8. Kish L., "Survey Sampling", John Wiley & Sons, 1965.
9. Kish L., "A Procedure for Objective Respondent Selection within the Household", *Journal of the American Statistical Association*, Vol. 44, No. 247 (Sep., 1949), pp. 380-387.
10. Levy P., Lemeshow S., "Sampling of Populations – Methods and Applications", John Wiley & Sons, 2008.
11. Nemeth, R., "Respondent Selection within the Household – A Modification of the Kish grid".
12. Salmon, C. & Nichols, J. "The Next-Birthday Method of Respondent Selection", *The Public Opinion Quarterly*, Vol. 47, No. 2 (Summer, 1983), pp. 270-276.
13. Sudman S., G. Kalton, "New Developments in Sampling of Special Populations", *Annual Review of Sociology*, Vol. 12 (1986), pp. 401-429.
14. Национален статистически институт, www.nsi.bg/census2011/