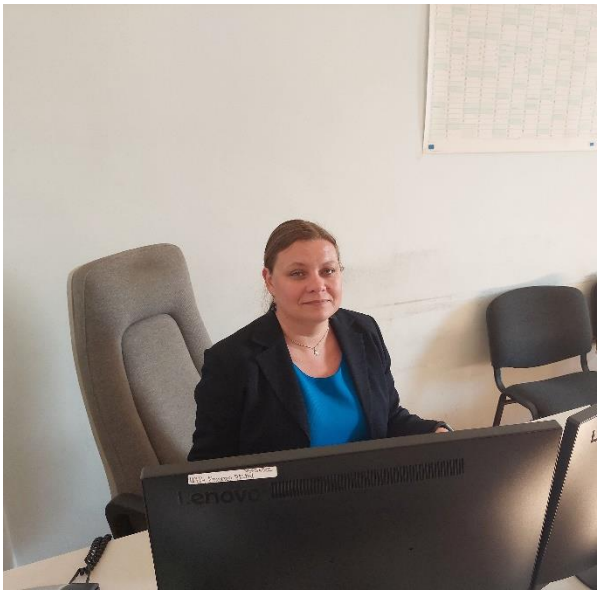
МЕТОДИ ЗА СТАТИСТИЧЕСКО СЪЧЕТАВАНЕ НА ДАННИ ОТ СЛОЖНИ ИЗВАДКОВИ ИЗСЛЕДВАНИЯ

*Цвета Цонкова**



I. Въведение

В последно време се отдава голямо значение на създаването на нови индикатори и инструменти за статистическо наблюдение, които да са в състояние да отговорят на нарастващите нужди от информация в различни аспекти от социално-икономическата област. В доклада на Европейската комисия относно измерването на икономическите резултати и социалния напредък (Stiglitz, Sen, Fitoussi[1]) се набляга на необходимостта от преразглеждане и актуализиране на настоящата система за събиране на статистически данни с цел справяне с новите обществени предизвикателства и в подкрепа на разработването на нови политики за решаване на проблемите в социално-икономическата област. Оттук произтича и необходимостта от наличието на интегрирана статистическа информация, която да обхваща различни аспекти на социално-икономическото развитие на обществото.

* Държавен експерт, отдел „Статистика на условията на живот“, дирекция „Демографска и социална статистика“, НСИ, e-mail: TTsonkova@nsi.bg

[1] https://ec.europa.eu/eurostat/documents/8131721/8131772/Stiglitz-Sen-Fitoussi-Commission-report.pdf

Официалната социална статистика е организирана около провеждането и анализа на специфични изследвания, обхващащи голяма част от нуждите на потребителите на тази информация: доходи, потребление, здравеопазване, образование, пазар на труда, използване на информационно-комуникационни технологии, социално включване и други. Въпреки това поради финансови и други пречки нито едно изследване не може да обхване самостоятелно всички посочени аспекти. В този контекст настоящият процес на модернизация на социалните изследвания се фокусира върху повишаването на общата им ефективност (от гледна точка на събирана информация, натовареност на респондентите, намаляване на разходите за провеждане), своевременното реагиране на нуждите на потребителите и подобряването на аналитичния потенциал на данните чрез използването на интегрирана система за социални изследвания.

Статистическото съчетаване на данни (известно още като синтез на данни, сливане на данни или синтетично съчетаване) е подход за предоставяне на обща статистическа информация на базата на променливи и показатели, събрани чрез два или повече източника. Източниците на информация могат да бъдат различни извадкови или изчерпателни изследвания, административни източници и/или „големи данни“. Потенциалните ползи от този подход са във възможността за засилване на допълващото използване и анализ на съществуващите източници на данни (например статистическа информация, която обхваща широк спектър от социално-икономически аспекти - бедност, заетост/безработица, потребление на домакинствата, здравен статус и др.) без допълнително увеличаване на разходите и на натоварването на респондентите. Следователно статистическото съчетаване на данни може да бъде разглеждано като инструмент за повишаване на ефективността на използване на информацията чрез прилагане на настоящите и утвърдени начини за събиране на данни.

Макар широко обсъждана и изследвана в световен мащаб, темата за статистическото съчетаване на данни от извадкови изследвания не е подробно разработвана в България. Настоящата статия има за цел да запознае читателите с методите за съчетаване на данни от сложни извадкови изследвания, които биха могли да бъдат приложени за разработването на модели за съчетаване на статистическа информация от различни социални изследвания, с което да се постигне изграждането на по-пълна картина на социално-икономическата действителност в България.

II. Същност на статистическото съчетаване на данни

В днешно време вземането на решения изисква колкото е възможно по-богата и навременна информация. Необходимата информация за управлението на икономиката и обществото може да бъде получена чрез провеждане на подходящи статистически изследвания, но в много случаи този подход е труден и неподходящ за постигане на поставените цели поради наличието на определени ограничения.

Когато събраните данни се отнасят за едни и същи единици (лица, домакинства, предприятия), различните набори от данни с информация за различни променливи могат да бъдат комбинирани помежду си чрез директно свързване на данните (data linkage), т.е. едни и същи единици в различните набори от данни се „срещат“ помежду си. За съжаление, свързването на данните невинаги е възможно, тъй като често различните набори от данни съдържат различни единици, за които няма достатъчна идентификационна информация (например липсващо ЕГН или ЕИК), която да се използва за „срещата“ на данните, които принадлежат към една и съща единица. Обикновено данните могат да бъдат „срещнати“ само за малък брой припокриващи се единици.

Възможното практическо решение на гореизложените ограничения е да се използва възможно най-обширно цялата информация, която вече е налична в различни източници на данни, т.е. да се извърши статистическо съчетаване на вече събрана информация (D’Orazio, 2006). Статистическото съчетаване на данни може да се използва, когато различните набори от данни съдържат различни единици, но с набор от общи (основни) променливи. Основната цел на статистическото съчетаване е да се оцени възможно най-добре връзката (например корелация, съвместно разпределение, условно разпределение и др.) между променливите, които не се срещат едновременно в различните набори от данни (целевите променливи). Например ако имаме набор от данни с информация за нивото на образование на лицата, техните пол, възраст и местоживеене и друг набор от данни с информация за професията на (други) лица, техните пол, възраст и местоживеене, целевите променливи са ниво на образование и професия. Можем да използваме информацията за пола, възрастта и местоживеенето, за да направим оценка на връзката - например на съвместното разпределение на образователното ниво и професията.

Най-общата рамка на статистическото съчетаване на данни може да бъде описана по следния начин:

Имаме два набора от данни A и B, които съдържат общи променливи $X = (X_1, X_2, \dots, X_p)$. Освен това наборът от данни A съдържа променливата Y, а наборът от данни B - променливата Z. Единиците в A и B са различни една от друга, като A съдържа n_A брой единици, а B - n_B брой единици (табл. 1).

1. Представителен набор от данни за статистическо съчетаване на данни от два източника

y_{11} y_{a1} y_{n_A1}	$x_{11} \dots x_{1p}$ $x_{a1} \dots x_{ap}$ $x_{n_A1} \dots x_{n_Ap}$		Набор от данни A
	$x_{11} \dots x_{1p}$ $x_{b1} \dots x_{bp}$ $x_{n_B1} \dots x_{Bp}$	z_{11} z_{b1} z_{n_B1}	Набор от данни B

Сивите полета в табл. 1 представляват ненаблюдаваната променлива съответно във всеки набор от данни. Целта на статистическото съчетаване на данни е да се направят изводи за връзката между целевите променливи Y и Z. Примери за тази връзка могат да бъдат съвместното разпределение на $f(Y, Z)$, корелационна матрица, таблица на двумерното разпределение или пълен синтетичен набор от данни.

Тази цел може да бъде постигната чрез оценяване на съвместното разпределение $f(X, Y, Z)$ на наборите от данни A и B, които се приемат като извадки от независими и идентично разпределени наблюдения от същото разпределение $f(X, Y, Z)$.

III. Методи за статистическо съчетаване на данни от сложни извадкови изследвания

В зависимост от поставената цел методите за статистическо съчетаване на данни се класифицират като макро- или микроподходи (D'Orazio, 2006). Също така в зависимост от допускането за съвместното разпределение те се обозначават като параметричен или непараметричен подход.

Първо, резултатът от метода за статистическо съчетаване може да бъде директно насочен към съвместното разпределение $f\ (X,\ Y,\ Z)$ или неговите характеристики (например съвместното разпределение на целевите променливи $f\ (Y,\ Z)$ или коефициента на корелация между тях). Това се нарича **макроподход.** В противен случай фокусът може да бъде поставен върху получаването на пълен синтетичен набор от данни (т.е. **микроподход**).

Второ, може да се приеме, че съвместното разпределение $f\ (X,\ Y,\ Z)$ се получава от параметричен набор от разпределения, т.е. приема се, че $f\ (X,\ Y,\ Z)$ се основава на фиксиран набор от параметри $\theta = f\ (X,\ Y,\ Z \mid \theta)$. За съвместното разпределение $f\ (X,\ Y,\ Z)$ може да се приеме и непараметричен модел с цел да се избегнат твърде строгите предположения и да се получи повече гъвкавост при използването на различни методи за статистическо съчетаване на данните. Когато методът за статистическо съчетаване на данни използва както параметричен, така и непараметричен модел в своята процедура, той се нарича „смесен подход“.

Елементарни микроподходи

Елементарният подход за статистическо съчетаване на данни от две сложни извадкови изследвания се състои в прилагане на непараметрични методи (метод хотдек на разстоянието, случаен или рангов метод хотдек), без да се взема под внимание моделът на извадката или теглата. След като бъде получен синтетичният набор от данни, статистическите анализи се извършват, като се вземе предвид моделът на извадката, който стои в основата на набора от данни, избран като получател (A), и съответното тегло на единиците в изследването (w_A).

Моделът на извадката може да се вземе предвид, като се формират донорски класове чрез използването на дизайн променливите (например стратифициращите променливи) заедно с най-подходящите общи променливи (Andridge and Little, 2010). Има вероятност този подход да не може да бъде приложен, ако дизайн променливите са само частично или изобщо не са налични. Освен това дизайн променливите, използвани при едно изследване, може да не са налични в другото и обратно, какъвто е обикновено случаят, когато данните се отнасят за изследвания с различен модел на извадката. Andridge and Little (2010) посочват, че при импутиране на липсващите стойности чрез случайния метод хотдек подборът на донорите може да се извърши с вероятност пропорционална на теглата на донорите (претеглен случаен метод хотдек).

Претегленият случаен метод хотдек използва теглата на единиците w_i в A и B при изчисляването на оценките на процентните пунктове на емпиричната кумулативна функция на разпределение чрез следните изрази:

$$\hat{F}^{(A)}(x) = \frac{\sum_{i=1}^{n_A} w_i^{(A)} I(x_{A,i} \leq x)}{\sum_{i=1}^{n_A} w_i^{(A)}} \text{ и} \quad (1)$$

$$\hat{F}^{(B)}(x) = \frac{\sum_{i=1}^{n_B} w_i^{(B)} I(x_{B,i} \leq x)}{\sum_{i=1}^{n_B} w_i^{(B)}} \quad (2)$$

По този начин се използват наличните тегла както от набора от данни (B), определен като донор, така и тези от набора от данни, определен като получател (A).

D'Orazio et al. (2012) сравняват резултатите от няколко елементарни микро-подхода за статистическо съчетаване на данни. Като цяло, когато се използват тегла при случайния и ранговия метод хотдек, тези методи се представят добре по отношение на запазване в синтетичния набор от данни на пределното разпределение на Z и на съвместното разпределение на $X \times Z$. Методът хотдек на разстоянието дава добри резултати само когато се постави ограничение при съчетаването на данните (т.е. един донор може да бъде използван само веднъж) и дадена дизайн променлива (например променлива, използвана за стратификация) се вземе предвид при формиране на донорски класове.

Сложни подходи - методи, които вземат под внимание теглото от извадката

В литературата има няколко метода за статистическо съчетаване на данни, които вземат под внимание теглото от извадката: подходът на Ренсен (Renssen, 1998), основан на калибрирането на теглата; обединяването на файлове, предложено от Рубин (Rubin, 1986), и подходът, основан на емпиричната вероятност (empirical likelihood), предложен от By (Wu, 2004). Сравнение между тези подходи може да се намери в D'Orazio et al. (2010).

Подходът на Рубин (обединяване на файлове) се състои в определяне на вероятностите за попадане, които единиците в извадка A биха имали, ако се използва моделът на извадка B ($\pi_a^B, a = 1, \dots, n_A$) и вероятностите за попадане, които единиците в извадка B биха имали, ако се използва моделът на извадка A ($\pi_b^A, b = 1, \dots, n_B$). Следователно полученият файл, съчетаващ двата набора от данни, ще има $n_A + n_B$ единици с вероятност за попадане:

$$\pi_h^{A\cup B} = \pi_h^A + \pi_h^B - \pi_h^{A\cap B}, \quad (3)$$

където $h = 1, \dots, n_A + n_B$, а последният член показва вероятността за попадане на единица в пресечната точка между двете извадки. В повечето случаи тази последна вероятност е незначителна, тъй като шансът дадена единица да попадне в две независими извадки, излъчени от една и съща генерална съвкупност, е близка до нула (обикновено извадките са сравнително малки в сравнение с размера на генералната съвкупност) и както е предложено от Rubin (1986), може да бъде елиминирана от формулата. За съжаление, подобен подход е подходящ само при теоретични извадки и не отчита намаляването на размера на извадката поради липсата на отговор от определен брой единици. Освен това би било доста трудно да се изчислят вероятностите за попадане $\pi_h^{A\cup B}$, защото за тези изчисления е необходимо: (1) познаване на моделите на извадките, използвани за избор на единиците съответно от A и B; (2) дизайн променливите, използвани за избор на единиците в A, да бъдат налични в A и B и (3) дизайн променливите, използвани за избор на B, да бъдат налични в A и B (Ballin et al., 2008). След като се изчислят вероятностите, остава проблемът да се оценят необходимите параметри при наличието на липсващи стойности, тъй като Z липсва в A, а Y липсва при B.

Подходът на Ренсен (калибриране) се основава на поредица от стъпки за калибриране на теглата в A и B. Всички етапи на обработка се изпълняват поотделно в двата източника на данни, като крайната цел е да се оцени даден параметър, отнасящ се до връзката между Y и Z. Калибрирането е често използвана техника при извадковите изследвания, при която се изчисляват нови тегла, възможно най-близки до първоначалните (дизайн или базови тегла), които изпълняват поредица от ограничителни условия относно сумите на набор от помощни променливи (Sarndal and Lundstrom, 2005). По-конкретно, ако w_k са „началните“ тегла, то крайните калибрирани тегла w_k^{cal} се извеждат като решение на задача за минимизиране:

$$min\left[\sum_{k\in r} D\left(w_k, w_k^{cal}\right)\right] \quad (4)$$

при спазване на следните ограничения (ако приемем, че има само една помощна променлива):

$$\sum_{k=1}^{m} w_c^{cal} x_k = \sum_{k=1}^{N} x_k \text{ и} \quad (5)$$

$$\sum_{k=1}^{m} w_k^{cal} = N, \quad (6)$$

където $D\left(w_k, w_k^{cal}\right)$ е мярка за разстояние. Подходът на Ренсен работи добре при използване на категорийни променливи или в случай на категорийни и много ограничен брой количествени променливи (D'Orazio et al., 2010).

На практика подходът, използван от By (емпирична вероятност), е подобен на калибрирането, като позволява да се изчислят нови тегла за единиците в A и за единиците в B, които отговарят на някои ограничения относно сумите на общите променливи X. Тези целеви суми могат да бъдат: (1) известни от трети източници; (2) оценени чрез комбиниране на оценки, получени по отделно от A и B (този подход е сходен на подхода на Ренсен), и (3) неизвестни и да не са оценени, а да бъдат зададени като равни, като допълнително ограничение в задачата за оптимизация (комбиниран подход). Като цяло подходът на By е по-гъвкав в сравнение с калибрирането (не се получават отрицателни тегла) и при комбинирания подход не е необходимо да се прави оценка на сумите на X променливите. От друга страна, прилагането му при сложни извадкови модели е трудно и не се взема под внимание намаленият размер на извадката поради наличието на липсващи отговори от страна на единиците (D'Orazio et al., 2010).

Непараметрични микрометоди

Обичайна практика е прилагането на определен набор от непараметрични подходи за импутиране, обикновено наричани „хотдек процедури за импутиране". Тези процедури се характеризират с факта, че запълват липсващите стойности с наблюдавани (истински) такива, поради което са широко използвани при статистическото съчетаване на данни от различни източници. Друг плюс на хотдек процедурите за импутиране е, че не се нуждаят от спецификация на дадено семейство разпределения (т.е. те са непараметрични) и не се нуждаят от оценка на функцията за разпределение или на някоя от нейните характеристики.

Като цяло методите, които се използват при статистическото съчетаване на данните от сложни извадкови изследвания, са непараметрични в своята същност (D'Orazio, M., M. Di Zio, M. Scanu, 2012). По тази причина са представени трите най-разпространени в практиката непараметрични метода - случаен метод хотдек (random hot deck), рангов метод хотдек (rank hot deck) и метод хотдек на разстоянието (distance hot deck).

При използването на методи хотдек за импутиране на данни в рамките на статистическото съчетаване на данни е необходимо да се припишат определени роли на

наборите от данни (A и B), които ще бъдат използвани. Единият набор от данни приема ролята на файла получател, като липсващите елементи от всеки запис в него се импутират (попълват) с помощта на подходящо избрани записи от другия набор от данни, донорския файл.

Изборът на кой от двата файла да бъде приписана ролята на получател и на кой ролята на донор зависи от много фактори. Най-важните са същността на изследваните явления и точността на информацията, съдържаща се в двата файла (D'Orazio et al., 2006).

Когато данните от двата източника могат да бъдат считани за еднакво надеждни (т.е. получени от представителни извадки на една и съща генерална съвкупност), при присвояването на подходящата роля на получател и на донор на съответните файлове се взема предвид и размерът на извадката на всеки от тях. Например ако броят на записите в двата файла се различава значително, обичайната практика е да се избере по-малкият файл като получател. Ако по-малкият файл се избере за донор, някои записи в донорския файл ще бъдат импутирани повече от веднъж в получаващия файл, като по този начин изкуствено се променя дисперсията на разпределението на импутираната променлива в крайния синтетичен файл.

Нека A е файлът получател, а B - донорът. Следователно целта е импутирането на целевата променлива Z във файла A чрез използването на наблюдаваните единици при B. Три метода хотдек за импутиране се използват при статистическото съчетаване на данни (Singh et al., 1993): случаен хотдек (random hot deck), рангов хотдек (rank hot deck) и хотдек на разстоянието (distance hot deck).

Случаен метод хотдек (Random hot deck)

Случайният метод хотдек се състои в произволен избор на запис от донорския файл за всеки запис във файла получател. Понякога случайният избор се прави в рамките на подходяща подгрупа единици в донорския файл. По-точно, единиците от двата файла се групират в хомогенни подмножества според дадени общи характеристики (единици, обитаващи един и същ географски район, лица с еднакви демографски характеристики и др.). Тези подмножества се наричат донорски класове. По този начин за дадено лице в даден географски район ще се считат за възможни донори само записи в същия географски район. Като цяло донорските класове се дефинират с помощта на една или повече категорийни променливи X, избрани в рамките на набора от общи променливи между A и B.

Нека да разгледаме следния елементарен пример. Нека наборът от данни A да съдържа 6 записа и три променливи: възраст, пол и нетен годишен доход на лицето (табл. 2). Нека наборът от данни B да съдържа 10 записа и три променливи: възраст, пол и регулярно използване на интернет от лицето. Следователно имаме набор от две общи променливи X ($X_1 = \text{възраст}$ и $X_2 = \text{пол}$) и две променливи, които не се наблюдават съвместно: Y - нетен годишен доход и Z - регулярно използване на интернет.

Нека A е файлът получател, а B - донорският файл. Ако дадена единица b е приписана на единица a, то липсващата стойност на Z за a ще бъде импутирана с наблюдаваната стойност на Z в b. A-тият запис в крайния синтетичен файл ще бъде (x_a, y_a, z_b). В случая един запис от B може да бъде приписан повече от един път на различни записи в A, т.е. играе ролята на донор повече от веднъж. Теоретично има $n_B^{n_A} = 10^6$ възможни подмножества записи, които могат да бъдат избрани като донори от B, и следователно 10^6 възможни стойности за регулярното използване на интернет.

Ако общата променлива „пол“ се използва за формирането на донорски класове, донорите от B трябва да бъдат избрани сред тези, които се намират в същия донорски клас по отношение на пола. Таблица 2 дава пример за разпределението на A и B по пол. По този начин броят на възможните донорски конфигурации намалява значително: $(n^B_{\text{мъж}})^{n^A_{\text{мъж}}} + (n^B_{\text{жена}})^{n^A_{\text{жена}}} = 4^4 + 6^2 = 292.$

1. Примерни данни за променливата „пол“ в A и B

a	Пол в A		B	Пол в B
1	Жена		1	Жена
2	Мъж		2	Мъж
3	Мъж		3	Жена
4	Жена		4	Мъж
5	Мъж		5	Жена
6	Мъж		6	Жена
-			7	Мъж
-			8	Мъж
-			9	Жена
-			10	Жена

Прогнозирането чрез случаен метод хотдек в рамките на донорски класове, дефинирани чрез X (за които се предполага, че са категорийни променливи), е еквивалентно на оценка на условното разпределение на Z при дадено X в B и подбор на наблюдение от него. Когато Z е количествена променлива, разпределението на Z при дадено X $f(Z, X)$ се оценява чрез емпиричната кумулативна функция на разпределението като оценка на съвместната кумулативна функция на разпределение:

$$\hat{F}_{Z/X(Z|X)} = \frac{\sum_{b=1}^{n_B} I(z_b \leq z) I(x_b = x)}{\sum_{b=1}^{n_B} I(x_b = x)} \ . \quad (7)$$

В този случай случайното изтегляне на наблюдение от донорския файл в клас *x* е еквивалентно на изтегляне на стойност от $\hat{F}_{Z/X}$. Същото важи и когато Z е категорийна променлива. Вместо функцията за кумулативно разпределение трябва да се вземат предвид оценките $\hat{\theta}_{k/i}$, $k = 1, \dots, K$. Методът на случайния хотдек съвпада с произволно изтегляне от това прогнозно разпределение.

Когато случайният метод хотдек се извършва без донорски класове, се приема, че Z и X са независими, като оценка на това предположение е възможна, използвайки данните в B. В този случай за генериране на стойностите, които ще бъдат импутирани, вместо оценките $\hat{F}_{Z/X}$ или $\hat{\theta}_{k/i}$ се използва пределното емпирично разпределение на Z в B.

Рангов метод хотдек (Rank hot deck)

При наличие на обща променлива X, измерена на ординалната скала, тя може да се използва при избора на донори от B за импутиране на записите относно променливата Z в A. В тази ситуация е възможно да се използва последователността във връзката между стойностите на X, което на практика представлява ранговия метод хотдек (Singh et al., 1993).

Единиците в двата файла са ранжирани поотделно спрямо стойностите на X. Ако приемем, че A е файлът получател и $n_B = kn_A$, където k е цяло число, то файловете могат да бъдат съчетани чрез свързване на записи с еднакъв ранг. Когато файловете съдържат различен брой записи, съчетаването се извършва като се вземе предвид кумулативната функция на разпределението на X във файла получател:

$$\hat{F}_X^A(x) = \frac{1}{n_A} \sum_{a=1}^{n_A} I(x_a \leq x), x \in X, \quad (8)$$

и в донорския файл:

$$\hat{F}_X^B(x) = \frac{1}{n_B} \sum_{b=1}^{n_B} I(x_b \leq x), x \in X. \quad (9)$$

След това всяко $a = 1, \dots, n_A$ се свързва с този запис b в B, за който:

$$\left| \hat{F}_X^A(x_a^A) - \hat{F}_X^B(x_b^B) \right| = \min_{1 \leq b \leq n_B} \left| \hat{F}_X^A(x_a^A) - \hat{F}_X^B(x_b^B) \right|. \quad (10)$$

Нека използваме следния пример за онагледяване на използването на ранговия метод хотдек за статистическо съчетаване на данните от два файла (A - получател, и B - донор). Таблици 3 и 4 съответно показват данните, налични в A и B, сортирани (ранжирани) по променливата „възраст“, която ще бъде използвана като съчетаваща променлива. Таблица 5 показва за всяко $a \in A$ най-близкия запис на B по отношение на неговия ранг. Крайният съчетан файл е показан в табл. 6.

2. Файл A с ранжирани по възраст записи

a	X_1 (Пол)	X_2 (Възраст)	Y (Нетен годишен доход)
1	Жена	26	11400
5	Мъж	29	11760
3	Мъж	32	19200
6	Мъж	43	21120
4	Жена	45	23400
2	Мъж	57	27600

3. Файл B с ранжирани по възраст записи

b	X_1 (Пол)	X_2 (Възраст)	Z (Регулярно използване на интернет)
8	Мъж	23	2
1	Жена	24	1
5	Жена	29	2
4	Мъж	32	1
2	Мъж	37	1
10	Жена	39	1
6	Жена	42	1
9	Жена	46	1
7	Мъж	48	1
3	Жена	54	2

4. Най-близкият запис на B за всеки запис в A според ранговете

a	$\hat{F}_X^A(x_a^A)$	$\hat{F}_X^B(x_{b^*}^B)$
1	1/6	$\hat{F}_X^B(x_1^B) = 2/10$
5	2/6	$\hat{F}_X^B(x_5^B) = 3/10$
3	3/6	$\hat{F}_X^B(x_2^B) = 5/10$
6	4/6	$\hat{F}_X^B(x_6^B) = 7/10$
4	5/6	$\hat{F}_X^B(x_9^B) = 8/10$
2	6/6	$\hat{F}_X^B(x_3^B) = 10/10$

5. Съчетан файл по ранговия метод хотдек

a	X_1^A	X_2^A	Y	Донор b	Z
1	Жена	26	11400	1	1
5	Мъж	29	11760	5	2
3	Мъж	32	19200	2	1
6	Мъж	43	21120	6	1
4	Жена	45	23400	9	1
2	Мъж	57	27600	3	2

Метод хотдек на разстоянието (Distance hot deck)

При метода хотдек на разстоянието всеки запис във файла получател се съпоставя с най-близкия запис в донорския файл, в съответствие с дадена мярка за разстояние, изчислена с помощта на общите променливи X. В най-простия случай на единична непрекъсната променлива X гореизложеното означава, че донорът за a-тия запис във файла получател A трябва да бъде избран, така че:

$$d_{ab^*} = \left|x_a^A - x_{b^*}^B\right| = \min_{1\le b\le n_B} |x_a^A - x_b^B|. \quad (11)$$

Като цяло, когато два или повече донорски записа са еднакво отдалечени от записа във файла получател, един от тях се избира на случаен принцип.

Този подход на метода хотдек на разстоянието се нарича неограничен, тъй като всеки запис в донорския файл B може да се използва като донор повече от веднъж.

Друг метод хотдек на разстоянието е ограниченият. Този подход позволява всеки запис в B да бъде избран като донор само веднъж. Ограниченият метод хотдек на разстоянието изисква броят на донорите да е по-голям или равен на броя на получателите $(n_A \le n_B)$. В най-простия случай на равен брой единици в двата файла $(n_A = n_B)$ моделът на донора трябва да бъде такъв, че функцията

$$\sum_{a=1}^{n_A}\sum_{b=1}^{n_B} d_{ab}w_{ab} \quad (12)$$

да е минимизирана при следните ограничения:

$$\sum_{b=1}^{n_B} w_{ab} = 1, a = 1, \dots, n_A \text{ и} \quad (13)$$

$$\sum_{a=1}^{n_A} w_{ab} = 1, b = 1, \dots, n_B, \quad (14)$$

където $w_{ab} \in \{0; 1\}$ и $w_{ab} = 1$, ако двойката (a, b) е съвпаднала, и $w_{ab} = 0$, ако не е (Kadane, 1978).

Проблемът с минимизирането на горепосочената функция става малко по-сложен, когато има повече донори, отколкото получатели $(n_A < n_B)$. В този случай ограниченията придобиват вида:

$$\sum_{b=1}^{n_B} w_{ab} = 1, a = 1, \dots, n_A \text{ и} \quad (15)$$

$$\sum_{a=1}^{n_A} w_{ab} \le 1, b = 1, \dots, n_B, \quad (16)$$

където $w_{ab} \in \{0; 1\}$, като е необходимо $\sum_{a=1}^{n_A} \sum_{b=1}^{n_B} w_{ab} = n_A$.

Минимизирането на съвкупното (агрегираното) разстояние след съчетаването на данните между двата файла съответства на търсенето на най-доброто съвпадение с поставено ограничение.

Основното предимство на метода хотдек на разстоянието с ограничение за съчетаване на данни в сравнение с този без ограничение е, че пределното разпределение на импутираната променлива Z се запазва в крайния синтетичен файл. Недостатък е, че средното разстояние на стойностите на донора и получателя на общите променливи X не се очаква да бъде по-голямо от това в неограничения подход (което е причина за отклонението от съчетаването при метода хотдек на разстоянието).

При хот дек метода на разстоянието също могат да се използват донорски класове, като те са особено полезни, когато има голям брой общи променливи, които правят изчисленията тромави. Ако дефинираме например променливата „пол“ като донорски клас, разстоянията, изчислени на базата на „възраст“, трябва да бъдат ограничени само до тези единици в двата файла с един и същи пол.

Едно от най-популярните разстояния, прилагани при метода хотдек на разстоянията за статистическо съчетаване на данни, е разстоянието на Маханалобис. То се дефинира като:

$$d_{ab} = (x_a - x_b)' \sum_{XX}^{-1} (x_a - x_b), \qquad (17)$$

където Σ_{XX} е ковариационната матрица на X. Когато Σ_{XX} е неизвестна, е възможно да се използва една от нейните оценки. Това разстояние отчита статистическата връзка между X променливите.

IV. Класификация на методите за статистическо съчетаване на данни от сложни извадкови изследвания

В статистическата литература няма ясно структурирана класификация на използваните методи за статистическото съчетаване. Според различни критерии на методите, използвани за съчетаване на данни от сложни извадкови изследвания, може да се направи класификация като тази в табл. 7.

6. Примерна класификация на методите за статистическо съчетаване на данни от сложни извадкови изследвания

Класификационен критерий	**Метод**
Според вида на подхода	• Микро • Макро
Според допускането за съвместното разпределение на общите и целевите променливи	• Параметрични • Непараметрични • Смесени
Според вземането под внимание на теглата от извадката	• Елементарни • Сложни - подход на Рубин, подход на Ренсен и подход на Ву
Според начина на избор на определен запис като донор	• Случаен хотдек • Рангов хотдек • Хотдек метод на разстоянието до най-близкия съсед
Според използването на групиращи променливи	• С донорски класове • Без донорски класове
Според поставянето на ограничаващи условия	• Неограничен • Ограничен

Предложената класификация не е изчерпателна по отношение на всички съществуващи методи за статистическо съчетаване на данни от различни източници, но е добра основа за разработване на по-детайлна структура на подобна класификация.

V. Заключение

В последните години статистическото съчетаване на данни става все по-актуално, тъй като потребителите на статистическа информация изискват по-подробна, по-точна и по-навременна информация за социално-икономическите събития. Наред със стремежа за намаляване на натовареността на респондентите, намаляване на разходите за провеждане на изследвания, все по-бързото развитие на информационните технологии и увеличаване на възможностите за обработка на голямо количество информация, ползите от интегрирани набори от данни стават все по-привлекателни по отношение на подобрени статистически изследвания, с помощта на които да се произвежда необходимата информация за вземане на политически решения.

Макар широко обсъждана и изследвана в световен мащаб, темата за статистическото съчетаване на данни от извадкови изследвания не е достатъчно разработвана в България както на теоретично, така и на практическо ниво. Националният статистически институт (НСИ) провежда множество извадкови и изчерпателни изследвания в различни области на статистиката - демографска и социална,

макроикономическа, бизнес, многоотраслова и други. Възможностите за съчетаване на данните от тях изглеждат на пръв поглед неизчерпаеми, но е необходимо да се обърне внимание на следните условия и ограничения:

1. Използваните статистически единици трябва да бъдат идентични или сходни - могат да бъдат съчетавани домакинства с домакинства, лица на определена възраст с лица на същата възраст, предприятия с предприятия и други.

2. Референтните периоди, за които се отнасят данните от изследванията, обект на статистическо съчетаване, трябва да бъдат еднакви или да бъде възможно да се приеме за вярно предположението, че генериращият модел не се променя от първото към второто изследване.

3. Когато се съчетават данни от извадкови изследвания, е препоръчително моделите на извадките да са сходни и да се отнасят до една и съща генерална съвкупност.

4. Данните от изследванията, които ще бъдат обект на статистическо съчетаване, трябва да бъдат хомогенни, т.е. използваните дефиниции на променливите, които ще бъдат използвани за съчетаването, да бъдат еднакви или достатъчно близки или да има възможност чрез подходящо трансформиране на първичните променливи те да бъдат хармонизирани до степен, подходяща за използването им в процеса на статистическото съчетаване на данни.

За съчетаване на данни от различни изследвания в домакинствата, провеждани от НСИ, особено подходящи са тези изследвания, които са основани на Регламент (ЕС) 2019/1700 на Европейския парламент и на Съвета от 10 октомври 2019 г. за създаване на обща рамка за европейската статистика за лицата и домакинствата, основана на индивидуални данни, събрани чрез извадки[2]. В европейския документ са заложени някои общи дефиниции и концепции за всички изследвания, включително и т.нар. целеви променливи (core variables) - въпроси, които се задават по един и същ начин, с еднаква номенклатура и по единна методология, което е едно от основните условия за приложение на статистическо съчетаване на данни от различни източници. В тези

[2] Регламент (ЕС) 2019/1700 на Европейския парламент и на Съвета от 10 октомври 2019 г. за създаване на обща рамка за европейската статистика за лицата и домакинствата, основана на индивидуални данни, събрани чрез извадки, за изменение на Регламенти (ЕО) № 808/2004, (ЕО) № 452/2008 и (ЕО) № 1338/2008 на Европейския парламент и на Съвета и за отмяна на Регламент (ЕО) № 1177/2003 на Европейския парламент и на Съвета и на Регламент (ЕО) № 577/98 на Съвета (текст от значение за ЕИП), PE/63/2019/REV/1.

изследвания се използват и еднакви статистически съвкупности (всички лица с обичайно местопребиваване в обикновени домакинства във всяка държава членка), както и сходни концепции за претегляне на данните и съответно възпроизвеждане на еднакъв брой на населението за съответна референта дата.

В рамковия регламент са включени седем изследвания, всяко едно от които предлага голям набор от данни и възможности за анализ. Комбинирането на данни чрез изследвания метод обаче разширява тези възможности до много по-високи нива, тъй като на практика се създава по-широк набор от данни. Например при свързване на данните за характеристиките на основната работа от наблюдението на работната сила и информацията за видовете неформално обучение, в което лицата участват с източник изследването на образованието и обучението на възрастни, ще бъде възможно да се анализира нуждата от специфични умения и компетентности, необходими на пазара на труда, а оттам и да се повиши конкретно практическата насоченост на формалната образователна система чрез актуализиране на съществуващите образователни програми, базирани на информирани решения.

Друг пример, свързан с възможностите за промяна на политиките на държавата, е при свързване на данни за непосрещнатите нужди от здравни грижи по финансови причини с източник „Европейско здравно интервю“, както и линията на бедност от изследването на доходите и условията на живот, с което би се установила необходимостта от промяна на финансирането от държавата при отделните медицински дейности и услуги с оглед осигуряване на адекватна здравна грижа за цялото население на страната.

ЦИТИРАНА ЛИТЕРАТУРА:

Andridge, R., R. Little (2010). A Review of Hot Deck Imputation for Survey Nonresponse, *International Statistical Review*, 78, 40-64.

Ballin, M., M. Di Zio, M. D'Orazio, M. Scanu, N. Torelli (2008). File Concatenation of Survey Data: a Computer Intensive Approach to Sampling Weights Estimation, Rivista di Statistica Ufficiale.

D'Orazio, M., M. Di Zio, M. Scanu (2006). *Statistical Matching: Theory and Practice*, Wiley.

D'Orazio, M., M. Di Zio, M. Scanu (2006). Statistical matching for categorical data: Displaying uncertainty and using logical constrains, *Journal of Official Statistics*, 22, 137-157.

D'Orazio, M., M. Di Zio, M. Scanu (2010). Old and new approaches in statistical matching when samples are drawn with complex survey designs, *Proceedings of the 45th "Riunione Scientifica della Societa' Italiana di Statistica"*, Padova 16-18 June 2010.

D'Orazio, M., M. Di Zio, M. Scanu (2012). Statistical Matching of Data from Complex Sample Surveys, *Proceedings of the European Conference on Quality in Official Statistics - Q2012*, 29 May - 1 June 2012, Athens, Greece.

Renssen, R. (1998). Use of statistical matching techniques in calibration estimation, *Survey Methodology*, 24, 171-183.

Rubin, D. (1986). Statistical matching using file concatenation with adjusted weights and multiple impitations, *Journal of Business and Economic Statistics*, 4, 87-94.

Sarndal, C., S. Lundstrom (2005). *Estimation in Surveys with Nonresponse*, Wiley.

Singh, A., H. Mantel, M. Kinack, G. Rowe (1993). Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption, *Survey Methodology*, 19, 59-79.

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method, *The Canadian Journal of Statistics*, 32, 15-26.