УНИВЕРСИТЕТ ЗА НАЦИОНАЛНО И СВЕТОВНО СТОПАНСТВО

ФАКУЛТЕТ „ПРИЛОЖНА ИНФОРМАТИКА И СТАТИСТИКА“

КАТЕДРА „СТАТИСТИКА И ИКОНОМЕТРИЯ“

Цвета Цонкова Цонкова

СТАТИСТИЧЕСКО СЪЧЕТАВАНЕ НА ДАННИ ОТ ИЗВАДКОВИ ИЗСЛЕДВАНИЯ ЗА АНАЛИЗА В СОЦИАЛНО-ИКОНОМИЧЕСКАТА ОБЛАСТ

АВТОРЕФЕРАТ

на дисертационен труд за присъждане
на образователна и научна степен „доктор“
по научна специалност „Статистика и демография“

Научен ръководител:

Доц. д-р Соня Дакова Чипева

София, 2022 г.

Дисертационният труд се състои от увод, три глави, заключение, списък на използваната литература и приложения. Той е в обем на 188 страници, от които 120 страници изложение, 5 страници списък на използваната литература и 63 страници приложения. Изложението съдържа 10 фигури и 30 таблици. Списъкът на използваната литература включва 56 заглавия на английски език и 3 интернет страници.

Дисертационният труд е обсъден и насочен за защита от катедра „Статистика и иконометрия“ при факултет „Приложна информатика и статистика“ на УНСС – София.

Авторът на дисертационния труд е бил задочен докторант към катедра „Статистика и иконометрия“ при УНСС, София, и работи като държавен експерт в отдел „Статистика на условията на живот“ в НСИ.

Защитата на дисертационния труд ще се състои на 25.05.2022 г. от 11:00 часа в зала 2032-А на УНСС на открито заседание на научно жури.

Материалите по защитата са на разположение на интересуващите се в библиотеката и в сектор „Научни съвети“ на УНСС – София.

I. ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

1. Актуалност на темата

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. По-конкретно рубриката на Европейската комисия „Отвъд БВП“, както и докладът на Комисията относно измерването на икономическите резултати и социалния напредък (Stiglitz, Sen, Fitoussi[1]) наблягат 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. От тук произтича и необходимостта от наличието на интегрирана статистическа информация, която да обхваща различни аспекти на социално-икономическото развитие на обществото.

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 тези аспекти. В този контекст, настоящия 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

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

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 ефективността на използване на информацията чрез прилагане на настоящите и утвърдени начини за събиране на данни.

Проблематиката за статистическото съчетаване на данни в световен мащаб може да бъде проследена от средата на 60^{те} години на миналия век, когато в САЩ е създаден изчерпателен набор от данни с информация за социално-демографските характеристики, доходите и приходите (събрани въз основа на подадените данъчни декларации) на семействата, чрез съчетаване на данни от два източника: така нареченото данъчно досие (Tax file) от 1966 г. и изследването на икономическите възможности от 1967 година. Процедурите за съчетаване на информация от отделни извадкови изследвания продължава да бъде интересна и обсъждана тема. В последните години статистическото съчетаване на данни става все по-актуално като важна причина за това е все по-задълбоченото изследване на така наречените „големи данни“. Друга важна причина, поради която статистическото съчетаване на данни става все по-интересно за изследователите като цяло и за националните статистически институти в частност, е, че потребителите на статистическа информация изискват по-подробна, по-точна и по-навременна информация за социално-икономическите събития.

Макар широко обсъждана и изследвана в световен мащаб, темата за статистическото съчетаване на данни от извадкови изследвания, според проучената литература по темата, не е подробно разработвана в България. Резултатите от проведеното изследване в рамките на дисертационния труд биха могли да допринесат за разработването на модели за съчетаване на статистическа информация от различни източници, включително данни от административни източници или т.нар. „големи данни“, с което да се постигне изграждането на по-пълна картина на социално-икономическата действителност в България.

Според направения SWOT анализ в рамките на стратегията за развитие на Националната статистическа система на Република България 2013-2017 г., изменена

чрез удължаване до 2020 г.[2], приложението на интегриран подход и съчетаване на информация от различни източници е недостатъчно. За справяне с този проблем едно от изброените основни предизвикателства, които ще определят развитието на Националния статистически институт (НСИ) и органите на статистиката в България според гореспоменатата стратегия, е съчетаването на различни източници на данни при производството на статистическа информация: административни данни, големи данни, данни от извадкови статистически изследвания. В този контекст, наличието на изследване, което да обобщава различни методи за съчетаване на статистически данни от извадкови изследвания и който да предлага конкретен модел за съчетаване на данни от провеждани от НСИ извадкови изследвания, би бил полезен в постигането на целите на българската статистика в практиката.

2. Изследователска теза, цели, задачи и хипотези на изследването

Изследователската теза на дисертационния труд е, че съществува обективна необходимост от разработване на технология за статистическо съчетаване на данни от различни източници за задълбочаване на анализа в социално-икономическата сфера в България.

Целта на разработката е да се оцени възможността и да се разработи технология за прилагане на методи за статистическо съчетаване на данни от две различни извадкови изследвания и да се апробира разработената технология за съчетаване на данни от извадковите изследвания „Използване на информационни и комуникационни технологии (ИКТ) от домакинствата“ и „Изследване на доходите и условията на живот“ (EU-SILC), за подобряване на възможностите за анализ чрез получаване на синтетичен набор от данни, който съдържа информация както за регулярното използване на интернет от лицата на възраст 16-74 години в домакинствата, така и за техния еквивалентен нетен разполагаем доход.

За реализирането на поставената цел са дефинира следните **задачи**:

- Да се изясни същността на понятието „статистическо съчетаване на данни“.
- Да се анализират и опишат съществуващи подходи и методи за статистическо съчетаване на данни от различни източници.

[2] https://www.nsi.bg/sites/default/files/files/pages/uplf/Strategy2013-2017_2020.pdf

- Да се проучи опита за съчетаване на данни от различни източници в световен мащаб.
- Да се дефинират възможности за приложение на статистическото съчетаване на данни от различни източници в българската статистическа практика.
- Да се разработи методология за статистическо съчетаване на данни от две извадкови изследвания с акцент върху различните подходи за хармонизиране и обединяване на масивите от данни.
- Да се апробира разработената методология, за да се съчетаят статистически данни от две сложни извадкови изследвания, провеждани ежегодно от Националния Статистически Институт – изследването „Използване на ИКТ от домакинствата“ и изследването на доходите и условията на живот (EU-SILC), като се акцентира върху качеството и ефективността на използваните методи.
- Да се направи описателен анализ на данните от получения интегриран масив от данни.
- Да се анализират резултатите от експеримента и да се формулират изводи и препоръки.

В рамките на изследването са формулирани следните **изследователски хипотези:**

- възможно е в българската статистическа практика да бъдат приложени методи за статистическо съчетаване на данни от две извадкови изследвания за получаване на синтетичен набор от данни;
- полученият синтетичен набор от данни е предпоставка за подобряване на анализа в социално-икономическата сфера.

3. Обект, предмет и обхват на изследването

Обект на изследването са данните от две извадкови изследвания, провеждани ежегодно от Националния статистически институт – използване на информационни комуникационни технологии (ИКТ) от домакинствата и изследването на доходите и условията на живот (EU-SILC).

Предмет на изследването е статистическото съчетаване на данни от две регулярни извадкови изследвания.

Изследването обхваща обикновени домакинства и техните членове на 16-74 години, живеещи на територията на страната в момента на наблюдение през 2020 година. Лицата, живеещи в колективни домакинства и в институции са изключени като цяло от целевата съвкупност.

Настоящият дисертационен труд не е изчерпателно изследване в областта на статистическото съчетаване на данни от различни източници. Той има определени ограничения в своя обхват, които се изразяват в това, че не представя изчерпателно всички съществуващи методи за статистическо съчетаване на данни. Изследването е ограничено до прилагане на методи за статистическо съчетаване на данни от две сложни извадкови изследвания.

4. Методология на изследването

Във връзка с постигането на основната цел на дисертационния труд, а именно да се оцени възможността и да се разработи технология за прилагане на методи за статистическо съчетаване на данни от сложни извадкови изследвания в България е предложен специфичен подход за съчетаване на данните от две извадкови изследвания, провеждани от НСИ, на основата на анализ и синтез на съществуващи в теорията и статистическата практика подходи и методи за статистическо съчетаване на данни от сложни извадкови изследвания. Предложената технология, приложена в емпиричната част на настоящия дисертационен труд обхваща непараметрични методи за статистическо съчетаване на данни, по-конкретно случаен хот дек метод, рангов хот дек метод и хот дек метод на разстоянието до най-близкия съсед, които са структурирани в методика за получаване на обединен масив от статистически данни.

За провеждане на изследването са използвани индивидуални данни от изследванията „Статистика на условията на живот“ (EU-SILC) и „Използване на информационни и комуникационни технологии от домакинствата“ (ИКТ), провеждани от НСИ. Използвани са данни за 2020 година. Данните са обработени и анализирани чрез използване на софтуерните продукти IBM SPSS Statistics 26 и R Studio.

II. СТРУКТУРА И СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

Списък на използваните съкращения

УВОД

ГЛАВА ПЪРВА: СТАТИСТИЧЕСКО СЪЧЕТАВАНЕ НА ДАННИ ОТ РАЗЛИЧНИ ИЗТОЧНИЦИ

1.1. Същност на статистическото съчетаване на данни

1.1.1. Рамка на статистическото съчетаване на данни

1.1.2. Неидентифицируемо съвместно разпределение

1.1.3. Предположение за условна независимост

1.2. Статистическо съчетаване на данни от сложни извадкови изследвания

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

1.3.1. Елементарни мирко подходи

1.3.2. Методи, които взимат под внимание теглото от извадката

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

1.5. Историческо развитие на статистическото съчетаване на данни от различни източници

1.6. Нарастващото значение на статистическото съчетаване на данни от различни източници в днешно време

1.7. Възможности за статистическо съчетаване на данни от различни източници в българската статистическа практика

ГЛАВА ВТОРА: ТЕХНОЛОГИЯ ЗА РЕАЛИЗИРАНЕ НА СТАТИСТИЧЕСКО СЪЧЕТАВАНЕ НА ДАННИ ОТ ДВЕ ИЗВАДКОВИ ИЗСЛЕДВАНИЯ

2.1. Проблеми при подготовката на статистическото съчетаване на данни

2.2. Технология за успешно статистическо съчетаване на данни от различни източници

2.2.1. Съгласуване на концепциите и определенията, използвани в два източника на данни

2.2.2. Съгласуване на изместени източници на данни

2.2.3. Съгласуване на несъвместими разпределения

2.2.4. Избор на съчетаващите променливи

2.2.5. Оценка на точността/качеството на получените от статистическото съчетаване на данни резултати

ГЛАВА ТРЕТА: МЕТОДИКА ЗА СТАТИСТИЧЕСКО СЪЧЕТАВАНЕ НА ДАННИ ОТ ИЗСЛЕДВАНЕТО НА ДОХОДИТЕ И УСЛОВИЯТА НА ЖИВОТ (EU-SILC) И ИЗСЛЕДВАНЕТО „ИЗПОЛЗВАНЕ НА ИНФОРМАЦИОННИ И КОМУНИКАЦИОННИ ТЕХНОЛОГИИ ОТ ДОМАКИНСТВАТА“ (ИКТ)

3.1. Анализ на масивите от данни

3.2. Хармонизиране на масивите от данни

3.3. Рамка за статистическо съчетаване на данните от EU-SILC и ИКТ

3.4. Избор на съчетаващи променливи

3.5. Използвани методи за статистическо съчетаване на данните от EU-SILC и ИКТ

3.6. Описателен анализ на получения синтетичен файл

ЗАКЛЮЧЕНИЕ

СПИСЪК НА ИЗПОЛЗВАНАТА ЛИТЕРАТУРА

Приложение 1: Таблици с резултати от SPSS

Приложение 2: Използван код в R

Приложение 3: Извадка от получения синтетичен набор от данни

III. КРАТКО ИЗЛОЖЕНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

В уводната част на дисертационния труд е обоснована актуалността и значимостта на изследваната проблематика и са дефинирани изследователската теза, целта, задачите, обекта, предмета, обхвата и хипотезите на изследването.

ГЛАВА ПЪРВА: СТАТИСТИЧЕСКО СЪЧЕТАВАНЕ НА ДАННИ ОТ РАЗЛИЧНИ ИЗТОЧНИЦИ

Глава първа е посветена на въпроси, свързани с решаването на първите четири задачи.

Точка „Същност на статистическото съчетаване на данни“ е посветен на решаване на първата задача, свързана с изясняване на същността на понятието „статистическо съчетаване на данни“.

Акцентирано е върху необходимостта от използването на статистическото съчетаване на данни от различни източници като инструмент за повишаване ефективността на използване на наличната информация чрез прилагане на настоящите и

утвърдени в статистическата практика начини за събиране на данни. Тази необходимост произтича основно от ограниченията при планирането и изпълнението на нови изследвания (то е времеемко, изисква голям паричен ресурс, натоварва допълнително респондентите, а от там и намалява качеството на получената информация), които да са в състояние напълно да покрият нуждите от информация за управлението на икономиката и обществото.

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.

Като понятие за същността на статистическото съчетаване на данни от различни източници в рамките на настоящия дисертационен труд е възприета дефиницията на D'Orazio, според която **статистическото съчетаване на данни е метод за свързване на записите от два набора от данни в случаите, когато точното съвпадение на отделните записи (по уникален идентификационен номер на изследваните статистически единици) не е възможно поради ограниченията на наличните данни. Статистическото съчетаване на данни използва променливи, общи за двата набора от данни, за да идентифицира подобни по характеристики записи (статистически единици), с цел да се генерира нов синтетичен набор от данни, който да позволява по-обогатен и гъвкав анализ, отколкото би било възможно с двата набора от данни по отделно.**

В рамките на тази точка е представена и информация относно понятията „рамка на статистическото съчетаване на данни", „неидентифицируемо съвместно разпределение", „предположение за условна независимост", „допълнителна информация" и „интервал на несигурност", които са основополагащи при методите за статистическо съчетаване на данни.

Рамката на статистическото съчетаване на данни най-общо представя структурата на синтетичния набор от данни, към който се стремим като показва целевите и съчетаващите променливи. **Целевите променливи** са тези, които са налични по отделно в двата набора от данни (получателя и донора). **Съчетаващите променливи** са тези, на базата на които се осъществява статистическото съчетаване на данните от двата

файла (получател и донор), които са общи за двата набора от данни (например пол, възраст, най-висока степен на завършено образование и други).

Неидентифицируемо съвместно разпределение означава, че съвместното разпределение на всички променливи в синтетичния набор от данни (съчетаващите и целевите променливи), което следва да бъде оценено при статистическото съчетаване на данни, не може да бъде идентифицирано само на базата на пределните разпределения на целевите и общите променливи по отделно в двата набора от данни (получател и донор).

Предположението за условна независимост (ПУН) е най-често използваното предположение при статистическото съчетаване на данни за решаване на проблема с неидентифицируемото съвместно разпределение. ПУН приема, че целевите променливи са независими една от друга и условно независими от общите променливи. Проверката на валидността на ПУН в практиката е изключително трудна и поради тази причина се търсят начини за неговото заобикаляне.

Наличието на **допълнителна информация** за съвместното разпределение на целевите променливи намалява значително необходимостта от допускане на предположението за условна независимост.

Интервалът на несигурност показва определен интервал, в който се намира истинското съвместно разпределение на целевите променливи в синтетичния набор от данни като за неговото изчисляване се използват параметри, определящи пределните разпределения на целевите и съчетаващите променливи по отделно в двата набора от данни.

Точки „Статистическо съчетаване на данни от сложни извадкови изследвания“, „Методи за статистическо съчетаване на данни от сложни извадкови изследвания“ и „Класификация на методите за статистическо съчетаване на данни от сложни извадкови изследвания“ са посветени на въпроси, свързани с решаването на втората задача, отнасяща се до анализа на съществуващите подходи и методи за статистическо съчетаване на данни от различни източници, като е акцентирано върху използваните такива при сложни извадкови изследвания.

Подходите, използвани при статистическото съчетаване на данни са разграничени от D’Orazio като микро и макро, параметрични, непараметрични и смесени.

Микро подходите са подходи, при които целта на статистическото съчетаване на данни е създаването на синтетичен набор от данни.

Макро подходите са подходи, при които целта е да се получи оценка за дадени параметри на генералната съвкупност (например коефициент на корелация) относно връзката между целевите променливи.

Параметричните подходи са подходи, за които се прави изрично допускане за съвместното разпределение на общите и целевите променливи, т.е. се приема, че то се основава на фиксиран набор от параметри. Обикновено този тип подходи се използват, когато целта е оценка за дадени параметри на генералната съвкупност относно връзката между целевите променливи.

Непараметричните подходи са подходи, за които не е необходимо да се прави изрично допускане за съвместното разпределение на общите и целевите променливи.

Смесените подходи са подходи, при които методът за статистическо съчетаване на данни използва както параметричен, така и непараметричен модел в своята процедура.

В резултат на обстоен литературен преглед, в първата точка са изведени и обобщени предимствата и недостатъците на статистическото съчетаване на данни от сложни извадкови изследвания (таблица 1).

Таблица 1: Предимства и недостатъци на статистическото съчетаване на данни от сложни извадкови

Предимства	Недостатъци
• Подобрява аналитичния потенциал на данните чрез използването на по-широк обем от информация без допълнително увеличаване на разходите и натоварването на респондентите. • Може да спомогне за подобряване на съгласуваността (по отношение на използвани понятия и дефиниции, извадкова рамка и др.) между провежданите изследвания, ако те предварително бъдат планирани по	• Качеството на съчетаването силно зависи от качеството и обхвата на източниците на данни. • Необходимо е предварително общите между двете изследвания променливи да бъдат хармонизирани от гледна точка на използвани дефиниции и класификации. • Всяка допусната грешка в процеса на статистическото съчетаване на данни се пренася върху последващия анализ на данните от синтетичния файл.

Предимства	Недостатъци
начин, който да позволява прилагане на методи за статистическо съчетаване на данните между тях.	

За да бъдат минимизирани недостатъците е необходимо да се решат въпросите, свързани с: подходящ избор на целеви променливи; идентифициране, хармонизиране и избор на съчетаващи променливи; избор на статистическа рамка за статистическото съчетаване на данните и проверка на качеството на данните в получения синтетичен файл.

Точка „Методи за статистическо съчетаване на данни от сложни извадкови изследвания“ представя подходите, които най-често се използват при съчетаването на данни от сложни извадкови изследвания.

Елементареният микро подход се състои в прилагане на непараметрични методи (хот дек метод на разстоянието до най-близкия съсед, случаен или рангов хот дек метод) без да се взима под внимание модела на извадката или теглата в двата набора от данни. След като бъде получен синтетичният набор от данни, статистическите анализи се извършват като се вземе предвид модела на извадката, който стои в основата на набора от данни, избран като получател, и съответното тегло на единиците в него.

Сложните микро подходи взимат под внимание теглото от извадките на двете изследвания. Съществуват няколко такива подхода: подход на Рубин, подход на Ренсен и подход на Ву.

Подходът на Рубин се състои в определяне на вероятностите за попадане, които единиците в извадката на получаващия набор от данни биха имали, ако се използва моделът на извадка на набора от данни, определен за донор и обратно, вероятностите за попадане, които единиците в извадката на набора от данни, определен за донор биха имали, ако се използва моделът на извадка на получаващия набор от данни. По този начин се изчислява каква би била вероятността за попадане на дадена единица в един теоретичен синтетичен файл, чийто обем представлява сбора от единиците в получаващия файл и във файла донор. Този подход има редица недостатъци, описани в литературата, най-важния от които е, че не отчита намаляването на размера на извадката поради липсата на отговор от определен брой единици.

Подходът на Ренсен се основава на поредица от стъпки за калибриране на теглата в двата набора от данни. Всички етапи на обработка се изпълняват по отделно в двата източника на данни като крайната цел е да се оцени даден параметър, отнасящ се до връзката между целевите променливи. Калибрирането е често използвана при извадковите изследвания техника, 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 (обикновено се изчисляват по такъв начин, че да възпроизведат генералната съвкупност по избрани признаци).

Подходът на Ву е подобен на калибрирането, използвано от Ренсен като позволява да се изчислят нови тегла за единиците в двата набора от данни, които отговарят на някои ограничения относно сумите на общите им променливи. Прилагането му обаче при сложни извадкови модели е трудно и не взима под внимание намаления размер на извадката поради наличието на липсващи отговори от страна на единиците[3].

Като цяло методите, които се използват при статистическото съчетаване на данните от сложни извадкови изследвания са непараметрични в своята същност[4]. По тази причина са представени трите най-разпространени в практиката непараметрични метода – случаен хот дек метод (random hot deck), рангов хот дек метод (rank hot deck) и хот дек метод на разстоянието (distance hot deck).

Случайният хот дек метод 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.

Ранговия хот дек метод се състои в избор на запис от донорския файл във файла получател на базата на дадена съчетаваща променлива, измерена на ординалната скала,

[3] D'Orazio, M., Di Zio, M., Scanu, M. (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.

[4] D'Orazio, M., Di Zio, M., Scanu, M. (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.

като по този начин се използва последователността във връзката между стойностите на съчетаващите променливи.

Хот дек метода на разстоянието се състои в избор на запис от донорския файл във файла получател като 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 за двата набора от данни променливи. Когато два или повече донорски записа са еднакво отдалечени от записа във файла получател, един от тях се избира на случаен принцип.

Изброените методи могат да бъдат разграничени по още два признака – по отношение на оформянето на донорски класове и по отношение на прилагането на ограничение относно броя пъти, в който един донор може да бъде използван за попълване на запис във файла получател. Методите могат да бъдат **с донорски класове** или **без донорски класове**. Понятието „донорски клас“ означава, че единиците от двата набора с данни се групират в хомогенни подмножества според дадени общи характеристики – единици, обитаващи един и същ географски район, лица с еднакви демографски характеристики и други. При използването на подход с донорски класове за дадено лице в даден географски район например ще се считат за възможни за донори само записи в същия географски район. Ако всеки запис в донорския файл може да бъде използван като донор повече от веднъж, то избраният метод е **неограничен**. Ако избраният подход позволява всеки запис в донорския файл да бъде избран като донор само веднъж, то той е **ограничен**.

Точка „Класификация на методите за статистическо съчетаване на данни от сложни извадкови изследвания“ представя примерна класификация на използваните в дисертационния труд методи за статистическо съчетаване на данни от сложни извадкови изследвания според различни класификационни критерии (таблица 2). Тя е изведена на базата на подробно описаните в предходните две точки подходи.

Таблица 2: 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 взимането 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 хот дек • Рангов хот дек • Хот дек 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, а е съсредоточена върху изложените в настоящия дисертационен труд такива. Въпреки това считам, че е добра основа за разработване на по-детайлна структура на подобна класификация.

Точки „Историческо развитие на статистическото съчетаване на данни от различни източници“ и „Нарастващото значение на статистическото съчетаване на данни от различни източници в днешно време“ са посветени на решаването на третата задача, свързана с проучване на опита за съчетаване на данни от различни източници в световен мащаб.

Темата за статистическото съчетаване на данни от различни източници се обсъжда и развива от 60^{те} години на миналия век като приблизително по едно и също време тя се анализира и се правят опити за приложение на методи за статистическо съчетаване на данни в практиката в САЩ и Канада от една страна и в Европа, от друга. Направленията, в които се работи от двете страни на океана са различни от гледна точка на вида данни, които се обработват, но методите, които се използват са сходни. По онова време в САЩ и Канада статистическото съчетаване на данни от различни източници е насочено към постигане на приложим в политиката анализ на данъчни и трансферни програми, за получаване на информация за общественото здраве и благосъстояние. Целта е била

получаването на синтетичен набор от данни като са използвани основно методи, основани на функция за разстояние, изчислена на базата на коефициенти за съгласуваност между целевите променливи. За разлика от развитието на методите за статистическо съчетаване на данни от различни източници в САЩ и Канада, чиято основна цел е подобрение на микроикономическия анализ на федерално ниво, в Европа то се случва основно поради изискването за медийно планиране от частни медийни институции, които се нуждаят от възможно най-богата информация за планиране на своите телевизионни излъчвания и печатни издания, за да отговорят на нуждите и предпочитанията на своите потребители. По онова време в Европа са разработени два основни метода за съчетаване на данни от различни източници. Единият е базиран на така наречената „топологична концепция“, при която данните от различните източници се групират в хомогенни подмножества според дадени общи характеристики и след това за донори се избират записи от тези групи на основата на дадена функция на разстояние (на практика това са хот дек методи на най-близкия съсед чрез обособяване на донорски класове). Вторият метод е базиран на йерархия от „бракове“, където прехвърлената (донорската) информация може да идва от повече от един донор. Този метод се основава на хи-квадрат функция за разстояние между получателя и донора. Тази функция за разстояние е в основата на анализа на съответствието, който картографира единиците от получателя и донорския файл в многомерно пространство. След това се изчисляват нови мерки за разстояние въз основа на координатите на картографирания получател и донор. Накрая се дефинират различни видове „бракове“ и се използват оптимизационни алгоритми за постигане на крайните резултати от съчетаването.

Всички тези усилия са основополагащи за развитието на статистическото съчетаване на данни от различни източници, но са били силно ограничени за приложение в практиката поради факта, че изчисленията, необходими за реално минимизиране на сумата от разстоянията за всички потенциални съчетавания често надвишава изчислителната мощност, налична по това време.

В днешно време, при възможностите, които предоставят новите технологии, статистическото съчетаване на данни става все по-актуално като важна причина за това е все по-задълбоченото изследване на така наречените „големи данни“ (Big data). Друга важна причина, поради която статистическото съчетаване на данни става все по-интересно за изследователите като цяло и за националните статистически институти в частност е, че потребителите на статистическа информация изискват по-подробна, по-

точна и по-навременна информация за социално-икономическите събития. Наред със стремежа за намаляване натовареността на респондентите, намаляване на разходите за провеждане на изследвания, все по-бързото развитие на информационните технологии и увеличаване 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 за вземане на политически решения информация.

В дисертационния труд са дадени два примера за реализиране на статистическо съчетаване на данни от различни източници в съвременната статистическа практика. Тези примери не изчерпват направените по темата експерименти, но са показателен пример за възможностите, които дават методите за статистическо съчетаване на данни, за обогатяване на анализа в социално-икономическата сфера.

Първият изложен пример е съчетаването на данните от „Изследването за доходите и богатството на домакинствата“ и извадковото изследване за потреблението на домакинствата в Италия, което е осъществено, използвайки стратифициран хот дек метод на разстоянието. Целта на това съчетаване е получаването на обща информация относно разходите на домакинствата и техните приходи. По този начин е възможно да се оцени съставът на домакинствата в рамките на матрицата за отчитане на разходите за социална защита в страната. Това е матрица, в която домакинствата се групират според набор от различни характеристики, като район на пребиваване и първичен източник на доходи. Във въпросната матрица за характеристиките на тези домакинства се записват първо размерът на разходите (въз основа на подробен списък с различни категории разходи) и второ, размерът на доходите (категоризирани по източник на дохода – от работна заплата, доход от самостоятелна заетост, доход от лихви, дивиденти, наеми и други).

Във Великобритания се реализира статистическо съчетаване на данните от две извадкови изследвания – изследването на доходите и условията на живот (EU-SILC) и наблюдението на домакинските бюджети (HBS) с цел да се направи синтетичен набор от данни, който да съдържа едновременно информация за доходите, разходите и материалните лишения на домакинствата. Използвани са три различни метода – непараметричен хот дек метод, параметричен метод и смесен метод. Изводът, направен от националния статистически офис на Великобритания е, че като цяло смесеният метод

е малко по-ефективен при възпроизвеждането на разпределението на разходите в синтетичния файл спрямо пределното разпределение на разходите в изходния файл.

Точка „Възможности за статистическо съчетаване на данни от различни източници в българската статистическа практика“ е посветен на решаването на четвъртата задача.

На база на проучената литература по темата е направен изводът, че темата за статистическото съчетаване на данни от сложни извадкови изследвания не е разработвана в България както на теоретично, така и на практическо ниво. По тази причина е отделено специално внимание на възможностите за приложение на статистическо съчетаване на данни от различни източници в българската статистическа практика. Акцентирано е върху изследвания, провеждани от Националния статистически институт като са дадени конкретни примери за приложение и са изведени съответните приноси към анализа в социално-икономическата сфера.

Например при свързване на данни за характеристиките на основната работа от наблюдението на работна сила[5] и информация за видовете неформално обучение, в което лицата участват с източник изследването на образованието и обучението на възрастни[6], ще бъде възможно да се анализира нуждата от специфични умения и компетентности, необходими на пазара на труда, а оттам и да повиши конкретно практическата насоченост на формалната образователна система чрез актуализиране на съществуващите образователни програми, базирано на информирани решения.

Друг пример, свързан с възможности 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[7] и линията на бедност от изследването на доходите и условията на живот[8], 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.

Изследваният научен метод би могъл да бъде използван успешно и от бизнеса за повишаване на финансовите резултати. Например при съчетаване на данни от изследването за образованието и обучението на възрастни и данни за профила на лицата,

[5] https://nsi.bg/sites/default/files/files/metadata/LFS_Methodology.pdf
[6] https://nsi.bg/sites/default/files/files/metadata/Edu_Meth_AES.pdf
[7] https://nsi.bg/bg/node/3365/
[8] https://www.nsi.bg/sites/default/files/files/metadata/SILC_Metodology.pdf

използващи регулярно интернет от изследването за използване на ИКТ[9], би дало възможност на предприятията, предоставящи образователни услуги по-точно да изберат гъвкави методи за предоставяне на онлайн или хибридни обучения.

Съчетаването на данни за потреблението от наблюдението на домакинските бюджети[10] и разпределението на времето на лицата за ползване на различните видове медии с източник изследването на бюджета на времето[11], би позволило на маркетинговите компании да планират по-ефективно ресурсите за реклама според целевата аудитория.

ГЛАВА ВТОРА:ТЕХНОЛОГИЯ ЗА РЕАЛИЗИРАНЕ НА СТАТИСТИЧЕСКО СЪЧЕТАВАНЕ НА ДАННИ ОТ ДВЕ ИЗВАДКОВИ ИЗСЛЕДВАНИЯ

Глава втора е посветена на решаване на петата задача, отнасяща се до разработването на методология за статистическо съчетаване на данни от две извадкови изследвания с акцент върху различните подходи за хармонизиране и обединяване на масивите от данни.

С цел провеждане на емпиричното изследване, състоящо се в статистическото съчетаване на данните от EU-SILC и ИКТ, е предложена **технология** за апробирането му. Тя се отнася до последователността на работа при неговото провеждане и се предопределя от същността на статистическото съчетаване на данни от различни източници. Технологията е представена на фигура 1.

[9] https://nsi.bg/bg/content/2805/
[10] https://nsi.bg/sites/default/files/files/metadata/HH_HBS_Metodology.pdf
[11] https://www.nsi.bg/census2011/pagebg2.php?p2=159&sp2=160

Фигура 1: Схематично представяне на технологията за успешно реализиране на статистическо съчетаване на данни

Изборът на целевите променливи се състои в идентифициране на тези променливи, които се наблюдават по отделно в двата набора от данни и за които се търси съвместна/обща информация. Тази задача е изключително важна, тъй като изборът на целевите променливи оказва влияние върху следващите стъпки от процеса на статистическото съчетаване на данни. По конкретно, съчетаващите променливи (т.е. променливите, които се използват за свързване на две извадкови изследвания) се избират според способността им да запазят пряката връзка между целевите променливи.

При идентифицирането на общите променливи, които се наблюдават и в двата набора от данни е необходимо извършването на хармонизиращи процедури, налагащи се от различие в използваните определения (дефиниции) и/или класификации. Очевидно, ако две подобни променливи не могат да бъдат хармонизирани в достатъчна степен, то ще е необходимо те да отпаднат от процедурата. Основно изискване за общите (съчетаващи) променливи е те да нямат липсващи стойности, както и наблюдаваните им стойности да бъдат възможно най-точни (да е налице ниска или липсваща грешка в измерването). Общите променливи в двата набора от данни следва да имат еднакво маргинално/съвместно разпределение, ако двата набора от данни са представителни извадки от една и съща генерална съвкупност.

След като общите променливи са изчистени от тези, които не могат да бъдат хармонизирани помежду си, е необходимо да бъдат избрани само тези, които могат да бъдат използвани за оценка на целевите променливи. За тази цел е възможно да се приложат някои статистически методи, чиято цел е да се открие връзката между дадени променливи, като различни статистически тестове или подходящи модели за обясняване на съответната връзка.

Изборът на съчетаващи променливи е тясно свързан с избраната рамка за статистическо съчетаване, т.е. необходимо е предварително да бъде избран метода, който ще бъде използван за статистическото съчетаване на данните – микро или макро насочен, параметричен, непараметричен, смесен и т.н.

След избора на подходяща рамка за статистическото съчетаване, което е от интерес, се прилага съответния метод за статистическо съчетаване на данните, за да се съчетаят извадковите набори от данни.

Накрая е необходимо да се направи оценка на качеството на получените резултати. Качеството на процедурата, използвана за съчетаване, може да се оцени като се анализира как синтетичният набор от данни:

- запазва пределното разпределение на импутираната променлива (референтната стойност е пределното разпределение, оценено от набора с данни, избран за донор);
- запазва съвместното разпределение на импутираната и съчетаващите променливи (референтната стойност е съвместното им разпределение, оценено от набора с данни, избран за донор).

За да се сравнят пределните или съвместните разпределения на променливите в синтетичния набор от данни по отношение на донора, може да се приложи „емпиричен подход“, който се състои в сравняване на съответните разпределения, оценени от синтетичния и от донорския набор от данни, чрез мерки за сходство/различие: индекс на различие, коефициент на припокриване на две разпределения, коефициент на Bhattcharya, разстояние на Хелингер. Изброените мерки за различие и сходство на разпределенията са тясно свързани помежду си и на практика са взаимнозаменяеми при сравнението на пределните или съвместните разпределения на променливите в синтетичния набор от данни по отношение на донора.

Следвайки методиката на Van der Laan[12] относно действията, които трябва да бъдат предприети с цел хомогенизиране на два отделни източника на данни, за да бъде тяхното съчетаване успешно, са описани подробно начините за съгласуване на изместени източници на данни, на несъвместими определения, методите за избор на съчетаващи променливи, както и начините за оценка на точността и качеството на получените резултати от статистическото съчетаване на данни.

ГЛАВА ТРЕТА: МЕТОДИКА ЗА СТАТИСТИЧЕСКО СЪЧЕТАВАНЕ НА ДАННИ ОТ ИЗСЛЕДВАНЕТО НА ДОХОДИТЕ И УСЛОВИЯТА НА ЖИВОТ (EU-SILC) И ИЗСЛЕДВАНЕТО „ИЗПОЛЗВАНЕ НА ИНФОРМАЦИОННИ И КОМУНИКАЦИОННИ ТЕХНОЛОГИИ ОТ ДОМАКИНСТВАТА“ (ИКТ)

В тази глава е проведена апробацията на изложената в предходната глава технология като е реализирано статистическо съчетаване на данните от изследването на

[12] Van der Laan, P. (2000), Integrating administrative registers and household surveys, *Netherlands Official Statistics*, 15, 7-15.

доходите и условията на живот (EU-SILC) и изследването „Използване на информационни и комуникационни технологии от домакинствата“ (ИКТ)“ през 2020 година.

Анализът на масивите от данни от двете изследвания показва, че EU-SILC и ИКТ са достатъчно сходни помежду си по отношение на:

- изследваната целева съвкупност – всички обикновени домакинства и техните членове, живеещи на територията на страната в момента на наблюдение, като лицата, живеещи в колективни домакинства и институции са изключени като цяло от целевата съвкупност. Различието по отношение на целевата съвкупност, изследвана от EU-SILC и от ИКТ, което се идентифицира е във възрастовата група на членовете на домакинствата, които подлежат на интервю – в EU-SILC това са всички членове на домакинството на 16 и повече години, а в ИКТ са членовете на домакинството на възраст 16-74 години. По тази причина използваният набор от данни от EU-SILC при статистическото съчетаване е ограничен до наблюдаваните единици на възраст 16-74 години.
- статистическите единици на наблюдение – и в двете изследвания наблюдаваните статистически единици са домакинствата и техните членове.
- модела на извадката – извадките и на двете изследвания са двустепенни гнездови стратифицирани случайни извадки с еднакъв брой и формирани по един и същ начин страти.
- периода на наблюдение – изследването EU-SILC се е провело в периода април-юни 2020 г., а ИКТ – в периода май-юни 2020 година.
- общите променливи – двете изследвания имат 11 общи променливи, чието качество и съгласуваност са добри.

Хармонизирането на масивите от данни от двете изследвания по отношение на използваните дефиниции и класификации на общите им променливи показва, че осем от единадесетте общи променливи могат да бъдат приведени във вид, който позволява използването им като съчетаващи. Три от променливите – гражданство, вид на договора и професия, не могат да бъдат хармонизирани поради различие в начина им на събиране (начина на задаване на съответните въпроси) в двете изследвания. След извършване на необходимата допълнителна обработка – прекодиране на някои променливи с цел

постигане на еднаква степен на детайлност, са посочени 8 променливи, чието сходство в използваните дефиниции и класификации позволява използването им като съчетаващи: (1) пол, (2) възраст, разделена в шест възрастови групи, (3) икономически статус, (4) държава на раждане, (5) най-висока степен на завършено образование, (6) брой членове на домакинството, (7) степен на урбанизация и (8) район на планиране по отношение на местоживеенето на домакинството. Тук са идентифицирани и целевите променливи, които да бъдат използвани при съчетаването на данните от EU-SILC и ИКТ, а именно „еквивалентен нетен разполагаем доход на лице от домакинството“ и „лица, регулярно използващи интернет“.

При определянето на рамката за статистическо съчетаване на данните, която да бъде приложена върху данните от EU-SILC и ИКТ са взети под внимание два основни фактора:

- крайната цел, а именно получаване на синтетичен набор от данни, който да съдържа информация на микро ниво;
- видът на данните, които ще бъдат съчетавани, а именно данни от сложни извадкови изследвания.

Взимайки под внимание тези фактори е определен методът, който да бъде използван в емпиричното приложение – непараметричен микро подход.

При избора на съчетаващите променливи са взети предвид два фактора, характеризиращи идентифицираните в етапа на хармонизиране общи променливи: (1) съгласуваността на разпределенията на тези променливи в двата набора от данни и (2) обяснителната сила на тези променливи спрямо избраните целеви променливи.

За оценка на степента, до която разпределенията на променливите могат да се считат за сходни в двата набора от данни (съгласуваност на разпределенията) са използвани сравнения на претеглените честотни разпределения на тези променливи в двата набора от данни и изчисляване на разстоянието на Хелингер, което се използва като мярка за сходство в разпределението на две променливи. Резултатите показват, че разпределенията на всички идентифицирани общи променливи, освен на променливите „икономически статус“ и „най-висока степен на образование“, са достатъчно сходни, за да бъдат включени в последвалия анализ относно избора на съчетаващи променливи. След подходящо прекодиране, състоящо се в обединяване на някои от категориите на променливите „икономически статус“ и „най-висока степен на образование“, техните

разпределения в двата набора от данни също показват сходство и са включени в по-нататъшния анализ.

За идентифициране на общите променливи, които оказват статистически значимо влияние върху избраните целеви променливи, са построени подходящи регресионни модели по отделно върху данните от EU-SILC и ИКТ. Резултатите от направените анализи показват, че от една страна променливите „пол“, „икономическа активност“, „размер на домакинството“, „най-висока степен на завършено образование“ и „степен на урбанизация“ оказват статистически значимо влияние върху общия разполагаем нетен доход на еквивалентна единица. От друга страна променливите „степен на урбанизация“, „възрастови групи“, „пол“, „икономическа активност“ и „най-висока степен на завършено образование“ оказват статистически значимо влияние върху променливата „лица, регулярно използващи интернет“. На базата на тези резултати са формирани две множества от съчетаващи променливи:

- обединението на променливите, които оказват влияние върху двете целеви променливи по отделно, т.е. множеството от съчетаващи променливи съдържа общо 6 променливи – „размер на домакинството“, „пол“, „степен на урбанизация“, „най-висока степен на завършено образование“, „икономическа активност“ и „възрастови групи“;
- сечението на променливите, които оказват влияние върху двете целеви променливи по отделно, т.е. множеството от съчетаващи променливи съдържа общо 4 променливи – „пол“, „степен на урбанизация“, „най-висока степен на образование“ и „икономическа активност“.

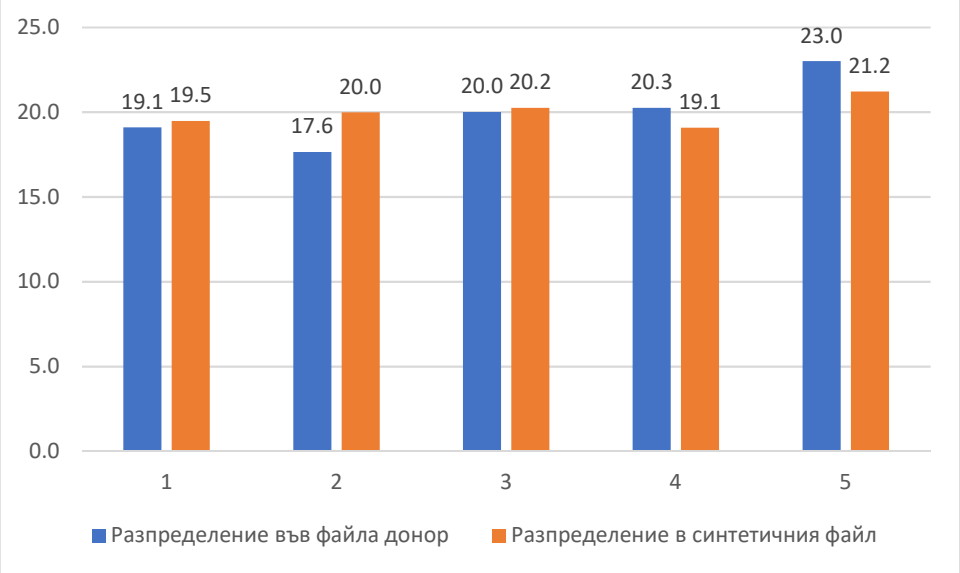
След направения избор на рамката за статистическо съчетаване на данните, на целевите и съчетаващите променливи, са приложени на практика методи за статистическо съчетаване на данните от EU-SILC и ИКТ с крайна цел получаване на синтетичен файл с микро данни. Експериментирани са 34 начина за съчетаване на данните. В тяхната основа стоят три метода – случаен хот дек метод, метод на разстоянието до най-близкия съсед и рангов хот дек метод. Различията идват от това дали съответният метод се използва претеглен или непретеглен, от това дали са зададени групиращи променливи и какъв е броя на използваните съчетаващи променливи (6, 4 или 2). Резултатите показват, че в случая на съчетаване на данните от EU-SILC и ИКТ за 2020 г. най-подходящия и успешен за реализирането на целта метод е този на

разстоянието до най-близкия съсед, когато за съчетаващи променливи се използват „пол“ и „най-висока степен на образование“. При използването на този метод са изпълнени условията за успешно съчетаване на данните от две извадкови изследвания:

1. Запазва се разпределението на целевата променлива от файла, който е определен за донор („общ нетен разполагаем доход на еквивалентна единица по квантилни групи“ от EU-SILC) както във файла донор, така и в получения синтетичен файл.
2. Запазва се съвместното разпределение на съчетаващите променливи („пол“ и „най-висока степен на образование“) и на целевата променлива от файла, който е определен за донор („общ нетен разполагаем доход на еквивалентна единица по квантилни групи“ от EU-SILC) както във файла донор, така и в получения синтетичен файл.

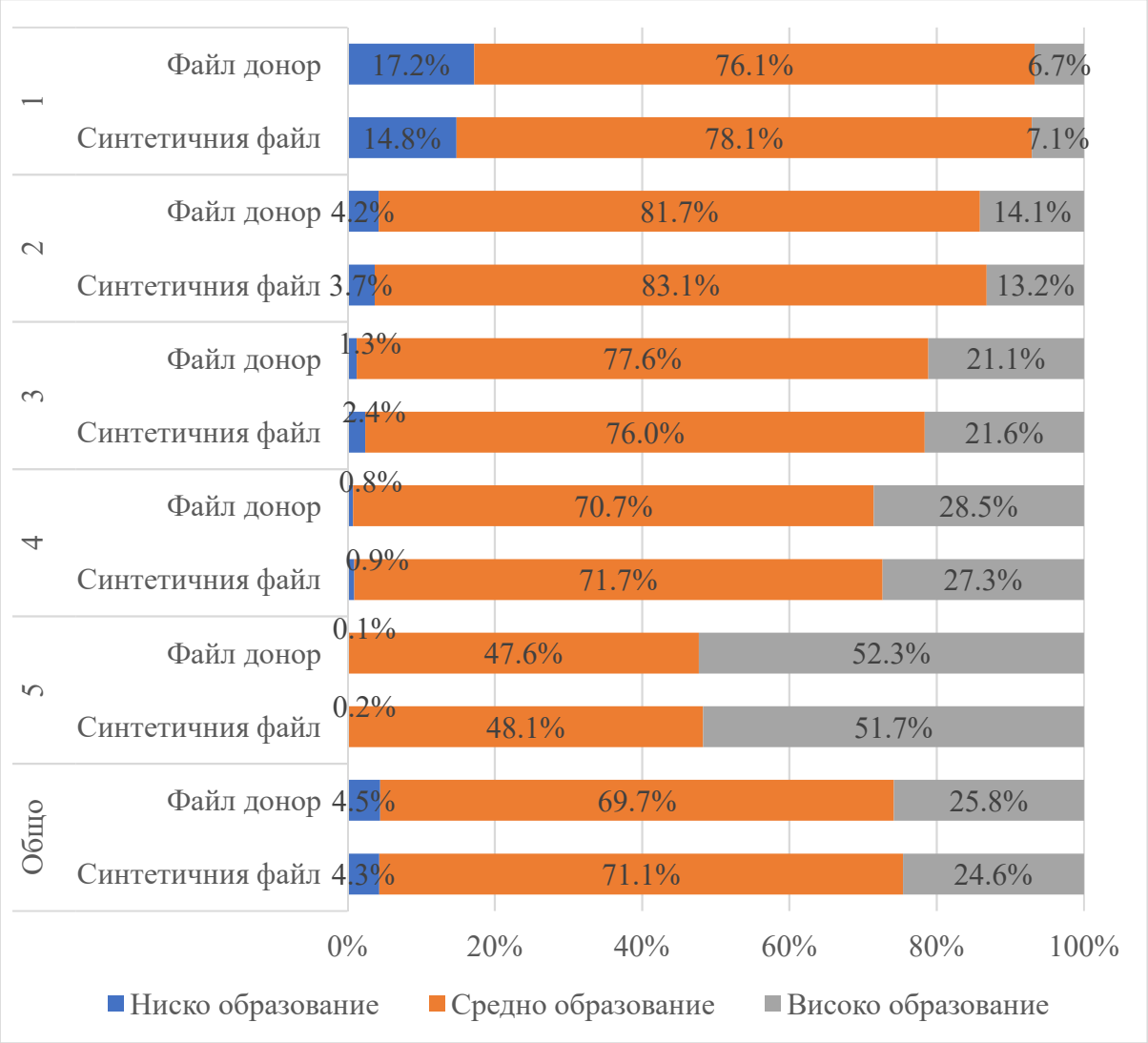
Фигура 2 показва изпълнението на първото условие за успешно съчетаване на данните от две извадкови изследвания, т.е. как в получения синтетичен файл е съхранено пределното разпределение на целевата променлива „общ нетен разполагаем доход на еквивалентна единица по квантилни групи“ съответно във файла донор (данните от изследването EU-SILC) и в синтетичния файл.

Фигура 2: Разпределение на променливата Z (общ нетен разполагаем доход на еквивалентна единица по квантилни групи) съответно във файла донор (данните от изследването EU-SILC) и в синтетичния файл

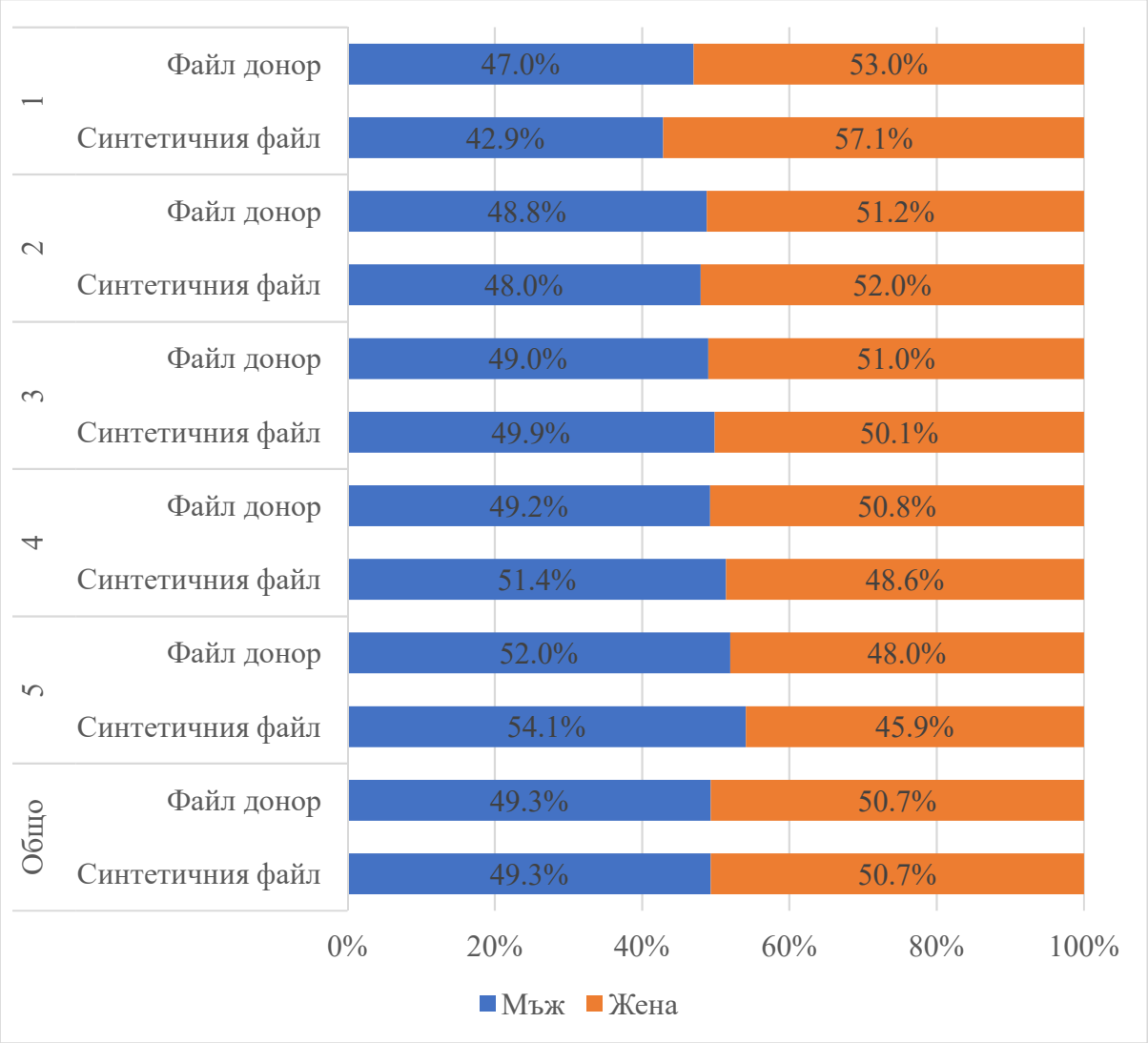


Фигури 3 и 4 показва изпълнението на второто условие за успешно съчетаване на данните от две извадкови изследвания, т.е. как в получения синтетичен файл е съхранено съвместното разпределение от една страна на променливата „най-висока степен на образование“ по квантилни групи, а от друга на променливата „пол“ по квантилни групи съответно във файла донор (данните от изследването EU-SILC) и в синтетичния файл.

Фигура 3: Разпределение на променливата „най-висока степен на образование“ по квантилни групи съответно във файла донор (данните от изследването EU-SILC) и в синтетичния файл



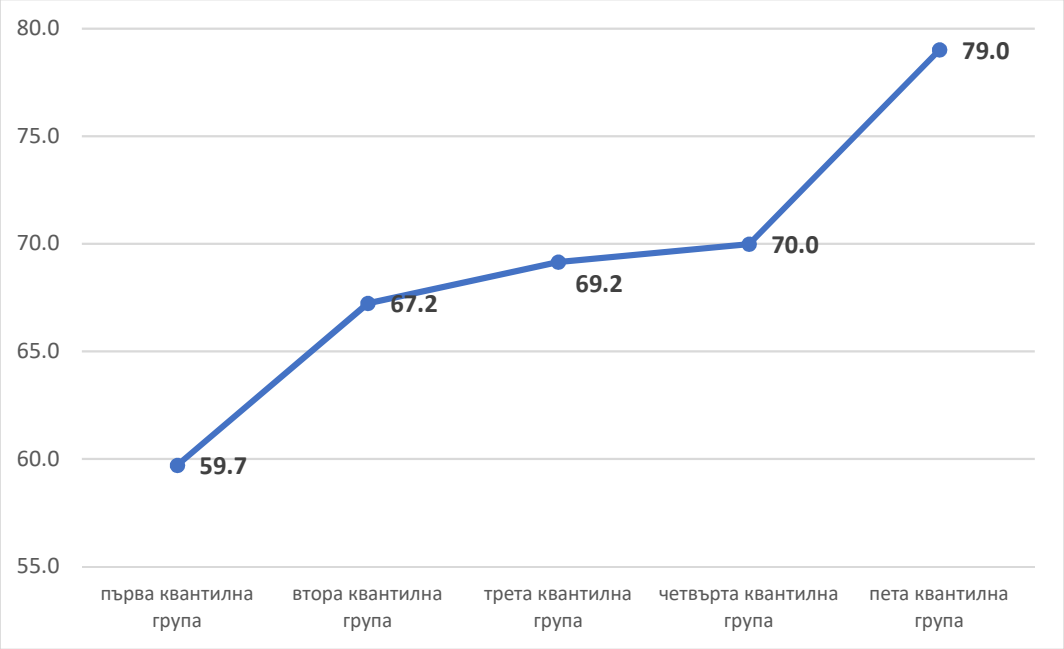
Фигура 4: Разпределение на променливата „пол“ по квантилни групи съответно във файла донор (данните от изследването EU-SILC) и в синтетичния файл



Последната точка в трета глава „Описателен анализ на получения синтетичен микро файл“ дава пример за описателен анализ на данните от получения синтетичен файл, който показва, че въз основа на тях може да се анализира детайлно разпределението на лицата на възраст 16-74 години, използващи регулярно интернет по най-висока степен на образование и по квантилни групи на нетния разполагаем доход на еквивалентна единица.

Основният извод направен въз основа на получените съчетани данни е, че въпреки че по-високото образование е основна характеристика за регулярното използване на интернет, се наблюдават и значителни различия по доходни групи (фигура 5).

Фигура 5: Относителен дял на лицата, регулярно използващи интернет по квантилни групи спрямо общия разполагаем нетен доход на еквивалентна единица



Докато 59.7% от в лицата в квантилната група с най-ниски доходи ползват регулярно интернет, то във втората група относителният дял се покачва със 7.5 процентни пункта. В третата и четвъртата групи този дял леко се увеличава, съответно с 1.9 и 0.8 процентни пункта, за да достигне 79.0% при квантилната група на лицата с най-високи доходи. Налице е сериозна диспропорция от почти 20 проценти пункта разлика при регулярните ползватели на интернет между лицата с най-ниски и тези с най-високи доходи.

Подобна информация не е налична само от един източник на данни и показва, че полученият синтетичен набор от данни е предпоставка за подобряване и обогатяване на анализа в социално-икономическата сфера.

Изследването за използване на ИКТ в домакинствата предоставя широк набор от надеждна и сравнима информация по основни социално-демографски характеристики на лицата, като пол, възрастови групи, образование и трудов статус. Съчетаването на комплекта от данни с променливи от изследването на доходите и условията на живот, разшири значително възможностите за анализ. В конкретния пример индикаторът за регулярно използване на интернет бе разпределен по квантилни групи на разполагаемия нетен доход, или данни които не са налични без да бъде направено съчетаване на данни от двете изследвания.

Представените променливи не са избрани случайно. През 2020 година на територията на Република България бяха въведени серия от противоепидемични мерки, с които използването на информационно-комуникационните технологии се превърна в необходимост за извършването на ежедневни дейности. Въвеждането на дистанционния режим на работа за голяма част от работниците и служителите, преминаването на онлайн форма обучение за учениците и студентите, организирането само на онлайн конференции, семинари и обучения, временното затваряне или ограничаването на дейността на някои обекти придадоха изключителна важност на технологиите и по-специално на регулярното използване на интернет. Откритите различия от почти 20 процентни пункта между групата на най-бедните и най-богатите лица, показват необходимостта от приемането на правителствени мерки за подпомагане на най-уязвимите групи от населението, като например въвеждане на допълнителни часове за обучение по ИКТ в учебната програма за училищата, осигуряване на възможност за безплатни курсове по компютърна грамотност за възрастни, подпомагане на бедните домакинства с техника за свързване с интернет, осигуряване на повече безплатен безжичен (wi-fi) интернет на публични места (паркове, административни сгради, читалища, обществен транспорт).

Цифровото приобщаване (digital inclusion) е част от политиката на целия Европейски съюз, с което се гарантира, че всеки може да допринесе и да се възползва от цифровия свят. Според Европейската комисия, „преходът към дигиталния свят ни донесе много нови и вълнуващи възможности. Не всеки обаче има равен достъп до тези възможности. За някои хора цифровият свят все още не е напълно достижим. За други това не е финансово достъпно. А други не бяха обучени на уменията да участват пълноценно.“[13].

ЗАКЛЮЧЕНИЕ

Статистическото съчетаване на данни може да бъде разглеждано като инструмент за повишаване ефективността на използване на наличната информация чрез прилагане на настоящите и утвърдени в статистическата практика начини за събиране на данни. В практиката много често изследователят е изправен пред избора да проведе своя анализ на базата на ограничените налични данни или да проведе ново изследване, с което да

[13] https://digital-strategy.ec.europa.eu/en/policies/digital-inclusion

събере интересуващата го информация. В по-голямата част от случаите вторият избор не е възможен основно поради финансови и организационни пречки.

Настоящият дисертационен труд е посветен на систематизиране на съществуващите методи за статистическо съчетаване на данни от сложни извадкови изследвания и дефинирането на подходяща технология, с която да се реализира успешно съчетаване на данни от такива изследвания.

В теоретико-методологичен план в разработения дисертационен труд са постигнати следните резултати:

1. Направен е анализ на възникването и развитието на идеята за статистическото съчетаване на данни от различни източници.
2. Анализирани са и са структурирани подходите за приложение на статистическото съчетаване на данни от сложни извадкови изследвания като са изведени основни проблеми при неговото реализиране.
3. Предложена е класификация на методите за съчетаване на данни от сложни извадкови изследвания според различни класификационни критерии.
4. Представени са възможности за приложение на статистическо съчетаване на данни от различни източници в българската статистическа практика.
5. Разработена е технология за статистическо съчетаване на данни от сложни извадкови изследвания с акцент върху различните подходи за хармонизиране и обединяване на масивите от данни.
6. Разработена е методика за статистическо съчетаване на данни от две извадкови изследвания, провеждани от официалната статистика в България, която е апробирана с данни за 2020 година.

На базата на проведеното емпирично изследване, състоящо се в апробация на предложената технология за статистическо съчетаване на данни от сложни извадкови изследвания, използвайки данни от изследването на доходите и условията на живот (EU-SILC) и изследването „Използване на информационни и комуникационни технологии от домакинствата“ (ИКТ)“ през 2020 година, може да се направи извода, че **формулираните изследователски хипотези се потвърждават**. Т.е. от една страна е възможно в българската статистическа практика да бъдат приложени методи за статистическо съчетаване на данни от две извадкови изследвания за получаване на

синтетичен набор от данни и от друга, полученият синтетичен набор от данни е предпоставка за подобряване на анализа в социално-икономическата сфера.

IV. СПРАВКА ЗА ОСНОВНИТЕ НАУЧНИ И ПРИЛОЖНИ ПРИНОСИ НА ДИСЕРТАЦИОННИЯ ТРУД

1. Идентифицирани са и са структурирани проблемите при статистическото съчетаване на данни от сложни извадкови изследвания и са изведени методи за реализиране на процеса при различни изходни условия.
2. Предложена е класификация на методите за съчетаване на данни от сложни извадкови изследвания според различни класификационни критерии.
3. Представени са възможности за приложение на статистическо съчетаване на данни от различни източници в българската статистическа практика.
4. Формулирани са етапите и е структуриран цялостния процес за съчетаване на данни от сложни извадкови изследвания.
5. Разработена е методика за статистическо съчетаване на данни от две сложни извадкови изследвания, провеждани от официалната статистика в България, а именно изследването за доходите и условията на живот (EU-SILC) и изследването „Използване на информационни и комуникационни технологии от домакинствата“ (ИКТ). Разработената методика е апробирана с данни за 2020 година.

V. СПИСЪК НА ПУБЛИКАЦИИТЕ ПО ТЕМАТА НА ДИСЕРТАЦИОННИЯ ТРУД

1. Цонкова, Ц. (2020), Статистическо съчетаване на данни от две извадкови изследвания. Пример с Nearest neighbor hot deck, Сборник с доклади от международна научна конференция „Новите информационни технологии и големите данни: възможности, перспективи при анализите и управленските решения в бизнеса, икономиката и социалната сфера“, Издателски комплекс – УНСС, 270-278.
2. Цонкова, Ц. (2022), Методи за статистическо съчетаване на данни от сложни извадкови изследвания, Сп.Статистика (под печат).