ПЛОВДИВСКИ УНИВЕРСИТЕТ „ПАИСИЙ ХИЛЕНДАРСКИ''
ФИЛОСОФСКО-ИСТОРИЧЕСКИ ФАКУЛТЕТ
КАТЕДРА „ПРИЛОЖНА И ИНСТИТУЦИОНАЛНА СОЦИОЛОГИЯ“

Донка Атанасова Кескинова

ПРОБЛЕМИ И ПОДХОДИ ПРИ ПРИЛОЖЕНИЕТО НА КЛЪСТЕРЕН АНАЛИЗ ЗА ИЗГРАЖДАНЕ НА СЪСТАВНИ ИНДИКАТОРИ В СОЦИОЛОГИЯТА

АВТОРЕФЕРАТ

на дисертационен труд за присъждане на образователната и научна степен „доктор”

Област на висше образование: 3. Социални, стопански и правни науки
Професионално направление: 3.1. Социология, антропология и науки за културата
Докторска програма: Социология (Приложение на статистическите методи в социологията)

Научен ръководител: **доц. д-р Елица Куздова Димитрова**

Пловдив, 2017

СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

ВЪВЕДЕНИЕ

„Клъстерен анализ“ е общо название на множество многомерни статистически методи (техники, алгоритми) за класификация (групиране) на обекти в хомогенни групи, които не са известни предварително.

Необходимост от класификация възниква в много научни области. Всяка от тях влага различен контекст и използва различен термин за нея. В дисертационния труд са въведени както термините таксономия, типология и сегментация, които се използват взаимозаменяемо в социалните науки, така и клъстеризация и класификация, които имат различно съдържание в статистиката.

Потребност от типология практически възниква във всяко ЕСИ. Въпреки възможността тя да се изгражда с клъстерен анализ, този анализ не се използва често в социологията. Причините се търсят в *практически проблеми с изпълнението на анализа, парадигмата на изразените с уравнения модели, размерността на анализираните данни, скалата на измерване на анализираните данни.* В дисертационния труд се аргументират като причини и *незадоволително представяне на методологията на клъстерния анализ пред социологията, слабо развита методология и методика на приложение на клъстерния анализ в социологията*, *недостатъчно ясно изявени възможности на клъстерния анализ в социологията.*

По отношение на последната причина, традиционно възможностите на клъстерния анализ както в социологията, така и в останалите научни дисциплини се описват с изследователските задачи, които могат да се решават с него, а не с продукта от него. В социологията продуктът от клъстерния анализ най-често е типология. Могат да се разграничат две форми на типология и съответно два типа класификации в социологията - *типологичен съставен индикатор и емпирична типология.*

В основата на това разграничение, от една страна, стоят степента на обобщение и познавателната същност на резултата. *Съдържателно* съставният индикатор дефинира нов признак, който не може да бъде регистриран непосредствено, но от който възниква необходимост при описанието на изучаваното социално явление. Емпиричната типология дефинира типове на основата на съдържателната интерпретация на резултата от клъстеризацията и е част от типологичния анализ като самостоятелен метод за изучаване на социалните явления. *Формално* двата случая могат да бъдат разграничени по броя на променливите в основата на клъстеризацията и по броя на измеренията, към които те принадлежат. Изграждането на съставен индикатор предполага включване в анализа на ограничен брой променливи от едно измерение, а емпиричната типология – голям брой променливи от различни измерения.

От друга страна, в основата на разграничението стоят и различните изисквания към резултата от клъстерния анализ, а от там и специфична методология на приложение на анализа. При съставните индикатори стремежът по-скоро е да се осигури висока степен на хомогенност на

клъстерите, равен принос на всяка от променливите, на основата на които съставният индикатор се изгражда, достатъчен брой случаи в разновидностите на новата променлива. При емпиричната типология акцентът е върху хетерогенността между клъстерите - типовете трябва да са носители на различни характеристики, а оттам и върху подбора на променливите и валидизацията на резултата.

Тъй като въведеното разграничение в продукта от класификацията и съответно резултата от клъстерния анализ в социологията е ключово за дисертационния труд, то е разгърнато по-подробно в т.1 от първа глава.

Клъстерният анализ в повечето случаи се обвързва с емпирична типология в социологията. *Не откриваме публикации за клъстерен анализ, в които да се прави разграничение между двата типа класификации и съответно двете форми на типология в социологията.* У нас няма, а в чужбина не ни е известно да има изследване, свързано с методологията на клъстерния анализ както с цел изграждане на типологични съставни индикатори, така и с цел емпирична типология в социологията. *В повечето случаи изследванията върху клъстерния анализ са насочени към конкретен методологически проблем, а не към конкретно приложение. Последното стои като задача пред съответната научна област, която трябва да изследва и осветли как резултатите от тези изследвания могат да се използват при конкретно приложение*. За разлика от емпиричната типология, където методологията на приложение на клъстерния анализ може да заимства от сегментацията в маркетинга, то при съставните индикатори възможностите за това са по-ограничени.

Предмет на изследване в дисертационния труд е приложението на клъстерния анализ за изграждане на съставни индикатори в социологията.

На основата на прегледа на проблемите, свързани с приложението на клъстерния анализ в социологията е формулирана следната **изследователска теза:** *ефективното използване на клъстерния анализ за изграждане на типологични съставни индикатори в социологията се ограничава от незадоволително представената и слабо развитата методология и методика на приложение на анализа.*

В тази насока, **целта на дисертационния труд** е да запълни този дефицит в методологията на анализа на данни в социологията у нас чрез извеждане и обосноваване на възможни методологически решения (подходи) при приложението на клъстерния анализ за изграждане на съставни индикатори.

В тази връзка са формулирани следните **изследователски задачи**:

- Да се осветлят възможностите на клъстерния анализ при анализа на данни в социологията от гледна точна на продукта (резултата) от анализа;
- Да се изследват възможностите на клъстерния анализ, като се проучи и опише методологията му, систематизират се резултатите от изследванията по всяко едно решение, пред които е изправен изследователят в хода на анализа, и се аргументират определени стратегии на приложението на клъстерния анализ в социологията;

- Да се изследват възможностите на клъстерния анализ за изграждане на съставни индикатори и се изведат методологически решения по отношение на различните етапи от изпълнението на анализа за изграждането на този тип индикатори.

Обект на изследване в дисертационния труд са класическите методи за клъстерен анализ, които традиционно се използват в социалните науки. Характеристиките на тези методи, както и мястото им сред останалите групи методи за клъстерен анализ, се дискутират в т. 2 от първа глава.

Проверяваните **изследователските хипотези** с настоящото изследване са:

- Уорд методът за клъстерен анализ е подходящ за изграждане на съставни индикатори (променливи) в социологията. Приложението му с Евклидово разстояние вместо квадратно Евклидово разстояние, което се налага от строги статистически ограничения, може да повиши ефективността на метода.
- Изборът на адекватен метод и мярка е необходимо, но недостатъчно условие за ефективното приложение на клъстерния анализ при изграждане на съставни индикатори. Подходът към променливите е от решаващо значение за ефективността на клъстеризацията.

В изследователската работа по дисертационния труд са използвани следните **подходи**: сравнителен анализ на емпирични и симулационни изследвания върху методи и процедури, свързани с клъстерния анализ; емпирични изследвания върху реални данни; статистически подход при оценка на ефективността на изследваните методологически решения.

При разработването на дисертационния труд са използвани общо 224 литературни източника, от които 38 на български и руски език и 186 на английски език, както и 15 линка към електронни ресурси.

Първа глава. КЛЪСТЕРНИЯТ АНАЛИЗ В СОЦИОЛОГИЯТА

Клъстерният анализ може да се използва за различни цели. Например: разработване на типологии и класификации; изследване на концептуални схеми за групиране на обекти; генериране на хипотези; проверка на хипотеза за присъствие в данните на определени типове; анализиране на структурата на данните; свързване на различни аспекти на данните един с друг; съкратено описание на изучаваната съвкупност от обекти; построяване на извадки; изучаване на връзки между признаци; анализ на „механизмите“, определящи характера на дадено явление. За методологията и методиката на приложение на клъстерния анализ обаче е от значение продуктът от класификацията в социологията. В тази насока в тази глава са конкретизирани възможностите на клъстерния анализ както от гледна точна на продукта от анализа, така и от гледна точка на ролята на този продукт за изучаване на социалните явления чрез съвместно използване на няколко многомерни статистически анализа.

1. Възможности на клъстерния анализ в социологията

Клъстерният анализ *създава нова категорийна променлива* на основата на няколко изходни променливи – първични или производни. В социологията тази нова променлива може (най-вече) да задава разновидностите на *типологичен съставен индикатор* или *емпирична типология*.

1.1. Изграждане на типологичен съставен индикатор

Емпиричните, както и теоретичните индикатори, имат своя йерархия – степен на обобщение. Най-ниско в тази йерархия са първичните (простите) индикатори, следвани от междинните, а най-високо са съставните индикатори. Степените в тази йерархична структура зависят от сложността на конкретната характеристика на изследвания обект, която индикаторите обективизират. Задачата на изследователя в ЕСИ е да сведе всички индикатори до прости. За да се получи информация за по-общ признак, който се наблюдава с множество прости индикатори, информацията от тези индикатори трябва да бъде обобщена в нов индикатор и съответно нова променлива в базата данни. Този индикатор/променлива в дисертационния труд се нарича *съставен/съставна* независимо дали е краен или междинен продукт.

Още през 80-те години на XX век е изявена възможността клъстерният анализ да се използва като инструмент за изграждане на съставен индикатор и по-точно, на съставна променлива, защото той дефинира нейните разновидности. Въпреки това, той не се препоръчва за тази цел през този период у нас. Причините са различните резултати от различните алгоритми; възможността получените структури да не съответстват на съдържателната представа на социолога и невъзможността в този случай да се оцени дали разликата е в резултат на неправилна понятийна представа на изследователя или е под влиянието на формалната структура на даден алгоритъм.

Наистина приложението на клъстерния анализ в социологията често води до нехомогенни групи, които трудно се поддават на интерпретация, както и до нестабилни групи, чието съдържание се променя при незначителни промени в подхода на прилагане на анализа, което налага субективен избор от страна на изследователя. Тези трудности обаче са недостатъци от гледна точка на една позитивистка представа, която мисли изследваните обекти като притежаващи устойчиви, предварително известни характеристики и намиращи се в стабилни отношения помежду си. От гледна точка на една социология, която еманципира обекта на изследване и конструира индикаторите си в диалог с него, тези трудности всъщност са предимство. Те разкриват многомерната природа на изследваните явления и карат изследователя да търси адекватни интерпретации, за да обясни хетерогенността и нестабилността на групите.

Така, макар и на по-късен етап - началото на XXI век, изграждането на съставни индикатори чрез клъстерен анализ навлиза в българската социология. Свидетелство за това е публикация на автора на дисертационния труд (в съавторство).

В заключение, за агрегирането на първичната информация при изграждане на съставни индикатори в социологията се използват три вида съставни измерители - индекс, скала и типология. *Индексът* обобщава резултати от първични индикатори, без да засяга тяхната взаимосвързаност. За разлика от индекса, с*калата* се построява на основата на айтеми, между които има логическа или емпирична връзка. *Типологията* е набор от категории, получени от сечението на два или повече признака. Ординалният характер на скалата и индекса ги прави предпочитани измерители, но в социологията има нужда и от типологични съставни измерители. Последните могат да се конструират „автоматично“ чрез клъстерен анализ. При подходяща методология на приложение, клъстерният анализ може да намери устойчиви комбинации от признаци, които да отговарят на логиката на изграждания съставен индикатор. Ако изследователят има колебание между няколко варианта за даден съставен индикатор, в резултат на различни методически решения в процеса на приложението на клъстерния анализ, то изборът трябва да се прави на основата на дискриминиращите им възможности спрямо други променливи от анализираните данни.

Изградените чрез клъстерен анализ типологични съставни променливи (индикатори) могат да се използват самостоятелно или да бъдат включени в друг многомерен статистически анализ. Използването на типологични съставни променливи в описанието на данните превръща едномерните и двумерните разпределения в многомерни, а в многомерния анализ те намаляват размерността на изходното пространство, като запазват многомерността на информацията и решават проблеми, свързани с методологически изисквания по отношение на броя на анализираните променливи, броя на случаите в категориите на променливите, независимост на променливите. Затова често клъстерният анализ се използва като „*трамплин*“ към друг многомерен анализ, което е израз на

така наречения *комплексен подход* при приложение на два статистически анализа.

Добре познато е последователното използване на клъстерен и множествен анализ на съответствията (multiple correspondence analysis). У нас има както изследвания върху приложението на множествения анализ на съответствията в социологията, така и публикации за ролята му при анализа на данни в социологията и за съвместното му приложение с клъстерен анализ. Този комплексен подход се прилага и в изследванията в дисертационния труд.

Получените съставни променливи могат да се използват като независими променливи и в други анализи, например в регресионния анализ, където ще редуцират броя на бинарните (дихотомните) променливи и/или ще решат проблема с колинеарността на променливите в регресионното уравнение. Комплексният подход обвързва клъстерния анализ с факторния анализ и други анализи извличащи латентни променливи. В този случай клъстерният анализ е втори, а анализите, които го предхождат, имат за задача да редуцират броя на променливите, на основата на които ще се клъстеризират случаите, а не да използват резултата от него.

1.2.Изграждане на емпирична типология

Типологията е многомерна класификация за дефиниране на „типове“ в социалните науки. *Типът* е конструкт, създаден на основата на комбинация на стойностите на няколко променливи. Има различни концептуализации на типологиите в социологията. Една[1] от тях разграничава традициите (подходите) при изграждане на типология едновременно по две измерения: качествена – количествена и евристична – емпирична. В по-старата – качествената, наричана още вербална традиция, типологията е концептуална (дедуктивна). От своя страна, в зависимост от статуса на типовете – дали са ментални конструкти или имат емпиричен корелат, качествената типология може да бъде евристична („идеален тип“) или емпирична („етнографски тип“ или тип, извлечен от качествено изследване). С компютъризацията се развива количественият подход в типологията, който е изцяло емпиричен. Тук типовете се извличат чрез многомерни статистически методи - клъстерен анализ или числова таксономия, от емпиричните данни. Процесът на построяване (изграждане) на емпирична типология се нарича емпирична типологизация и се определя като най-силният анализ в описателен план.

След навлизането на регресионните модели в социалните науки, конструирането на емпирични типологии в социологията става второстепенно. Мотивите за това са, че „типологиите са главно описателни, възникват в ранния етап на научния анализ, и по същество са сурови или наивни формулировки. За разлика от тях, появилите се по-късно модели се фокусират върху обяснение и прогнозиране а не на описание. Опасен мит е

[1] Bailey, K.D. Typologies. In: Borgatta, E.F., Montgomery, R.J.V. (eds.). Encyclopedia of Sociology, Second Edition, 2000, Vol. 5, 3180-3189, Macmillan Reference USA.

да се мисли, че социологията е „надраснала“ нуждата от типологии“ (пак там, с. 3185). В дисертационния труд са дадени примери за изграждане на емпирична типология чрез клъстерен анализ у нас.

Клъстерният анализ може да се използва и за потвърждаване на принадлежността към определен тип – съдържателно еднородна група от случаи, в резултат на което да се проведе по-обоснован статистически или визуален анализ на емпиричната информация. В този случай клъстерният анализ се използва паралелно на друг многомерен статистически анализ. Той е особено полезен като диагностично средство при анализи, изградени на предположение за хомогенност на данните като цяло или по групи, например факторен анализ, дискриминантен анализ, анализ на вариацията, линеен регресионен анализ.

Много популярно е паралелното използване на клъстерен анализ и многомерно скалиране. Различната чувствителност към допълващи се един друг аспекти на информацията, съдържаща се в данните, прави в много случаи уместно приложението на двата анализа към едни и същи данни. В случая, когато многомерното скалиране е водещ анализ, клъстерният анализ е спомагателно средство за интерпретирането на изобразените в равнината пространствени конфигурации в резултат от многомерното скалиране. Когато клъстерният анализ е водещ, тогава многомерното скалиране е спомагателно средство за описание на клъстерите чрез визуализирането им в двумерното пространство. Този подход е приложен и в дисертационния труд.

Резултативната променлива от клъстерния анализ, която разграничава типовете при емпирична типологизация, често се използва като зависима променлива в друг анализ. Този комбинационен подход е оправдан, само ако клъстерният анализ осигурява групите (типовете), а задачата на следващия анализ е да определи принадлежността на нови случаи към тях. Например дискриминантният анализ използва участвалите в клъстерния анализ променливи линейно, за да определи клъстерната принадлежност на нови случаи, наричано „*симбиотично партньорство*“ между двата анализа. Използването на втория анализ с цел доказване наличието на разлики между типовете, т.е. валидизиране на резултата от клъстерния анализ, е неприемливо.

Обобщение: И двата продукта, които могат да се получат с клъстерния анализ – типологичен съставен индикатор и емпирична типология, са от съществено значение за описанието на социалните явления и могат да допринесат за разработването на хипотези, концепции и теории в социологията. Това ни дава основание да определим клъстерния анализ като изследователска техника, със степен на важност за социологията, каквато имат и останалите описателни методи. В специализираната литература клъстерният анализ се обвързва с емпирична типология, но не и с типологични съставни индикатори, а именно те могат да бъдат инструмент за по-ефективно описание на данните във всяко социологическо изследване, независимо дали е подчинено на дедуктивен или индуктивен подход. У нас няма изследвания, които да разглеждат методологически проблеми,

свързани с приложението на клъстерния анализ както за изграждане на емпирични индикатори, така и за емпирична типология в социологията. Това очертава потребност от изследвания в областта на всяко от двете приложения на клъстерния анализ в социологията.

Акцентът върху клъстерния анализ в дисертационния труд по никакъв начин не омаловажава останалите статистически методи в социологията. През XXI век пред социологическата квантификация има много предизвикателства, въпреки че вероятно и в бъдеще социологията ще продължи да използва основно регресионен анализ на основата на различни линейни и логистични модели. Нашето очакване е, че по-често ще бъдат използвани и относително пренебрегваните методи за редуциране на пространството от типа на факторен анализ, многомерно скалиране, анализ на вариацията, канонична корелация и т.н. Сред тези методи трябва да намери място и клъстерният анализ. Още повече, че клъстерният анализ, основаващ се на връзки между случаи, няма алтернатива в лицето на методите за ординация, основаващи се на връзки между променливи, като факторен анализ, анализ на главните компоненти, многомерно скалиране, множествен анализ на съответствията и др., когато:

- Целта е да се получи една единствена променлива, с която да се определи еднозначно принадлежността на всеки случаи към даден тип, което не се постига с латентни променливи;
- Изходните променливи не са измерими или са измерими, но за целите на изследването не е необходимо да се установява степента на тяхното проявление, а само наличие или отсъствие на даден признак. В този случай регистрацията се свежда до един множествен въпрос, на който различните признаци са разновидности, а обобщението касае съвкупност от бинарни променливи, към които не са приложими методи за ординация;
- Редукцията обхваща малко на брой променливи и извличането на латентни променливи се обезсмисля;
- Целта е да се навлезе в дълбочина на типологията, т.е. в подтипове на типовете, което трудно може са се постигне с няколко измерения, получени с метод за ординация.

Трябва да отбележим обаче, че клъстеризацията е по-ефективна, когато се изгражда на основата на обективни характеристики, за разлика на методите за ординация, които са по-ефективни към субективни характеристики – ценностни ориентации и нагласи, в които връзките са много по-нестабилни и затова често се търсят на ниво променливи, а не на случаи.

Важно е да отбележим също така, че клъстеризацията се прави от статистическия метод, а типологията, в каквато и да е форма, от социолога. Затова той трябва да се чувства свободен в прехода от класификация към типология да обединява или да дели отделни клъстери, независимо от решението на метода.

2. Методи за клъстерен анализ в социологията

Едва ли има друг статистически анализ, който да е обобщено название на толкова много методи. Това се предпоставя както от липсата на универсална дефиниция за това, какво е клъстер или какво е „добър“ клъстер, така и от интердисциплинарното приложение на анализа. Всяка научна област обаче използва определени методи за своите нужди. В социалните науки това са *класическите (традиционните, стандартните) методи*, разработени в ранния етап от развитието на клъстерния анализ, но радващи се на широко приложение и днес. Именно тези методи се включени в изданията, представящи статистически методи пред социалните науки както у нас, така и в чужбина, и в универсалните софтуери за статистически анализ, използвани от тези науки.

На основата на различните парадигми и подходи, както и видове клъстеризации и клъстери, до които те водят, класическите методи за клъстерен анализ се определят като прототип или централно базирани, минимизиращи показатели с невероятностна интерпретация или основани на сходство, които пораждат твърди, непрепокриващи се и обхващащи всички случаи клъстери.

Класическите методи за клъстерен анализ са разработени през 60-те и 70-те години, а са изследвани усилено през 70-те и 80-те години на XX век. Днес те рядко са обект на изследвания. Налице е обаче развитие във времето на методологията на приложение на класическия клъстерен анализ. Докато първоначално тя се разглежда като последователност от стъпки, отговарящи на основните решения, които трябва да се вземат в процеса на анализа, то днес в нея важно място заема всяко едно от решенията, пред които е изправен изследователят на всяка стъпка от анализа. Акцентът е изместен от решението за избор на метод, върху решения, касаещи избора на променливи, начина на подреждане на случаите в базата данни, прилагане на формални критерии за определяне на броя на клъстерите и валидизация на резултатите, използване на формални критерии за сравнение на различни решения и построяване на „средно“ клъстерно решение. Именно тази методология на анализа е описана (втора глава) и приложена в изследванията в рамките на дисертационния труд (трета глава), а в резултат на нея са изведени методологически решения, които могат да повишат ефективността от приложението на клъстерния анализ за изграждане на съставни индикатори (променливи) в социологията.

Втора глава. МЕТОДОЛОГИЯ НА КЛЪСТЕРНИЯ АНАЛИЗ

Извеждането на методологически решения по отношение на приложението на клъстерен анализ за изграждане на съставни индикатори в социологията обуславя необходимост от задълбочено проучване на методологията на анализа. В главата са систематизирани възможните решения, пред които е изправен изследователят в процеса на изпълнение на клъстерния анализ, обосновани са причините да се предпочетат дадени

решения, както и последиците от тях върху резултатите от клъстеризацията. Направен е опит да се оцени адекватността на някои възможни подходи към конкретни изследователски задачи в социологията. Изведени са методологически проблеми, към които да се насочи изследването в дисертационния труд.

От съществено значение за методологията и методиката на приложение на клъстерния анализ са резултатите от сравнителните изследвания и имплементирането на анализа в софтуерите за статистически анализ. Затова решенията в клъстерния анализ са обвързани с изводите от сравнителни изследвания и възможностите за тяхната реализацията в IBM SPSS Statistics и Stata. Паралелното описание на методологията на клъстерния анализ с двата софтуера от една страна, разкрива алтернативни възможности пред изследователя, а от друга страна, противопоставя две стратегии по отношение на анализа. За компенсиране на дефицити по отношение на клъстерния анализ в тези софтуери се препраща към софтуерите R и SAS/STAT.

Изложението на методологията следва етапите в приложението на анализа.

1. Избор на променливи

В дисертационния труд изборът на променливи се дискутира в два аспекта – *релевантност* и *дискриминиращата роля.* След което се поставят ред въпроси относно избраните за основа на анализа променливи, за които няма еднозначни отговори. Те касаят скалата, на която да се анализират променливите; прилагането на стандартизация или трансформация на променливите; използването на латентни променливи (факторизация); използването на тегла. Решението за използването на тези преобразувания е изцяло в ръцете на изследователя, който трябва да го вземе, от една страна, на основата на броя, съдържанието, скалата и формата на разпределение на променливите, а от друга страна, на основата на използвания метод и мярка за близост. Едното от емпиричните изследвания в дисертационния труд, описано в първа точка от трета глава, е насочено към *извеждането на подходи към променливите* с цел по-ефективна клъстеризация от гледна точка на изграждане на съставни индикатори в социологията.

2. Избор на случаи

Класическите методи за клъстерния анализ не налагат никакви условия към анализираната съвкупност по отношение на обхвата на случаите (обектите) и данните за тях. Анализът може да се прилага към всяка изследвана съвкупност, вкл. всеки вид извадка.

Дискутират се три решения по отношение на анализираните случаи - отклоняващи се случаи, липсващи данни, начинът на подреждане на случаите в базата данни. За разлика от първите две, последното решение е нетрадиционно за статистическите анализи. Причините за влиянието на реда на случаите върху клъстеризацията са осветлени при разглеждане на алгоритмите на методите за клъстеризация, а последиците от различни

начини на подреждане е предмет на емпирично изследване в дисертационния труд.

3. Избор на мярка на близост

Изследователят трябва да избере мярка за близост между клъстеризираните случаи (обекти), която концептуално да съответства на характера на данните и целите на анализа. За тази цел са описани характеристиките, предимствата и недостатъците на различните видове мерки за близост и техни основни представители, както и е оценена приложимостта им към конкретни случаи в социологията. Смята се, че изборът на мярка за близост е свързан и със скалата на измерване на променливите, по-точно, че коефициентите на корелация и мерките за разстояния са приложими към интервално скалирани променливи, а коефициентите на асоциация към бинарни променливи. В дисертационния труд се доказва емпирично ефективността на мерките за разстояние и към бинарни променливи. Изборът на мярка зависи и от метода – връзка, разкрита от сравнителните изследвания на различните методите.

4. Избор на метод

Резултатите от клъстеризацията по-често отразяват свойствата на използвания метод, отколкото вътрешната структура на данните. Затова изборът на метод трябва да стъпи на добро познаване както на същността, така и на силните, и слабите страни на различните методи. С тази цел в раздела са описани методите в подкатегорията „Свързващи методи“ от агломеративните йерархични методи и „*K*-средните“ от итеративните. В заключителен раздел към двете групи методи са систематизирани математически свойства на методите, наричани *критерии (условия) за допустимост* и възможността те да се използват при клъстеризация в социологията. Представен е и двустъпковият клъстерен анализ в IBM SPSS Statistics, като възможност да се компенсира липсата на мярка за близост, приложима към променливи на различни скали в този най-често използван в социологията у нас софтуер.

Систематизирани са резултатите от сравнителните изследвания за ефективността на различните методи за клъстерен анализ. По отношение на *йерархичните методи* те разкриват превъзходството на два от тях – Уорд метода и метода на средното (междугруповото) свързване. Факторите, които оказват влияние за по-голямата ефективност на единия или другия от двата метода са:

- *мярката за близост* – Уорд методът е по-добър с Евклидово разстояние (ED), а методът на средното свързване е по-добър или еквивалентен на Уорд метода с коефициент на корелация на Пирсън;
- *обхватът (начинът на третиране на отклоняващите се случаи)* – Уорд методът е по-добър при пълен обхват, т.е. когато се клъстеризират всички случаи, а методът на средното свързване - при непълен обхват, т.е. когато отклоняващите се случаи се отстранят;
- *степента на препокриване* – Уорд методът работи добре при наличие на препокриващи се клъстери (възможността случаите да

принадлежат към различни клъстери), а методът на средното свързване при непрепокриващи се клъстери.

Изведените връзки не трябва да се приемат безрезервно, защото изключения има както сред обобщаваните изследвания, така и в изследванията за тяхното валидизиране.

От *итеративните методи* - методът на конвергентните k-средни, при които клъстерните центрове се обновяват след всяко преместване на случай, има тенденцията да възстановява добре клъстерната структура.

По отношението на *сравнението между йерархични и итеративни методи* изводът е, че итеративните методи с неслучайни начални центрове са толкова добри, колкото и най-добрите йерархични методи или по-добри от тях, тъй като ефективността им се влошава по-малко с увеличаване на обхвата, в сравнение с йерархичните методи.

Изводите от сравнителните изследвания са два. *Първият* е, че без йерархичен метод не може да се поучи добра клъстеризация. Следователно, *йерархичната клъстеризация трябва да бъде приложена задължително, независимо дали след това тя ще се оптимизира чрез итеративния алгоритъм на k-средните.* Тъй като Уорд метода и *k*-средните използват (по различен начин) една и съща оптимизационна функция, е логично те да бъдат използвани допълващо се – първо Уорд метода, след което *k*-средните. *Вторият* извод на основата на факторите, които влияят на ефективността на *Уорд метода,* е, че той е *много подходящ за социологията*, където: 1) характерът на променливите предполага по-често използване на разстояние пред корелация като мярка за близост; 2) клъстерите са припокриващи се; 3) използвайки клъстерния анализ като „трамплин“ към други многомерни анализи, изследователят е заинтересован да клъстеризира цялата база данни, без да изключва случаи.

Заключението от практическото приложението на клъстерния анализ в социологията е, че Уорд методът дава значително по-често хомогенни клъстери, които лесно могат да бъдат интерпретирани съдържателно. *Така интерпретативността на клъстерите, както и близкия им размер и достатъчен брой случаи в тях, правят Уорд метода подходящ инструмент за изграждане на съставни индикатори и емпирична типологизация в социологията.* Изследванията в рамките на дисертационния труд имат за цел да проверят този извод.

Накрая на раздела е дискутиран проблемът с мярката за близост в изходната матрица при Уорд метода. Изводът от сравнителните изследвания е, че Уорд методът работи по-добре с Евклидово разстояние (ED), а съгласно стратегията за съгласуваност, изискваща използване на една и съща мярка за близост между случаите и между клъстерите, Уорд методът трябва да се прилага с квадратно Евклидово разстояние (SED). Това противоречие, вероятно породено от начина на имплементиране на метода – автоматично повдигане на квадрат на ED, поставя пред изследователя *въпроса с какво разстояние е правомерно да бъде използван методът– Евклидово или квадратно Евклидово*, особено след като установи, че с ED се получават по-хомогенни клъстери. В тази връзка са изведени аргументи за използването

на Уорд метода с ED вместо с SED, а в емпиричните изследвания, от една страна се потвърждава ефективността на Уорд метода с ED, а от друга се изследва ефектът от използването на двете разстояния при клъстеризация по бинарни променливи.

5. Определяне броя на клъстерите

Съществуват два подхода за определяне броя на клъстерите - емпиричен и формален. При емпиричния подход изследователят прилага различни нива на клъстеризация и избира това от тях, при което клъстерите са ясно разграничени и интерпретативни. Този подход е по-широко застъпен в социалните науки, където не броят и йерархията на групите, а разкриване на зависимости е цел на клъстеризацията. Това не означава, че не трябва да се използват формални критерии. Те само могат да допринесат към емпиричното заключение, потвърждавайки го или подтиквайки изследователят да направи обоснован избор между емпиричното и формалното решение.

Формалният подход е свързан с използването на процедура за математическо или статистическо разграничаване на клъстерите, наричани „правила за спиране“ при йерархичните методи. В дисертационния труд са описани различни начини на използването на коефициент на сливане, както и множество *формални правила на основата на индекси*. Трудностите с използването на правилата за спиране произтичат от противоречивите им резултати и ограниченото им присъствие в универсалните софтуерни програми. В едно от емпиричните изследвания в дисертационния труд броят на клъстерите е определен на основата на сравняване на резултата от голям брой индекси.

6. Валидизиране на резултатите

Първият критерий за оценка на резултата от клъстеризацията е интерпретативността и полезността на клъстерното решение от гледна точка на целта на изследването, наричано *първична валидизация*. Вторият критерий е доколко резултатът притежава необходимите за клъстерния анализ качества. Именно това е предмет на валидизацията, като последна стъпка от приложението на анализа. Нивото на формализация, което ще се приложи при валидизацията, зависи от характера на данните, размера на базата данни и изследователя. В дисертационния труд е описана валидизация чрез вътрешни и външни тестове (критерии), тестване с външни променливи, повторен анализ, комбиниране на клъстерни решения от различни методи, приложение на анализа към друга извадка от същата съвкупност, Монте Карло изследване, валидизация от експерти. В рамките на изследванията в дисертационния труд са приложени външни тестове, тестове с външни променливи, комбиниране на клъстерни решения (консенсусни решения).

7. Заключение

Клъстерният анализ е свързан с избор: на променливи; на скала на измерване на променливите; на начин на преобразование на променливите

(вкл. стандартизация); на начини на подреждане на случаите в базата данни; на мярка за сходство/разстояние; на метод за клъстеризация; на брой клъстери. Няма друг статистически метод, в който изследователят да е изправен пред толкова много решения и който да няма еднозначен статистически критерий за тяхното оценяване. Това е причината клъстерният анализ да се разглежда като описателна техника, а не анализ, а резултатите от него днес да се определят като „*полезни*“, а не като „*естествени*“ групи от случаи (обекти).

В социологията по-полезно клъстерно решение е това, което води до по-голяма ефективност на последващия анализ - друга описателна или неописателна техника (метод), която да разкрие различно поведение на получените групи и да подпомогне генерирането или потвърждаването на хипотези. Затова приложението на клъстерния анализ трябва да е съпроводено от много тестове, докато се потвърди полезността на избраната стратегия за клъстеризация. Тези тестове трябва да стъпят на аргументиран избор от всички възможни решения. Описанието на методологията на клъстерния анализ в този раздел, от една страна, подпомага определянето на адекватна изходната стратегия при приложението на метода към конкретна изследователска задача, а от друга страна, извежда методологическите проблеми, пред които е изправен изследователят в хода на приложението на анализа.

Трета глава. ПРИЛОЖЕНИЕ НА КЛЪСТЕРЕН АНАЛИЗ ЗА ИЗГРАЖДАНЕ НА СЪСТАВНИ ИНДИКАТОРИ

В главата се извеждат възможни подходи за преодоляване на различни методологически проблеми при приложението на клъстерния анализ за изграждане на съставни индикатори, както и последиците върху резултата от клъстеризацията на методологични решения като избора на метод, използваната мярка за близост, промяната на скалата на измерване на променливите, прилагането на стандартизация, начина на подреждане на случаите в базата данни.

За целта са реализирани две емпирични изследвания върху изграждане на съставни променливи (индикатори) с клъстерен анализ. *Първото изследване* е насочено към „привеждане“ към социологическата практика на формалните подходи към променливите, на основата на които да се извърши клъстеризацията на относително хомогенна съвкупност (малък брой променливи, голям брой случаи), получена чрез случайна извадка. Това изследване стъпва на критична самооценка на приложението на клъстерния анализ в проекта TACTICS[2], в който е реализирано взаимнодопълващо се приложение на клъстерен анализ и множествен анализ на съответствията с цел идентифицирането на специфични конфигурации от отношения между дейци на основата на взаимозависимости между разновидностите на променливи от различни области. *Второто изследване* е насочено към решенията, които трябва да се вземат на други етапи от приложението на анализа, към силно хетерогенна съвкупност (малък брой случаи, голям брой бинарни променливи), получена чрез типологична извадка. То стъпва на проект[3], използващ клъстерния анализ за изграждане на съставна променлива за предмет на дейност на фирми на основата на кодове от класификацията на икономическите дейности. Създаването на съставна променлива, която еднозначно да описва предмета на дейност и разкрива типове индустриални групи от фирми, е първа стъпка на статистическия анализ. На следващ етап се използват мрежови анализ, регресионен анализ, множествен анализ на съответствията и многомерно скалиране за разкриване на структурни отношения и характеристики на получените индустриални типове фирми.

За всяко от изследванията са описани целта, с която се прилага клъстерният анализ, очакванията към резултата от клъстеризация, приложените методологични решения на всеки от етапите на анализа. По-долу са представени резултатите от тези изследвания.

1. Подходи при изграждане на съставни променливи

Първото изследване в дисертационния труд е насочено към извеждането на подходи при приложението на клъстерния анализ за изграждането на типологични съставни променливи в социологията. Тези подходи имат за цел получаването на клъстери с *ясно разграничено съдържание*, което да позволи на изследователя да ги интерпретира и наименува, с *принос на възможно повече от изходните променливи за тяхното дефиниране*, което да осигури минимална загуба на информация при нейното редуциране, с *относително равномерно разпределение на случаите в клъстерите и достатъчен брой случаи в тях,* което да позволи използването им в последващ анализ.

В рамките на това изследване с количествено измерени описателни характеристики, *е потвърдена по-голямата ефективност на Уорд метода с Евклидово разстояние* в сравнение с други 24 варианта (всички йерархични методи с мерки ED, SED и коефициент на корелация, три варианта на *k*-средните и двустъпковия клъстерен анализ в IBM SPSS Statistics) при изграждане на конкретна съставна променлива на основата на бинарни променливи.

След което усилията са насочени към повишаване на ефективността на клъстеризацията в рамките на конкретния метод и мярка – Уорд метода с Евклидово разстояние. В тази връзка са изследвани и обосновани възможни подходи към теоретично определените променливи за основа на изграждането съставна променливи. Ефективността на подходите се оценява чрез повишаване на дискриминиращите възможности на изградената съставна променлива спрямо променливи, неучастващи в клъстеризацията, т.е. тест с външни променливи. За тази цел са използвани обобщаващите характеристики – разграничителни мерки на променливи и

[2] http: //www.isomatic.co.uk/tactics.htm (20.05.2017)

[3] https://www.surrey.ac.uk/sbs/sar/centres/bcned/research/industrial_clusters_in_the_global_information_sector_20042006.htm (20.05.2017)

квантификации на разновидности, в множествения анализ на съответствията приложен към 25 съставни променливи, изградени с клъстерен анализ. Подходите включват:

- *Клъстеризация по множествена променлива след елиминиране на „масови“ разновидности*

Колкото повече разновидности има дадена множествена променлива, толкова повече са трудностите при прилагането на клъстерния анализ с цел формиране на съставна променлива, по която случаите да бъдат еднозначно определени. Подходите зависят както от разпределението на изследваната съвкупност по разновидностите на множествената променлива, така и от връзката (корелацията) между тях.

Когато една или повече *разновидности имат висок относителен дял не само в цялата изследвана съвкупност, но и в подсъвкупностите, дефинирани от другите разновидности, то те няма да имат разграничителна способност при клъстеризирането на случаите, а само биха пречили да се прояви тази на останалите разновидности*. Изключването от клъстеризацията на „масовите“ разновидности води до *увеличаване на обема на хомогенните клъстери,* обособени *по следващите по масовост разновидности*. Така клъстерите стават по-ясно разграничени и по-лесно наименуеми. Изключените променливи ще имат достатъчно висок относителен дял във всеки от получените клъстери, за да участват в определяне на съдържанието им, с което не се губи информация.

- *Клъстеризация по множествена променлива след елиминиране на „редки“ разновидности*

„Редките“ разновидности, т.е. тези с нисък относителен дял, също не биха имали разграничителна способност и не биха инициирали обособяването на самостоятелен клъстер, но за разлика от „масовите“ разновидности те могат да допринесат за намиране на теоретичното основание за групиране, което е предпоставка за адекватно наименуване на клъстера. Затова изследователят трябва да оцени разликите между вариантите „с“ и „без“ редки разновидности.

Изследването установява емпирично, че след изключването на редките разновидности от клъстеризацията, клъстерите имат по-ясни клъстерни центрове и се увеличава броя на клъстероопределящите променливи, т.е. тези които допринасят за дефинирането на клъстерите.

- *Клъстеризация по множествена променлива след обединяване на разновидности*

Различна е ситуацията, когато трябва да се клъстеризира множествена променлива с голям брой разновидности, между които няма „масови“ и „редки“. Ако изследователят има интерес да използва информацията от всички разновидности, то редуцирането на променливите не може да бъде чрез изключване. Подходът зависи от броя и корелацията между разновидностите на множествената променлива. При голям брой разновидности, между които няма корелация, възможен подход е обединяването на разновидности. Така и броят на променливите ще бъде

намален и информацията няма да бъде напълно загубена. Предприемайки подобна стратегия, изследователят е изправен пред задачата да открие теоретично обосновано обединяване, което да доведе до формиране на интерпретативни групи.

- *Клъстеризация по множествена променлива чрез латентни променливи*

Описаният подход с обединяването на разновидности е неефективен при множествена променлива с голям брой разновидности, със значими относителни дялове на всяка от тях и корелация между тях. В този случай групирането на разновидностите може да доведе до голямо препокриване на съдържанието на клъстерите и до наименуването им с етикети, разкриващи какви характеристики не притежават случаите в клъстера, вместо какви притежават.

Подходът за по-ефективна клъстеризация на *множествена променлива с голям брой и (умерено) корелиращи се разновидности е използването на латентни променливи, получени с факторен анализ (анализ на главните компоненти) върху бинарните променливи.* Този подход, наричан „тандем“ клъстерен анализ, нямаше да е изненадващ, ако променливите бяха поне на ординална скала, но прилагането му към бинарни променливи и то в рамките на множествен въпрос (променлива) е нетрадиционно.

- *Клъстеризация по множествена променлива с оригиналните разновидности*

След като изведохме четири подхода за изграждане на съставна променлива на основата на клъстеризация по множествена променлива, всеки от които насочен към редуциране на участващите в анализа разновидности на множествената променлива, сме длъжни да обърнем внимание, че има случаи, в които клъстерите могат да са по-полезни и без редуциране на разновидности, дори и те да са хетерогенни. Въпреки хетерогенността на клъстерите, социологът може да разпознае в тях групи, чиято идентичност просветва зад традиционно наложилите се класификации благодарение на клъстерния анализ.

- *Клъстеризация по количествени променливи*

Клъстерният анализ към интервално скалирани променливи по *разновидности, описващи структурата на даден индикатор*, напр. относителното разпределение на клиентите на фирма по категориите - малки и средни предприятия (МСП), големи фирми, мултинационални компании (МНК), води до получаване на клъстери, чието съдържание е силно повлияно от доминиращата разновидност, т.е. променливата с високи стойности, дори и след прилагането на z-стандартизация. Възможност да се преодолее този недостатък е *преминаването от интервална към ординална скала с възможно най-малко разновидности* (напр. 0=0%; 1=под 50%; 2=над 50%,). Новите ординално скалирани променливи могат да бъдат клъстеризирани като интервални или като бинарни, т.е. пълно множество фиктивни променливи. *Колкото по-слаба е скалата, толкова по- хомогенни*

са клъстерите, но това е за сметка на размера на един или повече хетерогенни клъстери.

- *Клъстеризация по ординално скалирани променливи след стандартизация*

При ординално скалирани променливи, независимо дали те са резултат от трансформация на интервална скала или оригинално са такива, възможен подход е прилагането на *z*-стандартизация. *Потискайки променливите с висока средна стойност, тя позволява да се изявят разграничителните възможности на променливата с ниска средна стойност. В резултат на което участващите в клъстеризацията променливите стават по-равностойни и независими от тяхното разпределение.*

- *Клъстеризация по променливи на слаби скали чрез въвеждане на тегло*

Прилагането на стандартизацията може да се окаже недостатъчно ефективен подход по отношение на *улавянето в един клъстер на случаи притежаващи редки, но важни разновидности.* От примера по-горе това са фирмите имащи за клиенти МНК. Опитът това да се постигне с намаляване на размерността на скалата за променливите със силно асиметрични разпределения, от три на две разновидности за МСП (1= до 50%; 2=над 50%) и МНК (0=0%; 1=над 0%), няма нужния ефект при последователно кодиране на категориите. Решението се оказва *увеличаване на диапазона (интервала) между кодовете чрез въвеждане на тегло за подценената категорията*, т.е. кодирането с 2, вместо с 1 на отговора „над 0% МНК“.

Следователно използването на тегло може да бъде полезно, когато важна характеристика се подценява поради ниския относителен дял на случаите, които я притежават, и невъзможността тя да бъде уловена с помощта на стандартизация.

- *Клъстеризация по качествени променливи след стандартизация*

Противоречивите становища за това дали трябва включените в анализа променливи да се стандартизират и ако да, то за променливи на какви скали се отнася това, е причина стандартизацията да бъде един от тестваните варианти при търсенето на адекватен подход към всички описани по-горе клъстеризации по множествени променливи. Изводът е, че когато традиционната *z-стандартизация е приложена към бинарни променливи в рамките на една множествена променлива*, *тя дава възможност да се изявят като клъстероопределящи и разновидности с нисък относителен дял, но това е за сметка на размитото съдържание на клъстерите и невъзможността да се уловят чисти комбинации с висока интензивност*. Емпиричните изследвания показват, че такива съставни променливи имат по-ниски дискриминиращи възможности спрямо тази получените без стандартизация, но на основата на по-хомогенни групи.

Различна е ситуацията при клъстеризация на основата на фиктивни бинарни променливи по качествени променливи с различна структура, напр. едната с относително равномерно разпределение, другата с едномодално

разпределение. В този случай клъстерното решение се доминира от модата на втората променливата. Ако желаем да направим двете променливи по-равнопоставени*, т.е. разновидности и от двете качествени променливи да имат принос за дефиниране на съставната променлива, то трябва да се приложи традиционната z-стандартизация към фиктивните бинарни променливи.*

- *Клъстеризация по количествена и качествена променлива*

При клъстеризиране на променливи на различни скали се препоръчва прилагане на стандартизация. Изследването доказва, че независимо от начина на стандартизация клъстерните решения с Уорд метода и мярка за разстояние (ED или SED) се повлияват силно от високите стойности на количествената променлива. Заключението е, че *стандартизацията на количествена променлива не е подходящ подход при асиметрично разпределение и голяма вариация в нейните стойности.*

Този недостатък може да се преодолее с трансформиране на интервалната скала на количествената променлива в ординална чрез прилагане на групировка, подход който изведохме като ефективен и при клъстеризация по количествени променливи. Изследването установява, че за да *бъдат по-равнопоставени разновидностите на новата променлива, а съответно и на оригиналната количествена променлива, те трябва да се кодират последователно.* В зависимост от скалата (номинална или ординална) на качествената променлива, двете променливи се анализират като интервални, с фиктивни променливи за номинално скалираната променлива. А в зависимост от разпределението - с или без стандартизация.

- *Клъстеризация по променливи за открити въпроси*

В социологическата практика честотата на разновидностите от класификатора, по който се типологизират отговорите на открит въпрос, често е много ниска, а процентът на неотговорилите или друга филтрираща разновидност е много висок. Това е предпоставка за трудности при изграждане на съставна променлива по открити въпроси. Преодоляването им изисква *типологизирането на изходните променливи да води до възможно най-малък брой разновидности* според ясни и еднозначно критерии.

Обобщение

Заключенията от емпирично изследване върху изграждане на съставни индикатори чрез клъстерен анализ са:

- При индивидуален подход към променливите по отношение на скалата им на измерване и преобразования (стандартизация, факторизация, претегляне), Уорд методът с Евклидово разстояние може да бъде добър инструмент за изграждане на типологични съставни индикатори.
- Клъстерният анализ е много подходящ за трансформиране на множествена променлива в единична (съставна) променлива. Ако между разновидностите на множествената променлива няма корелация, то клъстеризацията може да бъде на основата на оригиналните разновидности, след изключване на разновидности с висок относителен дял („масови“

разновидности), след изключване на разновидности с нисък относителен дял („редки“ разновидности), след обединяване на разновидности. При клъстеризация по множествени променливи с изключването на разновидностите с много висок и много нисък относителен дял се увеличават разграничителните способности на другите разновидности и се получават по-хомогенни клъстери. Обединяването на разновидности е алтернатива, когато останалите подходи не са подходящи и са налице съдържателни аргументи за обединяването на отделни категории на множествената променлива. При голям брой и умерено корелиращи се разновидности множествената променлива може да бъде заменена с латентни променливи. Съдържанието на клъстерите обаче трябва да бъде определено на основата на латентните променливи, а не на изходните за тяхното формиране променливи.

- Начинът за преодоляване на недостатъка на Уорд метода да се повлиява от високите стойности при клъстеризация на основата на интервално скалирани променливи е трансформиране на интервалната скала в ординална с ограничен брой разновидности, кодирани последователно. В зависимост от останалите променливи и броя на разновидностите на новата ординална скала за количествената променлива, тя може да бъде анализирана или като интервална, или с пълно множество от фиктивни променливи. На колкото по-слаба скала се анализират променливите, толкова по-хомогенни и съответно по-лесни за интерпретация са получените клъстери, но това е за сметка на по-голям хетерогенен клъстер/и.
- Стандартизацията на бинарни променливи в рамките на една множествена променлива изявява като клъстероопределящи и разновидности с нисък относителен дял, но това е за сметка на „замърсено“ (размито) съдържание на клъстерите и невъзможността да се уловят чисти комбинации с висока интензивност. Затова пък стандартизацията на фиктивни бинарни променливи, дефинирани по разновидностите на качествена променлива, измерена на номинална скала, изявява разграничителните възможности на всяка от променливите. Стандартизацията на количествени променливи не е подходящ подход при асиметрично разпределение и голяма вариация, тъй като тя преодолява разликата между скалите на променливите, но не и разликата между случаите по дадена променлива. Стандартизацията на ординално скалирани променливи, анализирани като интервални, прави променливите по-равностойни и независими от разпределението им. Така, при изграждане на съставни променливи чрез клъстерен анализ, стандартизацията не е полезно преобразование към количествени и бинарни променливи в рамките на множествена променлива, но е ефективен подход към ординално скалирани променливи, анализирани като интервални и номинално скалирани променливи, анализирани с пълно множество от фиктивни променливи.
- Ако стандартизацията не се справя с улавянето на случаи, притежаващи важна, но рядка (с нисък относителен дял) характеристика - разновидност по ординална или бинарна скала, то възможен подход е

въвеждането на тегло, увеличаващо диапазона между съответните кодове и предаващо по-голяма тежест на подценената разновидност. Обратно за да бъдат по-равностойни разновидностите на количествена променлива измерена на ординална скала те трябва да бъдат кодирани последователно, а не със средите на интервалите, които ще зададат тегла на съответните отговори.

Ограничения: Изведените подходи повишават ефективността на приложение на Уорд метода с Евклидово разстояние, с цел изграждане на съставни индикатори. Адекватността им към други методи и мерки е въпрос на допълнителни изследвания.

2. Разстоянието в изходната матрица при Уорд метода

Второто емпирично изследване върху построяването на съставна променлива на основата на голям брой бинарни променливи и малък брой случаи има за задача, от една страна, да потвърди приложимостта на Уорд метода към бинарни променливи, а от друга, да установи последиците от замяната на квадратно Евклидово разстояние (SED), налагано от строги статистически критерии при Уорд метода, с Евклидовото разстояние (ED). Изследвана е и връзката между *разстоянието* (ED или SED), *начинът на подреждане* на случаите (случайно; по свързана с резултата съдържателна променлива; оптимално подреждане, определено от софтуер на основата на голям брой пермутации), *включените в анализа променливи* (всички или редуцирано подмножество чрез изключване на тези с нисък относителен дял).

Установените зависимости от сравнението с коригирания Ранд индекс (adjusted Rand index) на 12 варианта на клъстеризация и 3 консенсусни решения (дървета) в резултата на тяхно комбиниране са:

- *Изборът на мярка – ED или SED, оказва по-силно влияние на клъстерното решение от начина на подреждане на случаите, но влиянието на реда при силно хетерогенна съвкупност е достатъчно силно, за да бъде пренебрегнато.*
- *Подреждането по съдържателна променлива, която се очаква да бъде свързана с резултата от клъстеризацията, води до по-стабилни клъстери, в сравнение с подреждане по независима променлива*. Това заключение може да бъде използвано по два начина. Когато съставната променлива - резултат от клъстерния анализ, се очаква да отразява даден аспект на описваното с нея явление, трябва да се търси вътрешна или външна за анализа променлива, по която да се подредят случаите в базата данни. Същевременно, когато се прави оценка за степента на стабилност на решението, трябва да се прилага подреждане по независима променлива.
- *При SED има по-голяма съгласуваност между пермо-клъстерното решение, получено след оптимално подреждане на случаите в базата данни на основата на множество пермутации, и другите варианти, отколкото при ED*.

- При ED съгласуваността е повлияна силно от това дали е между варианти с или без пермо-клъстери. От една страна, *малките промени (възходящо или низходящо) в подреждане по случайна или свързана с резултата променлива влияят много малко на решението, т.е. решенията с ED са по-стабилни.* От друга, голямото отклоняване на тези варианти от този с пермо-подреждане показва, че *стабилността може да бъде подвеждаща и намирането на оптималното подреждане при по-малките разлики (разстояния) е по-трудно.*
- Консенсусното решение при SED е средно аритметично на решенията, които съгласува, докато при ED то е повлияно от мнозинството. Това още веднъж потвърждава по-голямата съгласуваност на решенията с ED и обуславя *по-голяма необходимост на консенсусни решения при SED.*
- Редуцирането на променливите, основа на анализа, чрез отстраняване на тези с нисък относителен дял (в случая тези с честота 1), се отразява по-малко на решенията с SED, отколкото на тези с ED. Следователно *наличие на редки разновидности би оказало по-малко влияние върху съдържанието и стабилността на клъстерите при SED, отколкото при ED.*
- Решенията с ED са по-съгласувани с общото консенсусно решение върху резултатите от приложението на двете мерки – ED и SED. *Следователно, има по-голяма вероятност резултати, получени с ED, да бъдат по-близко до „средното“ за двете мерки решение.*

Обобщено, при клъстеризиране на основата на голям брой бинарни променливи, много от които с нисък относителен дял:

- Уорд методът с SED формира по-хетерогенни клъстери с по-голям дисбаланс по отношение на обема. Решенията са по-зависими от начина на подреждане, но повечето от тях е еднакво близко до средното (консенсусното) решение. Наличието на редки разновидности оказва по-слабо влияние върху резултата в сравнение с това при ED.
- Уорд методът с ED формира по-хомогенни клъстери, за сметка на ограничен брой (1-2) хетерогенни клъстера. Обемът на клъстерите варира по-малко и по-рядко се срещат малки клъстери, съдържащи уникални случаи. Подреждането оказва по-слабо влияе на клъстеризацията, но и по-трудно се попада на оптималното подреждане. Резултатът от клъстеризацията се влияе по-силно от наличието на редки характеристики, в сравнение с SED.

Направените изводи допринасят за осветляването на два некоментирани и изследвани проблема не само за Уорд метода, взети поотделно и взаимосвързани: *влиянието на мярката за близост* - Евклидово или квадратно Евклидово разстояние, с която се прилага методът *и редът* на случаите в базата данни.

3. Съдържателната оценка на резултата от клъстерния анализ

Резултат от приложението на клъстерния анализ за изграждане на съставна променлива върху силно хетерогенна съвкупност е извеждането на насоки за съдържателна оценка на резултата от анализа. За разлика от формалната оценка, която е на основата на числов показател, съдържателната оценка изисква проследяване на характеристиките на клъстерите от гледна точка на най-голяма адекватност на клъстерното решение по отношение на предмета на изследването.

Изходната точка на съдържателната оценка са формирането на „клъстерните центрове“ и дефинирането на понятието „клъстероопределяща“ характеристика в контекста на конкретната клъстеризация:

- *Клъстерните центрове* са средните стойности за променливите, основа на клъстеризацията. Когато променливите са бинарни, клъстерните центрове се задават с процентното разпределение на случаите от даден клъстер по променливите.
- *Клъстероопределяща е характеристика*, която се притежава от мнозинството от случаите в клъстера. Това изисква при бинарни променливи да се фиксира долната граница на процента от случаи в клъстера, които трябва да притежават дадена характеристика, за да бъде тя определяща за съдържанието на клъстера. Този процент може да варира от 100% до 50%, но точната му граница е въпрос на данни и избор на изследователя. Личната ни практика е да използваме минимум 65%.

Следващата стъпка в съдържателната оценка на резултата от клъстеризацията е да се проследи съдържанието на клъстерите с цел:

- *Да се оцени броят на случаите в клъстерите* от гледна точка на изискванията на анализа, който ще се прилага с резултата от клъстеризацията. В типичния случай – изграждане на съставни променливи по данни от представително социологическо изследване, клъстерът трябва да съдържа поне 50 случая, за може да бъде разглеждан самостоятелно и сравняван с други клъстери. Затова, в случай на клъстер с малък обем, трябва да се оцени възможността той да бъде присъединен към друг клъстер, независимо от нивото, на което те се обединяват, т.е. да се прескочи ниво/а само за тях.
- *Да се оцени влиянието на броя на клъстероопределящите характеристики върху дефинирането (наименуването) на конкретен клъстер*. Колкото по-малко са общите характеристики на случаите в клъстера, толкова по-лесно той се дефинира. Докато комбинацията от повече характеристики изисква да се определи общото между тях, за да се даде адекватно име на клъстера.
- *Да се оцени разсейването в клъстерите*, т.е. дали неклъстероопределящите характеристики допринасят за съдържателното дефиниране на клъстера или го „замърсяват“. Във втория случай, ако тези характеристики са важни, да се провери дали те не са клъстероопределящи в друг вариант.

- *Да се оцени загубата на информация*, т.е. има ли важни променливи или техни разновидности с достатъчно висок относителен дял, които не са клъстероопределящи в никой от формираните клъстери. Това може да е резултат от ниска разграничителна способност на променливата, но може да е повлияното от конкретно клъстерно решение.
- *Да се оцени препокриването на информация*, т.е. наличие на променливи, които са клъстероопределящи в повече от един клъстер. Ако съдържателната граница между такива клъстери е незначителна, то или трябва да се търси друго решение, или те да бъдат обединени, прескачайки ниво на клъстеризация.
- *Да се оценят хетерогенните клъстери* – клъстери без клъстероопределящи характеристики. Изследователят трябва да реши дали е необходимо тези клъстери да бъдат разбивани на по-ниско ниво, т.е. на подклъстери или ще бъдат запазени. Ако се разбият, може да се вземе решение някои от тези клъстери да се присъединят към даден първоначално получен хомогенен клъстер.
- *Да се оценят алтернативни клъстеризации* в резултат на различни решения. Това изисква да се установят стабилните клъстери в различните варианти. За нестабилните да се проследи начинът им на промяна – дали от тях се отделя малка част, т.нар. мигриращи случаи; дали се разцепват на нови самостоятелни клъстери; дали те са нови клъстерни лидери, към които се присъединяват случаи от други клъстери.

ЗАКЛЮЧЕНИЕ

Клъстерният анализ е метод за многомерна класификация на случаи, чиято групова принадлежност не е предварително известна. Продуктът от тази класификация в социологията най-често е типология. За методологията и методиката на приложение на клъстерния анализ при построяване на типология, е от значение да се разграничат две нейни форми и съответно два продукта от анализа - *типологичен съставен индикатор* и *емпирична типология.* В основата на това разграничение стоят познавателната същност, степента на обобщение и специфичните изисквания към резултата от класификацията. Двата типа класификации в социологията, както и ролята им за изучаване на социалните явления чрез съвместно използване на няколко многомерни статистически метода, се дискутират в първа глава от дисертационния труд.

Направеният обзор на специализираната методологическа литература за клъстерен анализ в социалните науки и разработките, свързани с приложението на метода в социологията, потвърждават изследователската теза на дисертационния труд за незадоволително представена и слабо развита методология и методика на приложение на анализа в социологията у нас. В тази насока дисертационният труд се опитва да запълни този дефицит по отношение на изграждане на типологични съставни индикатори чрез извеждане и обосноваване на методологически решения, които могат да повишат ефективността на приложението на клъстерния анализ в този случай.

Изследванията в дисертационния труд имат за цел да осветлят кога и кои от решенията, пред които е изправен изследователят в хода на анализа, могат да повишат ефективността на резултата от него и съответно да осигурят възможност за преход от класификация към типология. Всяко подобно изследване трябва да стъпи на задълбочено проучване на методологията на клъстерния анализ, подкрепена с резултати от сравнителни изследвания и възможности за компютърна реализация. В тази насока втора глава от дисертационния труд представя резултата от проучването на методологията на класическите методи за клъстерен анализ. То позволява да се аргументира изборът на едно пред друго решение в хода на приложението на анализа за изграждане на съставни индикатори, както и да се изведат решения, чийто ефект към определен тип данни се нуждаят от допълнителни изследвания.

Систематизацията на научната литература по проблемите на клъстерния анализ, както и изследванията в дисертационния труд показват, че *не може да има универсален подход при приложението на клъстерния анализ за изграждане на съставни индикатори, но има методологически решения към конкретни типове клъстеризации, които позволяват да се повиши ефективността на анализа.* Първото от тези решения е изборът на метод за клъстеризация. Въпреки голямото разнообразие от методи, свойствата им в резултат на заложения в тях алгоритъм, правят в повечето случаи реалния избор много ограничен. А изведените на основата на сравнителните изследвания фактори, които влияят на ефективността на методите с доказано превъзходство пред останалите - Уорд и средното (междугруповото) свързване, правят избора в социологията почти без алтернативен. Това важи особено при изграждане на съставни индикатори, където хомогенността и съответно интерпретативността, близкият размер и достатъчният брой случаи в клъстерите, които се получават при Уорд метода, го правят по-подходящ за целите на клъстеризацията. Към получените от него резултати може да се приложи итеративен метод за клъстеризация (*k*-средните), но успехът на подобна стратегия при съставни променливи не е гарантиран. *Изследванията в дисертационния труд потвърждават хипотезата за ефективността на Уорд метода при изграждане на съставни променливи (индикатори), но само когато клъстеризацията е на основата на променливи измерени на слаби скали – ординални, номинални и бинарни (дихотомни).* По отношение на мярката за близост, с която да се прилага Уорд методът, в дисертационния труд се обосновава, че използването на Евклидово разстояние вместо квадратно Евклидово разстояние, налагано от строги статистически ограничения, позволява да се преодолее силното въздействие на големите различия (разстояния), в резултат на което се получават по-хомогенни и интерпретативни клъстери. Това е особено важно за клъстеризация по бинарни променливи, при които алтернативният вариант за тази цел - Манхатово разстояние, съвпада с квадратно Евклидово разстояние. Изследването във втората част на трета глава установява, че изборът между Евклидово и квадратно Евклидово разстояние при изграждането на

съставни променливи трябва да се прави на основата на размерността и хомогенността на анализираните данните. Това изследване извежда и ефекта от начина на подреждане на клъстеризираните случаи при двете разстояния. Подреждането по съдържателна променлива, свързана с изгражданата съставна променлива, може да доведе до по-стабилни и съответно до по-полезни резултати от клъстеризацията.

В случай, че резултатът от Уорд метода не е задоволителен във формално и/или съдържателно отношение, то причината за това не трябва да се търси в приложения метод и мярка, а в недостатъчно ефективни решения по отношение на останалите стъпки в методологията на анализа.

Изследването в първа част от трета глава на дисертационния труд *доказва хипотезата, че подходът към променливите е от решаващо значение за ефективността на клъстеризацията и съответно качеството на съставната променлива (индикатор)*, оценено чрез дискриминиращите ѝ възможности спрямо други променливи. В тази насока в дисертационния труд са изведени и обосновани, в зависимост от структурата на данните, пет подхода към теоретично определените като основа на клъстеризацията бинарни променливи в рамките на една множествена променлива. По отношение на променливи на интервална скала - ефективният подход е трансформацията на скалата в ординална. Особено внимание в изследването е отделено на ефекта от стандартизацията, за който в методологическата литература за клъстерен анализ няма единно становище. Резултатите от изследването показват, че ефективността на стандартизацията при изграждане на съставни променливи е зависима от скалата на измерване на променливите. Тя може да е адекватен подход към ординално и номинално скалирани променливи, анализирани съответно като интервални и фиктивни променливи, но не и към интервални и бинарни променливи. Изведен е подход – използване на тегло, да се изявят разграничителните способности на подценени променливи или техни разновидности при изграждане на съставни индикатори.

Въпреки че предметът на дисертационния труд е изграждане на съставни индикатори чрез клъстерен анализ, резултатите от изследванията в него могат да послужат като основа на изследвания, насочени към изграждане на емпирична типология чрез клъстерен анализ в социологията.

Ограничения: Изследванията в дисертационния труд са емпирични. Ефективността на изведените подходи за изграждане на съставни индикатори към конкретни данни, както и направените изводи за стабилността на клъстерните решения в зависимост от мярката и подреждането на случаите, трябва да бъдат проверявани. Същото се отнася и за симулационните изследвания.

ПРИНОСИ

Приноси с научен характер:

1. Обосновано е изграждането на два типа класификации в социологията - на основата на типологичен съставен индикатор и емпирична типология. Обосновани са разликите в двата подхода и особеностите на изграждането им.
2. Обоснована и емпирично доказана е по-голямата ефективност на Уорд метода с Евклидово разстояние в сравнение с другите йерархични методи, а в определени случаи и от итеративните методи за клъстеризация, за изграждане на съставни индикатори на основата на променливи на ординални и номинални скали в социологията.
3. Изведени са подходи за изграждане на съставен индикатор (променлива) чрез клъстерен анализ, които позволяват по-ефективно редуциране на информация в ЕСИ и използването ѝ в синтезиран вид при търсене на връзки и зависимости между променливите. Тъй като някои от тези подходи се прилагат традиционно към количествени, но не и към качествени променливи, то те са „приведени“ към социологическата практика.
4. Изследвани са дискусионни методологически решения в клъстерния анализ и са представени емпирични доказателства, че:

4.1. Използването на Евклидово разстояние към бинарни променливи, вместо квадратно Евклидово разстояние, налагано от строги статистически ограничения при Уорд метода, е по-подходящ подход за изграждане на съставни променливи. То позволява да се получат по-хомогенни, по-чувствителни към редки характеристики, с по-малки отклонения в обема и по-стабилни при промяна на реда на случаите клъстери. Това ограничава както дискусията за това, с какво разстояние е правомерно да се прилага Уорд метода, така и дискусията за правомерността на използване на мерки за разстояние към бинарни променливи.

4.2. Прилагането на традиционната z-стандартизация при Уорд метода към променливи на ординални и номинални скали, анализирани като интервални и фиктивни, прави по-равнопоставени изходните променливи при дефиниране на съставната променлива. Подобен ефект не се постига при интервално скалирани променливи. Представени са доказателства, че стандартизацията не е ефективен подход и при бинарни променливи в рамките на една множествена променлива, където изявявайки като клъстероопределящи и разновидности с нисък относителен дял, тя не дава възможността да се уловят чисти комбинации с висока интензивност. С това се ограничава дискусията по отношение на това кога трябва да се прилага стандартизация при клъстеризация с цел изграждане на съставни променливи.

Приноси с научно-приложен характер:

5. За първи път у нас са систематизирани възможните решения в методологията на класическите методи за клъстерен анализ, последиците от тях върху резултата от анализа, причините да се предпочете едно пред друго решение, начините за компютърната им реализация. Това осигурява възможност за адекватна стратегия на приложение на анализа за решаване на конкретна изследователска задача както в социологията, така и в другите социални науки у нас.
6. Формулирани са насоки от девет стъпки за съдържателна оценка на резултата от клъстеризацията, които да осигурят по-ефективно приложение на клъстерния анализ и подпомогнат прехода от класификация към типология в социологията.

Списък на публикациите по темата на дисертационния труд

1. Кескинова, Д., Чалъков, И. Приложение на клъстерния анализ при изследване на отношения между социални дейци (Типични проблеми и подходи). *Социологически проблеми*, 3-4, 2001, 36-59. ISSN 0324-1572.
2. Todeva, E., David, K., Keskinova, D. The Complementarity of Clusters Analysis and Network Analysis to Map the Structure of Competencies in Global Information Sector. In*:* Serdult, U., Taube, V. (eds.) *Applications of Social Network Analysis* – ASNA 2005, WVB, Germany, 2008, 281-298. ISBN 978-3-86573-374-0.
3. Кескинова, Д. Клъстерният анализ в социологията. *Социологически проблеми*, 1-2, 2017 (под-печат). ISSN 0324-1572.