МЕТОДОЛОГИЯ ЗА ИЗВЪРШВАНЕ НА ОЦЕНКИ ОТ ИЗВАДКОВИ ИЗСЛЕДВАНИЯ НА НИВО ОБЩИНИ И НАСЕЛЕНИ МЕСТА ПОСРЕДСТВОМ КЛАСТЕРЕН И СТРУКТУРЕН АНАЛИЗ

Валерия Ангелова, Ивайло Гавазки***

I. Въведение

В последните години значително нараства търсенето на данни на все по-ниско териториално ниво - общини, населени места и дори квартали. Източник на голяма част от тези данни са репрезентативни изследвания. Методите за оценки на липсващи данни за малките териториални единици (Small Area Estimation Methods) намират приложение при производството на данни, чийто първичен източник са извадкови изследвания. По правило производството на данни на ниските териториални нива среща сериозни ограничителни условия. Едно от тях се отнася до стойността на разходите за осигуряване на тези данни. На колкото по-ниско териториално ниво се извършва статистическото изследване, толкова по-скъпо е то. Друг проблем възниква от необходимостта за осигуряване на статистическата конфиденциалност на данните.

В световен мащаб тенденцията е все повече да се ползват извадкови изследвания вместо изчерпателни. По-голямата част от извадковите изследвания, провеждани от НСИ, са конструирани така, че да осигуряват представителна и с достатъчна точност информация предимно до ниво области. Обикновено се използва двустепенна извадка, стратифицирана по административно-териториални области (NUTS3) и местоживеене (град, село), в резултат на което се получават 56 страти. Доскоро се считаше, че от такива проучвания не могат да се произведат представителни данни за по-малки единици. Според някои изследователи една от възможностите за постигане на надеждност при прилагане на извадкови методи за събиране на данни е увеличаването на обема на извадките, както и подсилване на извадките в определени региони, което обаче повишава себестойността на получения статистически продукт и не реша-

* Старши експерт, отдел „Регионална статистика и ГИС”, НСИ; e-mail: vangelova@nsi.bg.
** Главен експерт, отдел „Демографска статистика”, НСИ; e-mail: igavazki@abv.bg.

ва проблема с производството на данни за статистически/териториални единици, които не са попаднали в извадката. Стратификацията според административно-териториалното деление и според статута на населените места не отчита хетерогенността на териториалните единици. В статистическата практика у нас все още не се прилагат подходи, които да отчитат хетерогенността на населените места и по този начин да се разширяват възможностите за получаване на оценки при липсващи данни.

В европейската практика също все още няма достатъчно опит и емпирични знания за методите за оценки на липсващи данни за малки териториални единици (Small Area Estimation (SAE). През 2003 г. в рамките на проект на Световната банка „Проследяване, оценка и разработване на политики по въпросите на бедността" е извършен анализ на бедността на ниво община по данни на многоцелевото наблюдение на домакинствата от екип с участие на експерти от НСИ и Института за икономически изследвания при БАН. Това е първият опит в България за оценки на липсващи данни за малки териториални единици (SAE) на ниво община. Използваният от авторите метод за SAE, известен като „Картографиране на бедността", представлява разработването на регресионен модел по данни от извадковото изследване и „налагането" на този модел върху данни от преброяването на населението през 2001 година. Така се „присвояват стойности" по целевите променливи на всяка единица от преброяването (unit level model). В регресионния модел са заложени помощни променливи от Преброяването, които авторите на изследването наричат променливи - потенциални фактори. В този случай е необходимо помощните променливи добре да корелират с целевите променливи. Въз основа на получените оценки на измерителите на бедността в 262 общини е приложен кластерният анализ за тяхната типологизация. По същество това изследване използва кластеризацията само за обобщаване и представяне на резултатите от извършените оценки, а не за повишаване на прецизността и надеждността на самите оценки. Авторите правят извод, че „За условията на България се оказва, че чрез картографирането могат да се получат надеждни оценки на бедността на областно и общинско равнище, но не и на равнище населени места." [1]. В статията предлагаме друг подход за решение на този проблем.

[1] България: предизвикателствата на бедността 2003, НСИ, с. 93.

Много изследователи (Cameron, 1998; Trewin, 1999; Chambers, 2005) разработват и подобряват статистико-математическия аналитичен апарат, който е необходим при оценяването на липсващи данни за малки териториални единици, като всички акцентират върху това, че успешното извършване на оценките зависи от правилния избор на модел, съобразен с типа данни (бройни, бинарни или непрекъснати) и от разумно подбрания размер на малките териториални единици, за които ще се извършват оценките. Те са единодушни, че броят на единиците, които не попадат в извадката, трябва да е минимален. Това е сериозно ограничително условие, което следва да бъде преодоляно.

Страна извън ЕС с традиции в оценките на липсващи данни от извадкови изследвания за малки териториални единици е Австралия. През 2006 г. Австралийското статистическо бюро публикува „Ръководство по оценки на данни за малки териториални единици". Трудът е с висока теоретична и познавателна стойност, но не дава отговор на въпросите, поставени по-горе. Възприемаме идеята на авторите на ръководството, че най-важните помощни променливи, които следва да се заредят в модела, са демографските характеристики. Допуска се, че повъзрастовата плодовитост, смъртност и миграции са синтетични измерители на състоянието на много сфери на обществения живот.

Целта на настоящото изследване е да се предложи методология, подходяща за производство на оценки на данни от извадкови изследвания на ниво община и населено място. За постигане на поставената цел бяха изпълнени следните задачи: проучване на съществуващата научна литература, посветена на оценките за малки териториални единици, формулиране на основните проблеми при извършването на оценки на ниво община и населено място, избор на помощни променливи, дефиниране на хомогенни групи от населени места на базата на структурен анализ, описание на алгоритъма и анализ на практическите ползи от прилагането на предлаганата методология.

II. Практическа необходимост от нов подход към оценката на липсващи данни от извадкови изследвания

Необходимостта от изготвяне на собствена методология за производство на оценки за малки териториални единици възникна с оглед на

ангажимента на НСИ за изпълнение на дейностите по проект на Европейската комисия за събиране на данни за избрани европейски градове и техните агломерационни ареали (Urban Audit data collection 2012). За настоящата фаза на проекта за референтния период 2010 - 2012 г. ще бъдат събрани и изчислени данни за повече от 150 променливи за 18 населени места с население над 50 000 души и 58 общини, формиращи техните агломерационни ареали. Градовете, обект на изследване в рамките на Urban Audit, са определени от Евростат, а общините са дефинирани НСИ въз основа на данни от Преброяване 2011 за ежедневните трудови миграции. В ГИС среда са избрани общините, които дават повече от 15% от работната си сила на съответното централно селище и формират компактна територия, което е в съответствие с предписанията на Евростат. За 33 променливи, отнасящи се до заетостта, икономическата активност на населението, образованието и доходите на домакинствата няма друг източник освен репрезентативни изследвания - SILC (Social Inclusion and Living Conditions), LFS (Labour Force Survey), HBS (Households Budget Survey). Изключение за някои променливи прави годината на преброяването. Това налага избор на подходящ математико-статистически инструментариум за извършване на оценки на данните по целевите променливи на ниво община и населено място. Известните досега подходи не удовлетворяват нуждата от осигуряване на качествен статистически продукт на ниво населено място и не са способни да отчетат особеностите в териториалното разпределение на населението и развитието на селищната мрежа в България.

III. Алгоритъм за производство на данни за малки териториални единици с помощта на кластеризация на базата на структурен анализ

Идеята за решение на разгледаните проблеми се базира на хипотезата, че между административните области (NUTS2), както и в рамките на самите области, съществува хетерогенност по отношение на демографското, социалното и икономическото развитие на населените места[2].

[2] Разглежданата идея е представена с доклад за участие в конкурс за млади статистици на Международната асоциация за официална статистика (IAOS) през януари 2013 година.

Подходът се състои в кластеризиране на населените места в България по няколко помощни променливи в четири тематични направления: демография, икономическа активност, образование и икономика от Преброяването на населението 2011 и текущи изчерпателни изследвания, което ще позволи извършване на оценки на ниво общини и населени места в няколко стъпки:

1. Избор на помощни променливи, чиято връзка с целевите променливи е статистически значима. Помощните променливи бяха подбрани така, че да притежават добра „прогностична" способност. Подходящи променливи за кластеризацията са: възрастова структура, раждаемост, смъртност, имиграции, емиграции, брой на заетите лица (на възраст над 15 години, разпределени по петгодишни възрастови интервали), брой на заетите лица по 21 сектора (КИД - 2008), степен на завършено образование и нетни приходи от продажби на глава от населението. С най-голяма важност са демографските данни, които имат добри качества на спомагателна информация. Те са особено подходящи, когато има силни колебания в броя на населението и демографския състав на малките териториални единици. В България, където гъстотата на населението по общини варира от 2.4 души на кв. км в Трекляно до 3 315.8 души на кв. км в Пловдив, това е често срещан проблем.

2. Измерване на разстоянията между структурите във възрастовия профил на населението (в случаите, когато е наличен) за основните му характеристики, изброени по-горе, чрез формулата:

$$\cos\alpha = \frac{\sum_{i=1}^{n} p_{i1}p_{i2}}{\sqrt{\sum_{i=1}^{n} p_{i1}^2 \sum_{i=1}^{n} p_{i2}^2}},$$

където:

$\{p_{i1}\}_{i=1}^{n}$ е структурата по отношение на изследвания признак в конкретно населено място;

$\{p_{i2}\}_{i=1}^{n}$ - структура на същия признак спрямо средното за страната;

p_{i1} и p_{i2} - съответните относителни дялове на двете структури;

i - поредният структурен интервал;

n - броят на относителните дялове;

α - ъгловото разстояние между два вектора, които са точки от нормираното Евклидово пространство и представляват сравняваните структури;

$\cos\alpha$ - нормиран измерител, функционално зависим от Евклидовото разстояние между двете структури (Христов, 2000).

Тази обща постановка на въпроса в нашия случай беше конкретизирана посредством заместване на средните стойности за страната, изчислени по горната формула с еталонна структура. Такава реално не се наблюдава при демографските и социалните процеси, тя играе ролята на отправна точка, на начало на координатната система. Използването на еталонната структура увеличава аналитичните възможности на модела, като позволява да се правят сравнения между страните - членки на ЕС. При този подход в изчисленията беше използван следният конкретен измерител, отразяващ Евклидовото разстояние между структурата $\{p_{ij}\}_{i=1}^{n} = \{p_{ij}, i = 1,...,n\}$ на j-тото населено място и хипотетичната еталонна (равномерна) структура $\{{}_e p_i\}_{i=1}^{n} = \left\{ {}_e p_i = \frac{1}{n}, i = 1,...,n \right\}$:

$$\cos\alpha_j = \frac{\sum_{i=1}^{n} p_{ij}\,{}_e p_i}{\sqrt{\sum_{i=1}^{n} p_{ij}^2 \sum_{i=1}^{n} {}_e p_i^2}} = \frac{\sum_{i=1}^{n} p_{ij}\frac{1}{n}}{\sqrt{\sum_{i=1}^{n} p_{ij}^2 \sum_{i=1}^{n}\left(\frac{1}{n}\right)^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n} p_{ij}}{\sqrt{n\frac{1}{n^2}\sum_{i=1}^{n} p_{ij}^2}} = \frac{1}{n\sqrt{n\frac{1}{n^2}\sum_{i=1}^{n} p_{ij}^2}} = \frac{1}{\sqrt{n\sum_{i=1}^{n} p_{ij}^2}},$$

където:

$$\sum_{i=1}^{n} p_{ij} = 1;$$

p_{ij} са относителните дялове на единиците в отделните подсъвкупности i, ($i=1, ..., n$) спрямо общия брой на всички единици в съвкупността (j-тото населено място);

n е броят на всички относителни дялове.

3. Дефиниране на хомогенни групи от населени места с възможно най-малка вътрешногрупова и най-голяма междугрупова дисперсия.

Вътрешногруповото и междугруповото разсейване се определя по формулите:

$$\sigma^2_{\text{int}ra\text{-}group} = E\left(\xi_i - \mu\right)^2 \approx \frac{1}{n-k}\sum_{i=1}^{n}\left(\xi_i - \mu\right)^2 \approx \frac{1}{n-k}\sum_{i=1}^{n}\left(\cos\alpha_i - \frac{1}{n-k}\sum_{j=1}^{n}\cos\alpha_j\right)^2 = \hat{\sigma}^2_{\text{int}ra\text{-}group}$$

и

$$\sigma^2_{between\text{-}groups} = E\left(\xi_i - \mu\right)^2 \approx \frac{1}{k-1}\sum_{i=1}^{n}\left(\xi_i - \mu\right)^2 \approx \frac{1}{k-1}\sum_{i=1}^{n}\left(\cos\alpha_i - \frac{1}{k-1}\sum_{j=1}^{n}\cos\alpha_j\right)^2 = \hat{\sigma}^2_{between\text{-}groups}$$

където:

σ^2 е теоретичната - „истинска" стойност на дисперсията;

E - символ за математическо очакване;

μ - теоретичната - „истинска" стойност на математическото очакване;

ξ_i - случайни величини;

$\cos\alpha_i$ - реализации на случайните величини измерители на разстоянията между структурите;

α_i - ъгли между съответните структури;

i, j индикират интервалите;

n - брой на групите;

k -1 и n-k съответстват на степените на свобода.

Ако $\hat{F}_{empirical} = \dfrac{\hat{\sigma}^2_{\text{int}ra\text{-}group}}{\hat{\sigma}^2_{between\text{-}groups}} > F_{theoretical}$, се счита, че различието между средните на групите е статистически значимо, където теоретичната стойност $F_{theoretical}$ е взета от таблицата на F разпределението при ниво на значимост $\alpha = 0.05$ и степени на свобода ($k - 1$) и ($n - k$).

Беше избран подход за кластеризация на всички населени места, а не само на тези, включени в проекта Urban Audit, за да се осигури възможност за производство на оценки за малки териториални единици на следващата фаза от проекта, когато списъкът с населените места, които са обект на изследване, вероятно ще бъде разширен. Освен това, работейки с всички населени места, имаме възможност да изследваме генералната съвкупност.

4. Залагане в модела на непретеглени, агрегирани данни от извадковите изследвания за населените места, попаднали в конкретната извадка. Осигурената посредством кластеризацията хомогенност позво-

лява прилагането на теоретичната дефиниция на Лаплас за пресмятане на вероятности за сбъдване на определено събитие при реализирането на един статистически експеримент, в която се изисква в знаменателя на тази известна формула да стои общият брой равновъзможни събития. Тази дефиниция е приложима в случаите на крайни пространства от елементарни събития. В конкретния случай в знаменателя е поставена сумата от всички наблюдавани единици в кластера, а числителят представлява сумата от всички наблюдавани единици в кластера, които удовлетворяват изискванията на търсената целева променлива. Вероятностната мярка според дефиницията на Колмогоров (която е обобщение на известната дефиниция на Лаплас и е фундаментална, защото е основа на аксиоматичното изграждане на теорията на вероятностите, благодарение на което тази теория е с огромна научна стойност) има свойствата позитивност, нормираност и адитивност, от които следва, че така изчислената вероятност е една и съща за всички населени места в кластера. Тази вероятност се умножава по броя на единиците за всяко отделно населено място в избрания кластер и в резултат се получава търсената оценка за съответната целева променлива.

Логическа схема на алгоритъма

Кластеризация $\Rightarrow$ Хомогенност $\Rightarrow$ Равновъзможност $\Rightarrow$
Изчисляване на вероятностната мярка $\Rightarrow$ Използване на нейните свойства $\Rightarrow$
Приложимост към отделните населени места в кластера

5. Извършване на оценка, валидиране и представяне на резултатите за малките териториални единици, обект на изследване в проекта Urban Audit, посредством тематични карти. Картографирането на резултатите осигурява визуална и количествена оценка на грешката и позволява на анализаторите и потребителите да видят евентуални неочаквани пространствени модели или аномалии в териториалното разпределение на стойностите на оценките.

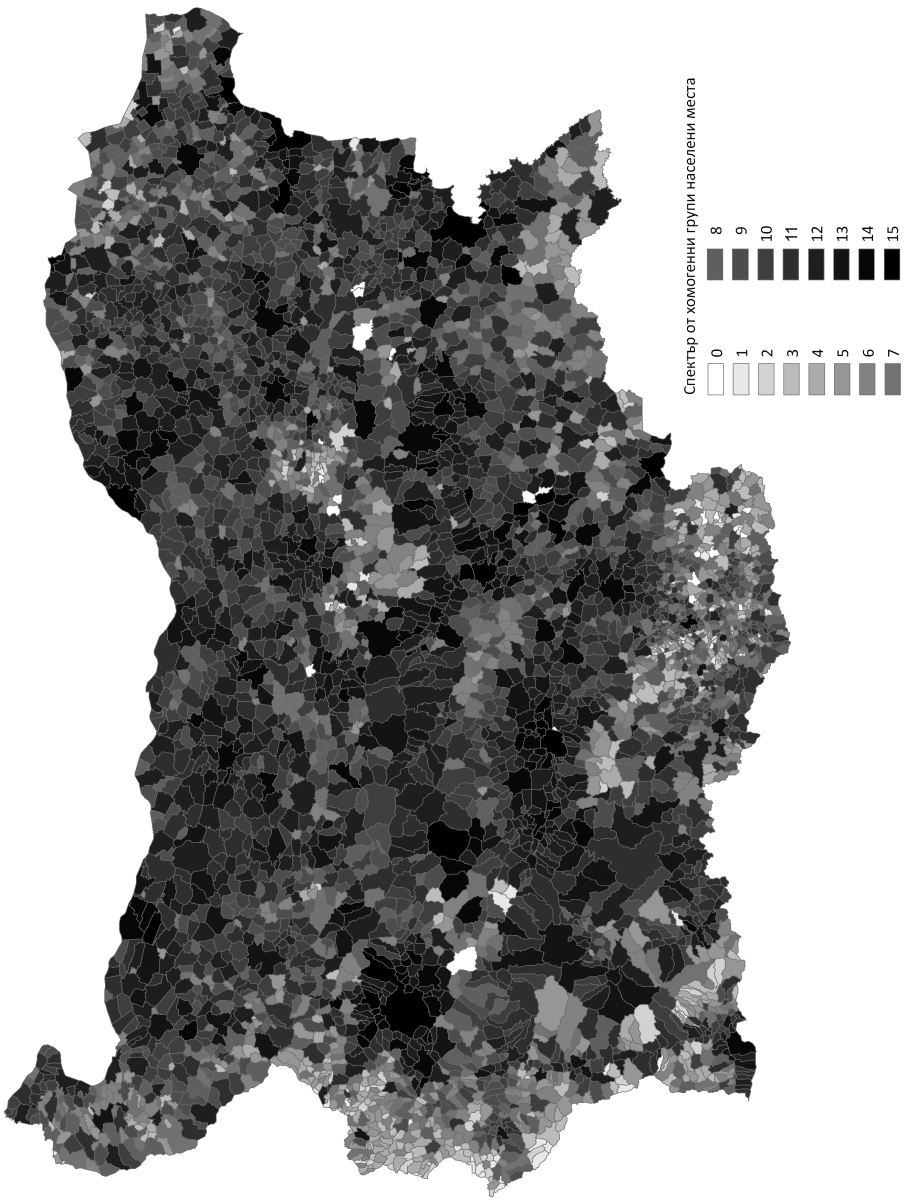
III. Проблеми при реализацията на метода и анализ на практическите ползи от приложението му

В научната литература въпросът за оптималния брой на кластерите е широко дискутиран. Приема се, че броят им може да е най-малко

3 и максимално 15. Когато обаче се работи с голям брой териториални единици (5 302 населени места към критичния момент на Преброяване 2011) и се използват значителен брой помощни променливи, кластерите стават по-чувствителни, спектърът на наблюденията се разширява и е възможно да се получат повече от 15 хомогенни групи.

След проведена серия от експерименти беше установено, че по всяка отделна променлива населените места следва да формират по 3 кластера с оптимална хомогенност. Така получените кластери по всичките 9 избрани помощни променливи бяха „кръстосани". Поради разнообразието на използваната информация кластерите се разделят на помалки групи населени места, в които протичащите демографски и социално-икономически процеси имат близки по стойност параметри, т.е. имат една и съща природа. Теоретичният максимален брой на кластерите е 3^9, което значително надвишава броя на населените места, както и реалните практически нужди, но показва, че проведената по този начин кластеризация има достатъчна разделителна способност. Целта беше да се постигне такъв мащаб, при който във всеки кластер да са налице достатъчен брой наблюдения от извадковите изследвания. Разработеният от нас метод дава възможност за производството на такъв голям брой кластери от порядъка на 3^9, въпреки че такъв брой в настоящото изследване няма практическа стойност.

В резултат на нашата работа бяха получени 16 хомогенни групи от населени места. Кластерите са представени на фиг. 1. Картографирани са само населените места, които притежават собствени землища.



Фиг. 1. Хетерогенност на населените места в България по отношение на някои демографски, социални и икономически характеристики

Прилагането на този подход увеличава точността на оценките в следните направления:

- Редуцира се влиянието на стратификацията на извадката върху пространствения обхват на данните при отчитане на географския фактор. Във всеки кластер ние изследваме населени места, считайки ги за практически идентични, индиферентни по отношение на тяхната пространствена локализация в различните административни области. Приемаме населените места за идентични, въпреки че в рамките на кластерите също има хетерогенност. Тя обаче е от друг порядък. Посредством приложението на кластерния анализ ние преодоляваме разстоянията и ограниченията, наложени от физическите и времевите измерения на географското пространство, оперирайки в друго, топологично пространство, изградено от материални обекти (сами по себе си изследваните структури представляват точки, принадлежащи на n - мерното Евклидово пространство). Ако приемем, че тези хомогенни групи от населени места представляват страти на извадката, то в такъв случай, вследствие на хомогенността вътре в кластерите, би било без значение къде точно в стратите са разположени гнездата. Това е една възможност да се намали обемът на извадката и въпреки това да се повиши надеждността на получаваните оценки за общините и населените места. В условията на хетерогенност възниква систематична грешка и колкото повече са наблюденията, толкова повече тя се мултиплицира. Много изследователи се стремят към по-голям обем на извадката, без да си дават сметка за ефекта върху качеството на данните. Прилагането на разработената от нас методология е възможност за намаляване на систематичната грешка. Нашата препоръка е броят на наблюденията в извадковите изследвания леко да превишава техния необходим минимален теоретичен брой. Разбира се, това само по себе си е немаловажна изследователска задача.
- Проблемите, породени от считания досега за малък обем на извадката и териториалните единици, за които няма наблюдения, са решени. Предлаганият подход повишава качеството на данните за населените места, за които има наблюдения и осигурява информация за населените места, които не са попаднали в извадката.

- Така получените групи от населени места са хомогенни по отношение на помощните променливи, които от своя страна корелират с променливите, на които се правят оценки. Това дава възможност за производство на достатъчно надеждни оценки за малки териториални единици при ниски разходи;
- Дава възможност различните извадкови изследвания да заработят синхронизирано в една система.

IV. Резултати от оценката на LFS[3] променливата „Относителен дял на рано напусналите образователната система"

Методологията за оценки за малки териториални единици беше апробирана с данни от LFS по специфичния показател „Относителен дял на рано напусналите образователната система лица на възраст от 18 до 24 години от населението на същата възраст" за 2011 година. Спряхме се на този показател, тъй като той е един от измерителите на напредъка към постигане на националните цели на стратегията „Европа 2020".

За целта беше изградена база данни, съдържаща входящата информация. За производството на тази променлива кластерите от 0 до 10 бяха обединени с цел работа в мащаб, при който има достатъчно наблюдения от извадковото изследване (в 4 393 населени места в кластери от 0 до 7 няма попаднали гнезда в извадката на LFS). Тъй като LFS е тримесечно изследване, бяха изчислени средногодишните вероятности за реализация на конкретното събитие (отпадане от образователната система) по кластери. Вероятностите на някои рязко отклоняващи се случаи, които попадат извън границите на доверителния интервал, бяха изгладени с помощта на полиномна крива от втора степен. Вероятностите бяха умножени по средногодишното незакръглено население на възраст 18 - 24 години по населени места. Резултатът е средногодишна оценка на целевата променлива за всички населени места в България.

Проверката за кохерентност на оценките на ниво LAU1 и LAU2, получени посредством предлаганата методология, и данните на ниво страна, получени в рамките на LFS, показа, че те са напълно съгласува-

[3] Заетост и безработица - годишни данни 2011, НСИ, 2012.

ни. Нашите резултати за 2011 година, агрегирани на национално ниво, и данните от LFS са следните:

Пол	LFS	Оценки по предлагания метод
Общо	**11.8%**	**11.8%**
Мъже	11.2%	11.1%
Жени	12.6%	12.5%

Показателен за демографската ситуация, в която се намира страната, е фактът че в 1 118 населени места изобщо няма население на референтната възраст 18 - 24 години. В 5 административни области, 25 общини и 271 населени места е постигната националната цел, заложена в стратегия „Европа 2020" - 11% дял на преждевременно напусналите образователната система. В 364 населени места, от които 66 са градове, една четвърт от населението на възраст 18 - 24 години напуска преждевременно образователната система и остава с образование по-ниско от средното (ISCED - 1, 2 или 3C). Сред пространствените единици, обект на изследване от Urban Audit, се открояват общините Кричим и Куклен, с дял на рано напусналите образователната система над 25%, и градовете Видин и Велико Търново - над 18%. Тези отклонения в стойностите за посочените общини могат да се обяснят с някои специфични особености в демографската структура на населението, географското разположение, развитието на социалната и производствената инфраструктура и други. Тяхното установяване е предмет на допълнителен анализ.

V. Заключение

Тъй като през последните години търсенето на статистическа информация за по-малки административно-териториални единици нарасна, планираме да използваме тази технология не само за нуждите на проекта Urban Audit. Нашата основна цел беше да разработим методология за извършване на оценки, която да има методичен характер, да е стабилна и да дава решение на широка гама от изследователски проблеми, свързани с производството и разпространението на данни на ниво общини и населени места. В хода на работата беше установено, че

тази методология също така подпомага и изготвянето на демографски прогнози на ниво населено място, като дава решение на проблеми, възникващи поради недостатъчния брой демографски събития в малките населени места.

Предлаганата от нас методология за оценка на липсващи данни за малки териториални единици е оригинална, неприлагана досега в този си вид. Тя дава възможност данни, чийто първичен източник са извадкови изследвания, да се произвеждат на ниво населено място, което в България се прави за първи път. В бъдеще работата по предлаганата методология ще продължи в посока на подобряване на нейния методичен характер, оптимизиране на алгоритмите и разширяване на обхвата на нейната приложимост.

ЦИТИРАНА ЛИТЕРАТУРА:

НСИ (2005). България: предизвикателствата на бедността 2003. Анализ на многоцелевото наблюдение на домакинствата.

НСИ (2012). Заетост и безработица - годишни данни 2011.

Христов, Е. (2000). Влияния на промените на повъзрастовата смъртност и възрастовата структура на населението върху изменението на брутната смъртност (методологични решения и емпиричен анализ). Население, Списание на института по демография, с. 22 - 47.

A Guide to Small Area Estimations (2006). Australian Bureau of Statistics.

Cameron, A. C., P. K. Trivedi (1998). Regression Analysis of Count Data, Cambridge: Cambridge University Press.

Chambers, R. (2005). Calibrated Weighting for Small Area Estimation, Southampton Statistical Sciences Research Institute, Methodology Working Paper, M05/04.

Trewin, D. (1999). Small Area Statistics Conference, Survey Statistician, 41, p. 8 - 9.

МЕТОДОЛОГИЯ ИЗГОТОВЛЕНИЯ ОЦЕНОК НА БАЗЕ ВЫБОРОЧНЫХ ОБСЛЕДОВАНИЙ НА УРОВНЯХ ОБЩИН И НАСЕЛЕННЫХ ПУНКТОВ С ИСПОЛЗОВАНИЕМ КЛАСТЕРНОГО И СТРУКТУРНОГО АНАЛИЗОВ

Валерия Ангелова, Ивайло Гавазки***

РЕЗЮМЕ Методы для оценки данных о небольших территориальных единицах применяются для преодоления небольшого размера выборок и изготовления достаточно надежных данных, в сравнении с результатами, полученными при использовании директных оценок с самого выборочного обследования. Авторы применяют кластерный анализ на основе структурного анализа для определения гомогенных групп населенных пунктов по некоторым демографическим, социальным и экономическим характеристикам. Работа в одной гомогенной среде позволяет произвести данные о малых территориальных единицах посредством использования вероятностей реализации того или иного события в каждом из кластеров и их наложения на среднегодовую численность населения в населенных пунктах.

* Старший эксперт, отдел „Региональная статистика и ГИС”, НСИ; e-mail: vangelova@nsi.bg.
** Главный эксперт, отдел „Демографическая статистика”, НСИ; e-mail: igavazki@abv.bg.

METHODOLOGY FOR SMALL AREA ESTIMATIONS OF SAMPLE SURVEYS' DATA AT LAU1 AND LAU2 LEVEL USING CLASTER AND STRUCTURAL ANALYSES

Valeria Angelova, Ivaylo Gavazki***

SUMMARY Small Area Estimation methods are used to overcome the problems caused by the small samples sizes at producing the small area estimates and production of enaugh reliable data in comparison with direct survey estimates obtained from the sample in each small area. The authors apply the cluster analysis on the base of structural analysis in order to define homogenous groups of settlements according to some demographic, social and economic characteristics. Working in a homogeneous environment allows the production of small area estimates using the probability for realization of a particular event in each cluster and its matching with the average annual population of the territorial units.

* Senior Expert, Regional Statistics and GIS Department, NSI; e-mail: vangelova@nsi.bg.
** Chief Expert, Demographic Statistics Department, NSI; e-mail: igavazki@abv.bg.