МЕТОДИЧНІ ЗАСАДИ ОБ'ЄДНАННЯ ДАНИХ З РІЗНИХ ДЖЕРЕЛ ДЛЯ АНАЛІЗУ ДЕМОГРАФІЧНИХ І СОЦІАЛЬНО-ЕКОНОМІЧНИХ ПРОЦЕСІВ

В.Г.Саріогло,
кандидат технічних наук,
старший науковий співробітник
Г.І.Терещенко,
молодший науковий співробітник,
Інститут демографії та соціальних досліджень НАН України

Останніми роками науковці, фахівці державних, громадських та комерційних структур, які працюють в галузі дослідження та моніторингу демографічних і соціально-економічних процесів, все чіткіше усвідомлюють необхідність використання інформації, отриманої з різних джерел. Цьому сприяють, зокрема, наступні тенденції у зміні принципів розробки соціальної політики [1]:

- зміщення акцентів при оцінці соціального розвитку та ефективності соціальних програм на регіональний та локальний (місцевий) рівень;
- зростання актуальності оцінки умов життя окремих верств населення – населення у стані бідності або злиденності, етнічних груп тощо;
- підвищення ролі багатовимірних, комплексних індикаторів соціального розвитку, таких, наприклад, як індекс людського розвитку;
- застосування методів моделювання при розробці заходів соціальної політики;
- дедалі ширше використання методів прогнозування і прогнозних оцінок;
- розвиток інформаційної бази, нові можливості застосування сучасних інформаційних технологій та ін.

Зростання попиту з боку користувачів призвело до того, що в галузі розробки методологічних підходів, методів, методик і процедур комплексного використання соціальної інформації нині відбуваються революційні зміни. Особливо це стосується проблем інтеграції даних на мікрорівні, про що свідчить й постійне зростання кількості публікацій з відповідних питань.

УДК 314.02

Значну увагу фахівці приділяють питанням створення спеціалізованих (демографічних, соціальних) інтегрованих інформаційних систем корпоративного та, навіть, національного рівнів [2]. Важливим елементом таких систем є засоби, що забезпечують можливість об'єднання даних з різних джерел.

Окремі методики об'єднання даних впроваджуються в технологію обробки результатів вибіркових обстежень населення, що дозволяє підвищити надійність оцінок показників та узгодити їх з наявною зовнішньою інформацією [2].

У даній роботі розглядаються основні підходи до об'єднання інформації з різних джерел для підвищення ефективності вимірювання соціально-економічних і демографічних показників. Головна увага приділяється підходам до об'єднання даних на субнаціональному рівні. Методологічні та методичні аспекти вказаних підходів в Україні та пострадянських країнах недостатньо висвітлені в спеціальній літературі, хоча окремі методики та процедури вже застосовуються на практиці, зокрема при організації державних вибіркових обстежень населення.

Слід зазначити, що найбільш поширеним і опрацьованим підходом до забезпечення можливості оцінки показників на регіональному рівні є об'єднання агрегованих даних. Так, для оцінювання демографічних показників на регіональному рівні у міжперепис-ний період ще з 50-х років минулого століття застосовуються спеціалізовані симптоматичні розрахункові методики, в основі яких лежать так звані традиційні демографічні методи [3]. Ці методи призначені для оцінювання показників по малих територях на базі об'єднання даних переписів з поточними адміністративними даними. Останні використовуються як симптоматичні показники для малих територій при аналізі відповідних процесів: актуальні адміністративні дані по певній території про кількість народжених, кількість померлих, характеристики міграційних процесів та ін. [4]. До найбільш відомих демографічних методик належать такі: методика життєвих відношень (Vital Rates Method), методика композиційного оцінювання, методика житлових одиниць тощо.

Спеціальні підходи до вимірювання показників зайнятості та безробіття на базі комплексного використання інформації з різних джерел з початку 60-х років XX століття застосовуються в США. Наприклад, у поточному вибірковому обстеженні населення (ПОН), що проводиться в США головним чином для щомісячного вимірювання рівнів безробіття і зайнятості населення, для отримання оцінок за окремими штатами та адміністративними округами застосовуються методи моделювання [5, 6]. При цьому будуються два види моделей: для рівня зайнятості та для рівня безробіття.

У моделі оцінки безробіття поряд з даними ПОН враховується така додаткова інформація, як дані з реєстру страхування по безробіттю, чисельність зайнятих, що визначається за платіжними документами, результати обстеження зайнятості, яке проводиться в несільськогосподарських галузях в ході поточного обстеження зайнятості, тощо. Перелічені дані використовуються в агрегованому вигляді. При моделюванні на рівні штату застосовуються статистичні процедури, що дають можливість коригувати оцінки з урахуванням наявних контрольних величин. Наприклад, місячні оцінки зайнятості та безробіття на рівні штатів коригуються таким чином, щоб їх річні середні показники дорівнювали відповідним показникам ПОН.

У США застосовуються також методи оцінки зайнятості та безробіття для субштатного рівня – так званий метод довідника, який до 1973 року був єдиним офіційним підходом для отримання оцінок зайнятості та безробіття на рівні штатів і на місцевому рівні [5, 6]. При оцінюванні безробіття на локальному рівні цей метод полягає у формуванні ряду оціночних "будівельних блоків", в яких категорії безробітних класифікуються за

їхнім попереднім статусом (рис.1). При цьому для отримання інформації по відповідних блоках використовуються наявні на місцевому рівні джерела даних, що забезпечує високу економічну ефективність підходу.

Оцінка безробіття в поточному місяці за методом довідника являє собою сукупну оцінку всіх трьох складових будівельних блоків. "Охоплена" категорія складається з таких груп безробітних: тих, хто в теперішній час отримує допомогу за програмою страхування по безробіттю; тих, хто не має права на допомогу. Оцінка чисельності першої групи безробітних не є проблемою, оскільки ці дані отримують з відповідних реєстрів.

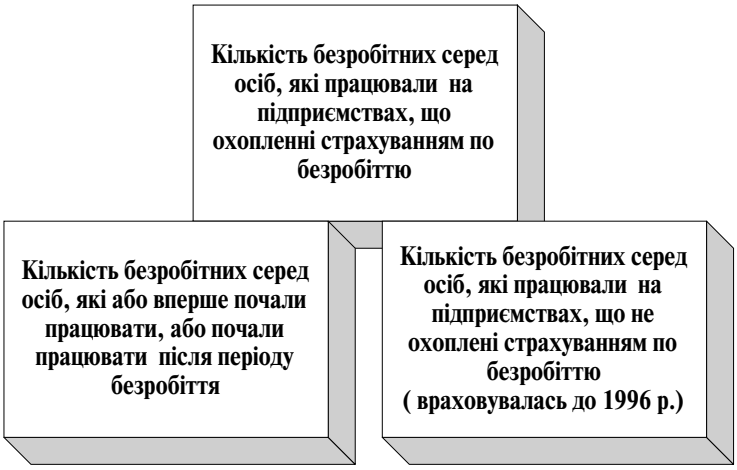

Рис. 1. **"Будівельні блоки" при оцінюванні чисельності безробітних за методом довідника**

Зазначимо, що до осіб, які в останній час працювали на підприємствах, що не охоплені страхуванням по безробіттю (рис.1), відносять хатніх робітниць, безоплатно працюючих робітників сімейних підприємств, осіб, які займаються індивідуальною трудовою діяльністю, тих, що працюють в некомерційних організаціях, місцевих органах самоврядування та деяких сільськогосподарських галузях. Серед економічно активного населення кількість осіб цієї категорії мала тенденцію до зменшення, насамперед, за рахунок розвитку системи страхування на випадок безробіття. Тому, починаючи з 1996 року, вона не враховується при оцінках чисельності безробітних на місцевому рівні.

Оцінки кількості безробітних осіб, які втратили право на отримання допомоги, та тих, хто не має права на її отримання, базуються на оцінках фактичного числа безробітних в теперішній час плюс оцінка чисельності тих, кого слід додати до цих оцінок з попередніх періодів. Наприклад, оцінки за методом довідника кількості безробітних серед тих, хто вперше почав працювати (і не має права на допомогу), є функцією ряду параметрів, а саме: місяця року; поточної кількості кваліфікованих осіб серед безробітних і поточної чисельності кваліфікованої робочої сили; частки молодих людей у чисельності населення працездатного віку та оцінок відповідних часток економічно активного населення у попередні періоди. Оцінка на місцевому рівні середньої чисельності безробітних осіб серед тих, хто вперше почав працювати в поточному місяці, визначається за формулою [6, с. 28]:

$$UT=A(X+E)+BX,$$

де UT – середня за поточний період чисельність безробітних серед тих, хто вперше почав працювати;

X – середня чисельність безробітних за поточний період;

E – середня чисельність зайнятих за поточний період;

A і B – синтетичні коефіцієнти, що враховують історичні (за попередні періоди) оцінки часток економічно активного населення, сезонні коливання безробіття та ін.

Оцінка загальної чисельності зайнятих за методом довідника базується на даних, які отримують з кількох джерел. Головним джерелом інформації є обстеження підприємств, що проводиться за програмою поточного обстеження зайнятості на федеральному та місцевому рівнях. Важливим джерелом даних є обстеження зайнятості на підприємствах, що проводяться самими штатами. Використовуються також деякі інші дані. Метою опрацювання усіх цих даних є отримання оцінок загальної кількості найманих працівників, не пов'язаних з сільськогосподарським виробництвом, які працюють на підприємствах певного району. Ці оцінки зайнятості за місцем роботи узгоджуються з оцінками за місцем проживання. Розраховуються коефіцієнти коригування для кількох категорій зайнятих на основі співвідношень, встановлених за результатами останнього перепису населення. Екстраполяцією результатів перепису оцінюється, зокрема, чисельність працівників, що належать до сільськогосподарських робітників; працюючих не за наймом; безоплатно працюючих на сімейних підприємствах; працюючих вдома та ін. Всі перелічені складові, доповнені даними про погодинну оплату та заробітну плату, становлять оцінку показника зайнятості за методом довідника.

Таким чином, з наведеного зрозуміло, що ступінь використання наявної інформації при оцінюванні показників зайнятості та безробіття в США є дуже високим, особливо при оцінюванні показників на рівні штатів та адміністративних округів.

Аналогічні підходи в США застосовуються при оцінюванні показників бідності. Подібна методика використовується також у Канаді. Окремі підходи реалізовані у Франції, Великій Британії, Швеції та в інших країнах.

Впровадження аналогічних підходів в соціальну статистику України є наразі дуже актуальним.

Розглянуті раніше підходи стосуються, головним чином, агрегованих показників для різних рівнів агрегації даних. Ці методи розвиваються і стають все більш досконалими.

На особливу увагу заслуговують методи інтеграції даних на мікрорівні, тобто на рівні осіб або домогосподарств. Ці методи є найбільш актуальними для фахівців, оскільки дають змогу отримувати узгоджені масиви мікроданих [1]. При інтеграції даних з різних джерел застосовують такі основні методи:

- приєднання даних (data merging);
- зв'язування записів (record linkage);
- статистичне злиття, з'єднання даних (statistical matching; data fusion).

Перші два підходи застосовуються за умов, коли з різних джерел отримують дані по одних і тих самих одиницях або по однакових ознаках. Наприклад, у вибірковому обстеженні економічної активності населення, що проводилося Держкомстатом України за схемою ротації з 1999 по 2003 роки, у вибірці щоквартально змінювалось 25% домогосподарств (тобто, четверта частина домогосподарств вибувала з обстеження, і стільки ж

нових домогосподарств входило до вибірки) [7]. Квартальні оцінки показників розраховувалися на основі результатів опитування приблизно 30 тисяч відібраних домогосподарств. Для побудови річних оцінок показників в єдиний масив зливалися результати опитування за чотири квартали. З урахуванням того, що кожного кварталу анкети опитування були ідентичні, схема процедури приєднання даних є досить простою (рис. 2).

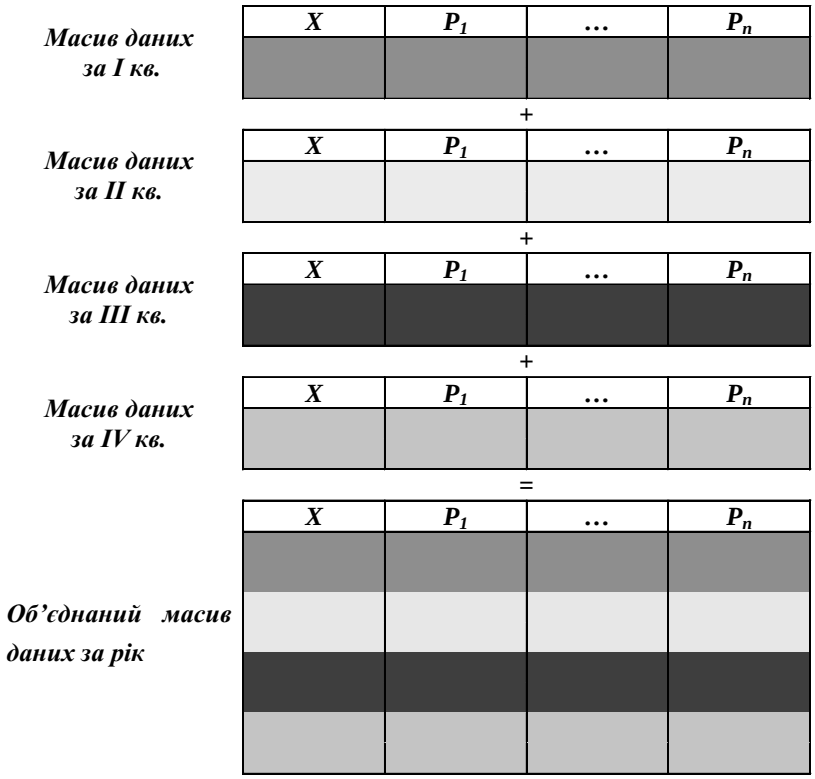

Рис. 2. **Схема об'єднання квартальних даних обстеження економічної активності в єдиний річний масив методом приєднання даних**

Звернемо увагу на те, що йдеться про одиниці однакового рівня – домогосподарства і про ідентичний набір ознак. Процедура приєднання даних реалізується багатьма стандартними статистичними пакетами програм і, зокрема, програмою SPSS.

Процедура зв'язування записів застосовується, коли по одних і тих самих одиницях отримують дані з різних джерел, тобто до певного масиву даних необхідно приєднати додаткові змінні (рис. 3). Прикладом застосування цього методу може бути щоквартальне вибіркове обстеження Держкомстату України умов життя домогосподарств, в якому дедалі більше застосовуються так звані модульні обстеження. Вони проводяться за окремою тематичною анкетою і охоплюють, як правило, всі або певну частину (підвибірку) домогосподарств, що були обстежені з питань умов життя. Для можливості приєднання

даних з умов життя (наприклад, даних щодо складу, доходів або витрат домогосподарств) до даних модульного обстеження необхідно передбачити спеціальні ознаки – ключі, за якими вдасться потім зв'язати записи за домогосподарствами.

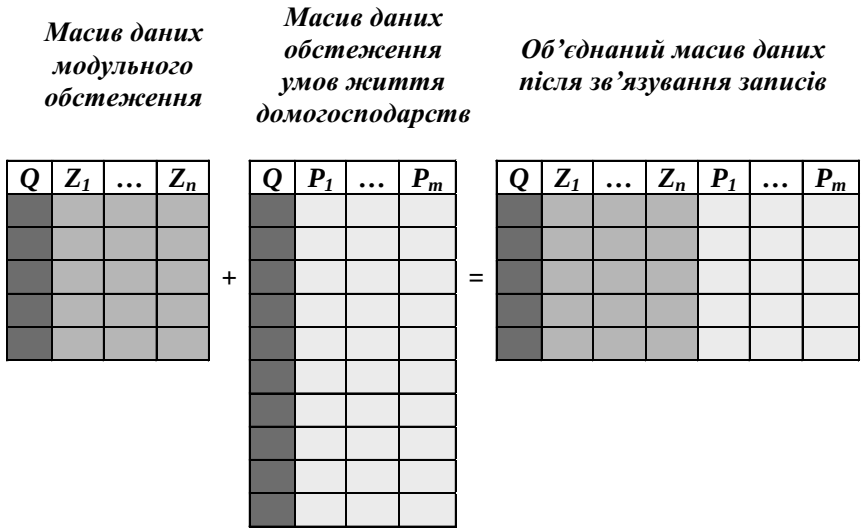

Рис. 3. **Схема зв'язування записів модульного обстеження з даними обстеження умов життя домогосподарств**

Процедура зв'язування записів також реалізується багатьма стандартними статистичними пакетами програм. Два розглянуті підходи широко застосовують при обробці даних вибіркових обстежень у відповідних підрозділах Держкомстату України, науковці, соціологи та статистики при дослідженнях.

Найбільш методологічно і технологічно складним є третій підхід, який дозволяє об'єднувати дані з різних джерел за різними блоками обстежених одиниць та різними блоками ознак (рис. 4). При цьому не вимагається наявність ознак-ключів. Таке об'єднання даних спирається на аналіз та врахування статистичних властивостей наборів даних. Застосування методики злиття даних дає можливість отримувати єдині масиви даних на основі різних обстежень, таких, наприклад, як переписи населення та вибіркові обстеження населення (домогосподарств).

Слід зазначити, що результати переписів і вибіркових обстежень фактично доповнюють одне одного. Переписи є джерелом даних по всіх одиницях спостереження, але по дуже обмеженому колу "фундаментальних" ознак. Спеціалізовані обстеження дають набагато більш детальні дані, але по певній частині (вибірці) одиниць. Лише на підставі комплексного аналізу цих даних з'являється можливість ефективного моніторингу заходів соціальної політики. При цьому дані переписів дають змогу встановити основні характеристики одиниць спостереження та їх розміщення, а результати обстежень – визначити взаємозв'язки та взаємодію ознак [5].

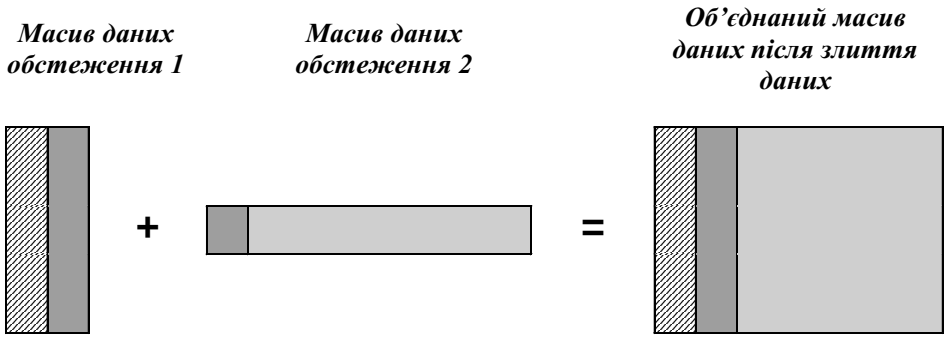

Рис. 4. **Схема статистичного злиття даних двох різних обстежень**

Методи статистичного злиття даних дають можливість оцінювати соціальні показники на основі більших за обсягом масивів інформації, ніж масиви, отримані за даними конкретних обстежень. Так, злиття даних обстежень умов життя домогосподарств і економічної активності населення дозволило би оцінювати показники, що вимірюються в обох обстеженнях (наприклад, окремі показниками зайнятості), на значно більших масивах даних – щоквартально об'єднані масиви включали би результати обстеження майже 35 тис. домогосподарств замість близько 25 тис. для обстеження економічної активності та близько 10 тисяч для обстеження умов життя домогосподарств. Саме для реалізації такої можливості зараз Держкомстат України проводить роботи з гармонізації програм цих обстежень.

Можливість застосування методології злиття даних передбачає виконання певних правил підготовки даних [8], а саме:

- гармонізацію статистичних одиниць;
- гармонізацію періоду статистичних обстежень;
- порівнянність охоплення населення;
- гармонізацію ознак;
- гармонізацію класифікацій;
- контроль пропущених значень;
- контроль походження змінних;
- контроль змінних на послідовність.

Гармонізація статистичних одиниць передбачає необхідність з'ясування, чи всі дані з різних джерел можна віднести до одиниць одного рівня. Наприклад, у вибіркових обстеженнях часто обстежують домогосподарства, а переписи та реєстри населення містять інформацію по окремих особах, тому треба передбачити можливість переходу до статистичних одиниць одного рівня.

Гармонізація періоду статистичних обстежень полягає в узгодженні періоду (моменту часу) отримання даних.

Аналіз охоплення населення дає змогу з'ясувати, чи однакові групи населення об'єднані, бо дані перепису населення містять інформацію про все населення, а дані, скажімо, об-

стеження економічної активності населення містять лише інформацію про працездатне населення, тобто населення у віці 15–70 років.

Гармонізація змінних. В усіх джерелах інформації всі змінні, що об'єднані, мають бути однаково визначені.

Гармонізація класифікацій. В усіх джерелах інформації всі змінні, що об'єднані, повинні бути однаково класифіковані.

Перевірка на пропущені значення. Дана процедура дозволяє з'ясувати, чи не містять змінні пропущені значення.

Походження змінних. Внаслідок об'єднання можуть з'явитися деякі нові змінні.

Перевірка змінних на послідовність дає можливість з'ясувати, чи зберегли змінні потрібну послідовність в об'єднаному масиві.

Оскільки перший масив, зображений на рис. 4 (масив даних обстеження 1), використовується для отримання невідомої частини другого масиву, перший масив називають донором, а другий – реципієнтом. Завдання полягає фактично у заповненні частини масиву - реципієнта за допомогою спеціальної процедури відновлення відсутніх даних – процедури імпутації [9]. Найбільш поширеними методами імпутації є заповнення відсутньої частини масиву: середніми з наявних значень; методом «cold deck» імпутації (заповнення без додбору); методом «hot deck» імпутації (з добором); методом «найближчого сусіда»; заповнення на базі статистичних моделей; багаторазове заповнення, або багаторазова імпутація. Слід зауважити, що кожна методика імпутації має свої переваги і недоліки. Для їх ефективного застосування в конкретній ситуації необхідно виконати спеціальні дослідження. Як правило, при об'єднані даних застосовують кілька методик отримання відсутньої інформації.

Яскравим прикладом комплексного використання даних переписів та вибіркових обстежень населення на основі третього підходу до об'єднання даних є методика так званої "картографії" бідності – спеціальна методика оцінювання показників бідності для малих територій, що забезпечує можливість візуалізації та детального аналізу просторового розподілу бідних верств населення для певної країни та її регіонів [10]. Корисність цього підходу для розробників соціальної політики та дослідників особливо висока, коли є можливість аналізу бідності для найнижчих рівнів агрегації даних – окремих населених пунктів або районів великих міст.

Головний методологічний принцип картографії бідності полягає в об'єднанні даних вибіркових обстежень, в яких вимірюються показники для оцінки бідності, і даних перепису населення (або даних реєстру населення). Враховуючи, що обстеження не можуть забезпечити надійних оцінок бідності на місцевому рівні, а переписи не надають інформації для оцінки бідності, об'єднання даних – фактично єдиний шлях до можливості аналізу бідності на локальному рівні. Процедура об'єднання даних забезпечує визначення (імпутацію) показників бідності (рівня доходів або споживання) для домогосподарств, які охоплені переписом населення. Для цього застосовуються, як правило, методи статистичного моделювання. Моделі будуються на даних вибіркового обстеження. При цьому зрозуміло, що факторні (екзогенні) змінні повинні вимірюватись і у вибірковому обстеженні, і в переписі та бути гармонізовані, що вже відзначалося раніше.

За результатами виконаних досліджень доцільно зробити **такі висновки**. Для належного вимірювання й аналізу демографічних і соціально-економічних процесів на регіональному та локальному рівнях необхідне ефективне використання інформації, отрима-

ної з різних джерел. Сучасні методологічні підходи до комплексного використання інформації на різних рівнях агрегації даних та на мікрорівні застосовуються в багатьох країнах світу при оцінці ряду важливих показників. Окремі розробки в цьому напрямі є і в Україні. Для забезпечення можливості застосування методів об'єднання інформації при розрахунках демографічних і соціально-економічних показників на регіональному та локальному рівнях для офіційного використання необхідно виконати певний обсяг досліджень і провести відповідну роботу з користувачами. При плануванні масштабних державних обстежень населення України необхідно передбачати можливість комплексного використання їх результатів на регіональному та локальному рівнях.

Джерела

1. *Bakker B.F.M., Al P.G.* Re-engineering social statistics by micro-integration of different sources: an introduction//Netherlands Official Statistics. – 2000. – Vol. 15.

2. *Материалы* встречи группы экспертов по инновационным методикам для переписей населения и крупномасштабных демографических обследований. – Гаага: NIDI, 22–26 апреля 1996 г.

3. *Rao J.N.K.* Small Area Estimation. – John Wiley&Sons, 2003.

4. *Рон Превост.* Интеграция статистических данных (результатов обследований) с (административными) данными регистров / Экономический и Социальный Совет. – ООН, 2000. – Е.

5. *Local* Area Unemployment Statistics. Program Manual. – U.S. Department of Labor, Bureau of Labor Statistics, August 2001.

6. *Методологическое* руководство бюро статистики труда США. – BLS Handbook of methods. Russian Language, April 1997.

7. *Статистичний* збірник. Економічна активність населення України 2002. – К.: Держкомстат України, 2003.

8. *Van der Laan, P.* Integration administrative registers and household surveys // Netherlands Official Statistics. – 2000. – Vol. 15.

9. *Саріогло В.Г.* Особливості застосування процедур імпутації при обробці даних вибіркових обстежень домогосподарств // Проблеми статистики. – К., 2001. – Вип. 3.

10. *Harold Alderman, Miriam Babta, Gabriel Demombynes, Nthabiseng Makhath, Berk*. How low can you go? Combining census and survey data for mapping poverty in South Africa. – World Bank. – November 2001.