УДК 657

Пугачова М.В., к.т.н., с.н.с.,
директор Науково-проектного інституту
статистичних технологій
НТК статистичних досліджень

МЕТОДОЛОГІЧНІ ЗАСАДИ ІМПУТАЦІЇ ДАНИХ КОН'ЮНКТУРНИХ ОБСТЕЖЕНЬ ПІДПРИЄМСТВ

Проаналізовано і класифіковано причини пропусків даних в обстеженнях, розроблено і перевірено на реальних даних цілий комплекс алгоритмів імпутації даних кон'юнктурних обстежень для всіх типів показників.

Вступ. Для більшості статистичних обстежень існує проблема пропусків даних. Причинами тому можуть служити як невідповіді респондентів в цілому на запитання анкети (тобто явна чи неявна відмова від участі в обстеженні), або небажання відповідати на окремі запитання анкети. Ще однією, і можливо самою прикрою, причиною пропусків даних може стати неуважність оператора або програмний (технічний) збій під час роботи з програмою введення даних. Ми вважаємо, що ця причина найбільш неприємна, тому що втрачаються дані, які надійшли від респондентів, тобто та інформація, з якою можна було би працювати. Звісно, втрату даних через помилки оператора намагаються контролювати у всіх системах автоматизованого введення, але відрізнити причину відсутності відповіді респондента на конкретне запитання через те, що відповідь не була надана від випадку, коли помилився оператор і не ввів надану відповідь, не завжди представляється можливим.

Проте, надалі будуть розглядатись ситуації більш цікаві для статистиків: компенсація невідповідей, що виникли через відмову респондента від участі в обстеженні та не заповнення окремих полів в анкеті, тобто відмову від відповіді на запитання. Їх будемо називати глобальною та локальною відмовами. Організатори тих обстежень, участь у яких не є обов'язковою за законом, це, наприклад, обстеження ділової активності підприємств (кон'юнктурні обстеження, КО) або соціологічні чи медичні опитування, часто стикаються не тільки з локальними, але й з глобальними відмовами. Тим не менш, якщо респондент постійно бере участь в обстеженнях (як це, наприклад, рекомендує методологія КО, за якою слід працювати з постійною, наскільки це можливо, панеллю респондентів-підприємств), але за якими-то причинами пропускає чергове опитування, доцільним, на нашу думку, було би здійснення імпутації (заповнення) пропусків саме його відповідей за визначеними заздалегідь алгоритмами.

Глобальні відмови виникають через такі причини:

– респондент не отримав анкети для заповнення; цю причину просто усунути;

– респондент отримав анкету, але забув її заповнити вчасно; для цього повинен бути розроблений механізм повторних нагадувань;

– респондент взагалі відмовляється надалі брати участь в обстеженнях; у цьому випадку імпутація недоцільна;

– респондент декілька періодів поспіль не бере участі в опитуваннях, тобто це – тимчасова глобальна відмова (наприклад, через тимчасову відсутність менеджера, який раніше відповідав на запитання анкети); у цьому випадку можлива пізніша імпутація даних у попередні пропущені періоди, якщо передбачений їх перерахунок, або одночасна імпутація у кожне поточне обстеження, якщо респондент повідомляє про тимчасове припинення участі в опитуваннях. Але при цьому потрібно враховувати причини, через які респондент відмовляється тимчасово брати участь в обстеженні: якщо це, наприклад, короткострокове припинення роботи підприємства через відсутність сировини, комплектуючих, то чи навряд слід заповнювати його невідповіді, особливо ті, що стосуються виробничої діяльності;

– респондент зникає (це відноситься до випадків, коли респондентами є юридичні особи, які можуть ліквідуватись, змінити статус тощо); у цьому випадку імпутація теж недоцільна.

Локальні відмови можуть виникати через такі причини:

– респондент не знає відповіді на конкретне запитання через те, що анкету заповнює не та особа, для якої її було розроблено. Наприклад, в КО передбачається, що анкету заповнює керівник підприємства або особа, що належить до складу вищого менеджменту підприємства і володіє інформацію щодо питань анкети. Такі пропуски можна заповнити шляхом імпутації, розробивши для цього систему алгоритмів заповнення;

– респондент не хоче відповідати на запитання, наприклад, через підозру, що його відповідь може нанести шкоду (спричинити неприємності). Ця ситуація може виникнути через некоректне формулювання запитань анкети (наприклад, керівник підприємства не вірно зрозумівши питання, може запідозрити спробу втручання у фінансову діяльність підприємства або просто формулювання питання викликає у респондента роздратування), спроби організаторів обстеження зібрати інформацію про сферу діяльності респондента, яка, на його думку не може бути висвітлена (наприклад, тіньова діяльність підприємства) тощо. Такі пропуски не завжди логічно заповнювати штучно, а іноді краще залишити без змін як джерело додаткової інформації щодо конкретного питання.

Роботи з обробки даних із пропусками з'явилися порівняно недавно, серед найбільш популярних можна назвати [1-4]. Автори розподіляють методи обробки даних з пропусками на чотири групи, що перетинаються [5].

1. *Метод виключення некомплектних об'єктів.* При відсутності у деяких об'єктів значень яких-небудь змінних простим прийомом є видалення таких некомплектних об'єктів з аналізу й обробка даних без пропусків (див., наприклад, [6]). Цей підхід легко реалізовується й може бути задовільним при невеликій кількості пропусків. Однак іноді він призводить до серйозних зсувів і зазвичай не дуже ефективний.

2. *Методи із заповненням.* Пропуски заповнюються й отримані "повні" дані обробляються звичайними методами. Щоб отримати коректні висновки, у стандартні методи аналізу варто ввести модифікації, що дозволяють відрізняти заповнені пропуски від реальних даних. Ці модифікації відносно прості в узагальненні з багаторазовим заповненням кожного пропуску.

3. *Методи зважування.* Висновки за даними вибіркових обстежень із пропусками звичайно побудовані на вагах плану, що обернено пропорційні ймовірності вибору, а ці методи змінюють ваги, щоб врахувати відсутність значень.

4. *Методи, засновані на моделюванні.* Широкий клас методів ґрунтується на побудові моделі утворення пропусків. Висновки одержують за допомогою функції правдоподібності, що побудована за умови справедливості цієї моделі, з оцінюванням параметрів методами типу максимальної правдоподібності. Переваги такого підходу полягають у тому, що він гнучкий і дозволяє відмовитися від методів, розроблених для окремих випадків.

Заповнення - це загальний і гнучкий метод розв'язання задач при наявності пропусків у спостереженнях. Проте йому властиві недоліки. У [7] відмічено: "Ідея заповнення й звабна, і небезпечна. Дослідник може заспокоїтися й прийти до приємного висновку, що врешті-решт його дані не містять пропусків. Небезпека цього підходу в тім, що він не дозволяє відрізняти ситуації, де задача не дуже важка й може бути коректно вирішена таким способом, від ситуацій, де звичайні оцінки за реальними і підставленими даними сильно зміщені".

До методів заповнення пропущених значень у вибіркових обстеженнях звичайно відносять такі [5, 8]:

1) *Заповнення середніми* за присутніми значеннями у вибірці.

2) *Заповнення з упередженим підбором* (або метод "hot deck" імпутації). Це метод, при застосуванні якого підстановка вибирається для кожного пропущеного значення за оцінкою розподілу. У більшості випадків емпіричний розподіл задається наявними значеннями, тому при заповненні з підбором вставляються різні значення з даних для схожих об'єктів без пропусків.

3) *Заміна* – метод обробки пропусків на етапі збору даних під час обстеження. Об'єкт, від якого не отримано відповіді, замінюється на інший, не включений до вибірки. Успішність використання цього методу залежить від схожості цих двох об'єктів за певним (необхідним) набором ознак.

4) *Заповнення без підбору* (або метод "cold deck" імпутації). Пропуск заповнюється постійним значенням із зовнішнього джерела, наприклад, значенням з попереднього спостереження з цього обстеження.

5) *Заповнення за регресією* - це заповнення пропусків значеннями, отриманими за регресією пропущених для даного об'єкту змінних на присутні, яка обчислюється за комплектними об'єктами (тобто тими, відносно яких є повна інформація).

6) *Стохастичне заповнення за регресією* ґрунтується на заміні пропуску значенням, яке підставляється при заповненні за регресією, в сумі із залишком, який відображає невизначеність передбачуваного значення.

7) *Композиційні методи* ґрунтуються на ідеях декількох методів. Наприклад, можна об'єднати заповнення з підбором і заповнення за регресією, обчислюючи передбачуване регресією значення та додаючи потім залишок, який випадковим чином вибирається з емпіричних залишків для передбачених величин при формуванні значень для підстановки.

8) *Методи багаторазового заповнення* передбачають заповнення пропуску декількома значеннями. Суттєвим недоліком методів одноразового заповнення, за думкою їх дослідників (наприклад, [5, 9]), є те, що звичайні формули призводять до систематично занижених оцінок дисперсії, навіть, якщо обчислені пропущені значення отримані з використанням вірної моделі. При багаторазовому заповненні отримуються правильні оцінки дисперсії, які можна отримувати звичайними методами аналізу повних даних.

Постановка проблеми. КО підприємств відрізняються від інших статистичних обстежень тим, що дані, які збираються, мають, в основному, якісний характер. Тобто респонденти заповнюють анкету, що надсилається їм регулярно, і містить запитання, які з'ясовують їх думки щодо стану підприємства та його перспектив. Специфічний характер даних КО обумовлює і методи їх обробки.

На цей час майже не існує робіт, пов'язаних із заповненням пропущених даних КО, як виняток можна тільки навести [10-12], де наводяться алгоритми імпутації для різних типів запитань, як якісних, так і кількісних, для окремих підприємств. Іншим шляхом заповнення даних може бути використання логлінійних моделей, як описано у [5, 13, 14]. Таким чином, для імпутації пропущених даних КО якісного характеру, можливо йти двома шляхами:

перший – розробити процедури і алгоритми заповнення пропусків для даних за окремими підприємствами, другий – розробити алгоритм заповнення пропусків для агрегованих даних, тобто використовувати частоти для варіантів відповідей, розраховані по групах підприємств.

Метою дослідження є створення процедур імпутації для окремих підприємств. З метою викладення методологічних засад із створення та застосування процедур імпутації даних для покращення інформації, що отримується з кон'юнктурних обстежень підприємств, слід розглянути такі питання: визначити всі типи показників, що використовуються в КО; визначити системи показників з анкет для окремих груп видів діяльності підприємств (промисловість, будівництво, торгівля, транспорт, сільське господарство), для яких можуть бути застосовані процедури імпутації, розроблені на основі алгоритмів, окремих для різних типів показників; визначити групи показників, для яких не можуть бути застосовані процедури імпутації; описати алгоритми імпутації для різних типів показників; визначити методи оцінювання якості результатів застосування алгоритмів імпутації з метою отримання висновків щодо доцільності її впровадження.

Викладення основного матеріалу дослідження. Можна визначити такі типи показників: запитання з триваріантними відповідями; дихотомічні запитання; запитання з багатоваріантними відповідями; кількісні запитання. Крім того, слід також розглянути особливі випадки, коли з метою заповнення пропущених значень одночасно аналізуються декілька пов'язаних за смислом запитань.

Також слід виділити окремо групи показників, для яких з тих чи інших причин неможливо застосовувати процедури імпутації значень. Такими причинами можуть бути наступні:

- інша періодичність збирання даних. Це стосується, наприклад, питань про стан інвестування на підприємствах, що ставляться двічі на рік і тому, через досить великий проміжок часу між двома відповідями поспіль (півроку), за який ситуація на підприємстві може значно змінитись, надійно відтворити пропущену відповідь за двома іншими, отриманими через рік, складно;
- нерегулярність включення деяких не гармонізованих з європейськими питань до анкет, що пов'язана із зміною економічної кон'юнктури;
- необхідність визначення незаповнених кодових полів анкети (коди ЄДРПОУ, КВЕД, КОАТУУ, КОПФГ, КФВ) та показників, що відносяться до атрибутивних характеристик респондента (наприклад, кількість працюючих на підприємстві). Такі пропуски у даних повинні заповнюватися за окремими процедурами, в основному передбаченими у процесі введення даних;

- особливий характер показників, обумовлений їх економічним смислом. До них відносяться, наприклад, запитання щодо змін цін на товари підприємства або зміни рівня торгових націнок. Така інформація повинна, на наш погляд, відображати реальний стан і відбивати погляди тільки тих респондентів, хто захотів відповідати на запитання.

Нижче пропонуються алгоритми імпутації, розроблені автором для всіх наведених вище типів показників, придатні саме для умов української економіки.

Головними гіпотезами, що лежать в основі розроблених алгоритмів, є такі: по-перше, заповнення пропусків для кожного окремого підприємства панелі респондентів краще відтворює реальну ситуацію, ніж заповнення одночасно для цілої групи підприємств за результуючими розрахунками балансів. До того, це дозволяє здійснювати групування підприємств за різними ознаками, не зважаючи на те, як було здійснено імпутацію. По-друге, вважається, що відповіді того самого респондента на те саме запитання у двох кварталах поспіль можуть змінюватись, причому зміни можуть бути кардинальними (з відповіді "збільшення" на відповідь „зменшення" чи навпаки), тобто допускається будь-яка комбінація двох відповідей поспіль для кожного підприємства. Це обумовлено і змінами в економіці країни, і впливом сезонних факторів на деякі види економічної діяльності та на деякі окремі показники. По-третє, відповіді підприємств того ж самого виду діяльності та найближчих за розміром аналогічні.

У загальному випадку щодо наявності невідповідей слід розглядати 3 ситуації, позначаючи поточний період (квартал) через Т:

1) необхідно заповнювати пропуски значень у періоді Т-2;

2) необхідно заповнювати пропуски значень у періоді Т-1;

3) необхідно заповнювати пропуски значень у періоді Т.

Зрозуміло, що при цьому слід розглядати всі комбінації відповідей-невідповідей, що виникають. Вони можуть бути такими (див. табл. 1):

Таблиця 1. *Комбінації варіантів наявності-відсутності даних для трьох періодів поспіль*

	Період		
	Т-2	*Т-1*	*Т*
1.	R	R	R
2.	R	R	**NR**
3.	R	**NR**	R
4.	R	**NR**	**NR**
5.	**NR**	R	R
6.	**NR**	R	**NR**
7.	**NR**	**NR**	R
8.	**NR**	**NR**	**NR**

Примітка: R - наявність відповіді, NR – невідповідь

Очевидно, що у подальшому будуть розглядатись комбінації 2-7.

Відновлення відповіді у період Т-2 напряму здійснюється за алгоритмами, наведеними нижче для комбінацій 5 та 6, та може бути здійснено для комбінації 7 при наявності відповіді у період Т-3.

Відновлення відповіді у період Т-1 напряму здійснюється у випадках 3, 4 та 7 (процедури для 4 та 7 аналогічні, але протилежні за напрямком дії).

Відновлення відповіді у період Т напряму здійснюється у випадках 2 та 6.

Відновлення відсутніх відповідей логічно здійснювати тільки за умови, що питома вага отриманих відповідей, на основі яких реалізуються алгоритми імпутації, становить не менше 70 %. В іншому випадку викривлення інформації може бути суттєвим. Під час здійснення процедури імпутації у наступних періодах до уваги беруться тільки ті відповіді, що отримані безпосередньо від респондентів, а ті, що були імпутовані, до розрахунків не залучаються.

Заповнення пропусків для запитань з триваріантними відповідями

Більш надійно, на нашу думку, не просто переносити відповіді окремого респондента з попереднього опитування на наступне, якщо існує пропуск, як пропонується у [11], а для заповнення пропусків спиратись там, де можливо, на дані щодо двох попередніх відповідей цього респондента, використовуючи, як було зауважено вище, інформацію стосовно відповідей найближчих за ознаками і найближчих за попередніми відповідями підприємств. В інших випадках можна використовувати одну відповідь, надану цим підприємством раніше або пізніше. Такий метод, за наведеною вище класифікацією, можна віднести до методів з упередженим підбором (“hot deck” імпутації).

Якщо позначити перший варіант триваріантної відповіді (наприклад, “збільшення”) через “+”, другий (“без змін”) - як “=”, а третій (“зменшення”) – як “-”, то процедура перенесення відповідей буде здійснюватись за алгоритмом, поданим у табл. 2.

Таблиця 2. *Алгоритм заповнення пропусків для триваріантних запитань*

Номер комбінації варіантів	*Т-2*	*Т-1*	*Т*	*Дія*
2	R	R	NR	заповнення Т з відповідною імовірністю на основі Т-2 і Т-1
3	+	NR	+ або =	+ в Т-1
3	+	NR	-	= в Т-1
3	=	NR	R	= в Т-1
3	-	NR	+	= в Т-1
3	-	NR	-	- в Т-1
4	R	NR	NR	заповнення Т-1 з відповідною імовірністю на основі Т-2
5	NR	R	R	заповнення Т-2 з відповідною імовірністю на основі Т-1 і Т
6	NR	R	NR	заповнення Т з відповідною імовірністю на основі Т-1
7	NR	NR	R	заповнення Т-1 з відповідною імовірністю на основі Т

Слід зазначити, що імовірність для заповнення невідповідей визначається частотами наявних комбінацій відповідей для підприємств з близькими атрибутами. Наприклад, для комбінації 2 визначаються частоти отримання всіх варіантів відповідей "+", "=", "-" (P+, P-, P=) у період T для різних комбінацій відповідей у періоди T-1 і T-2. Для заповнення невідповіді обирається така відповідь, що має найбільше значення частоти, тобто $\max\{Pj\}$, при $j^{\in\{+,-,=\}}$. У випадку, коли частоти двох відповідей співпадають, обирається та, що відповідає загальній тенденції за видом діяльності та розміром підприємства. Для комбінації 5 алгоритм аналогічний. Для комбінацій 4, 6 та 7 визначаються імовірності отримання всіх трьох варіантів відповідей на основі найближчої відповіді.

Заповнення пропусків для дихотомічних запитань

Процес відновлення цього типу запитань слід окремо здійснювати для тих, які знаходяться під впливом сезонних факторів, і тих, для яких ці фактори не діють. Для другої групи запитань пропонується алгоритм перенесення, наведений у табл. 3. Для тих, що мають сезонну компоненту, перенесення слід здійснювати з урахуванням сезонних змін, визначених для відповідного виду економічної діяльності. Це можна зробити, наприклад, шляхом розрахунку питомої ваги кожної з двох відповідей.

В українських обстеженнях поки що відсутні дихотомічні запитання, що ставляться регулярно і мають сезонний вплив, тому наведений у табл. 3 алгоритм стосується тільки випадку, коли сезонний фактор не враховується.

Таблиця 3. *Алгоритм заповнення пропусків для дихотомічних запитань*

Номер комбінації варіантів	*T-2*	*T-1*	*T*	*Дія*
2 або 6	R або NR	R	NR	перенесення T-1 в T
3 або 4	R	NR	R або NR	перенесення T-2 в T-1
5	NR	R	R	перенесення T-1 в T-2
7	NR	NR	R	перенесення T в T-1

Заповнення пропусків для запитань з багатоваріантними відповідями

Імпутація відповідей здійснюється тільки для комбінацій 2, 3 та 5 табл. 1, тобто алгоритм має такий вигляд, як представлено у табл. 4. Це обумовлюється тим, що варіанти відповідей окремих підприємств на багатоваріантні запитання не змінюються так швидко у часі, як на інші типи запитань, тому наявність однакових варіантів відповідей у двох обстеженнях поспіль дає підставу перенести їх і на третє обстеження (комбінації 2 та 5), а якщо є однакові відповіді у обстеженнях через одне, то можна з великою долею імовірності стверджувати, що імпутація однакових варіантів з цих обстежень у обстеження, що проводилось у період між ними, теж вірна (комбінація 5).

Таблиця 4. *Алгоритм заповнення пропусків для багатоваріантних запитань*

Номер комбінації варіантів	*Т-2*	*Т-1*	*Т*	*Дія*
2	R	R	NR	перенесення співпадаючих у Т-1 та Т-2 відповідей в Т
3	R	NR	R	перенесення співпадаючих у Т та Т-2 відповідей в Т-1
5	NR	R	R	перенесення співпадаючих у Т та Т-1 відповідей в Т-2

Заповнення пропусків для кількісних запитань

Для цього типу запитань пропуски пропонується в основному заповнювати середніми значеннями показника, отриманими для відповідного виду економічної діяльності та розміру підприємства (див. табл. 5).

Таблиця 5. *Алгоритм заповнення пропусків для кількісних запитань*

Номер комбінації варіантів	*Т-2*	*Т-1*	*Т*	*Дія*
2 або 6	R або NR	R	NR	розрахунок на основі інформації з Т-1 та Т
3	R	NR	R	у Т-1 ставиться середнє значення Т-2 і Т
4	R	NR	NR	розрахунок на основі інформації з Т-2 та Т-1
5	NR	R	R	розрахунок на основі інформації з Т-1 та Т-2
7	NR	NR	R	розрахунок на основі інформації з Т та Т-1

Заповнення пропусків у особливих випадках

До особливих випадків слід віднести ті, в яких логічно пов’язані два або більше запитань. Імпутація повинна бути здійснена у межах однієї анкети (тобто для одного респондента у визначений період часу). Якщо однієї з двох пов’язаних відповідей не вистачає, то інша є джерелом для її заповнення. Тобто у цьому випадку для заповнення пропуску у відповіді одного респондента не залучається додаткова інформація, отримана від інших респондентів.

Наприклад, для обстежень промисловості – це можуть бути питання щодо достатності завантаження виробничих потужностей та двох факторів, що стримують виробництво, а саме – нестача устаткування та застаріле обладнання, якщо потужностей не вистачає або, якщо потужностей більш, ніж достатньо – то таких факторів як нестача сировини і матеріалів, відсутність налагодженої системи збуту продукції, низький платоспроможний попит на продукцію, нестача кваліфікованих кадрів, нестача електро-, енергозабезпечення, високі тарифи природних монополій.

Кожний з перерахованих вище факторів, що стримують виробництво, може стати джерелом інформації для запитання про достатність завантаження виробничих потужностей. Для цього можна запропонувати такий алгоритм як представлено у табл. 6.

Таблиця 6. *Алгоритм заповнення пропусків для особливих випадків (запитання щодо стримуючих факторів та завантаженості виробничих потужностей)*

Номер комбінації варіантів	*T-2*	*T-1*	*T*	*Дія*
2 або 6	R або NR	R	NR	перенесення значення з T-1 у T за алгоритмом
3 або 7	R або NR	NR	R	перенесення значення з T у T-1 за алгоритмом
4	R	NR	NR	перенесення значення з T-2 у T-1 за алгоритмом
5	NR	R	R	перенесення значення з T-1 у T-2 за алгоритмом

Наприклад, для комбінацій 2 та 6 імпутація здійснюється таким чином: якщо у багатоваріантному запитанні щодо стримуючих факторів підприємством у період T обрана відповідь про нестачу устаткування та/або про застаріле обладнання, і у попередньому періоді (T-1) респондентом було зазначено, що потужностей не вистачає або достатньо (тобто "-" або "="), то у періоді T відсутню відповідь на запитання щодо достатності завантаження виробничих потужностей слід заповнювати значенням з T-1 періоду. Якщо у T-1 була відповідь про надлишок виробничих потужностей, то у T заноситься відповідь "=".

При наявності відповідей про недостатній попит, нестачу сировини, відсутність системи збуту продукції тощо у T заноситься відповідь про надмірні виробничі потужності (тобто "+"), якщо у T-1 була відповідь "+" або "=".Якщо була відповідь "-", то заноситься "=".

Аналогічні алгоритми пропонується застосувати для інших комбінацій варіантів.

Оцінка якості застосування процедур імпутації

Останнім етапом застосування процедур імпутації, що працюють на основі запропонованих алгоритмів, є визначення методів оцінки якості нових даних з метою отримання висновків щодо доцільності впровадження таких процедур. В якості такого методу оцінки пропонується здійснити порівняння з відповідними статистичними даними. Наприклад, візьмемо вихідні дані та результати застосування процедури імпутації для цих даних для окремого виду діяльності "виробництво машин та устаткування" (розділ 29 за КВЕД) щодо показника "зміни обсягів реалізованої продукції у поточний період" (баланси відповідей) і розрахуємо коефіцієнти кореляції цих двох рядів даних з індексом промислового

виробництва для цього виду діяльності, який має смисл ланцюгового індексу. Балансом у методології КО називається різниця питомих ваг позитивних і негативних відповідей (див., наприклад, [15]). У табл. 7 представлено результати таких порівнянь.

Таблиця 7. *Порівняння даних обстежень і офіційної статистики*

	I-2005	*II-2005*	*III-2005*	*IV-2005*	*Коефіцієнт кореляції*
Індекси промислового виробництва	96,1	115	104,9	98,1	
Баланси до імпутації	-1	37	32	8	0,91984
Баланси після імпутації	-2	42	35	11	0,92377

Висновки:

1. У статті подано алгоритми заповнення пропусків, розроблені автором для різних типів показників КО підприємств, таких як запитання з триваріантними відповідями, дихотомічні запитання, запитання з багатоваріантними відповідями, кількісні запитання та для особливих випадків, коли одночасно розглядаються дані щодо двох логічно пов'язаних запитань.

2. Розрахунки показали, що запропоновані алгоритми імутації для триваріантних відповідей (при наявності відповідей цього підприємства у двох попередніх періодах та у одному попередньому періоді) надають набагато точніші результати, ніж алгоритми, наведені у [11].

3. Розроблено окремий алгоритм для особливих випадків (для двох взаємопов'язаних за смислом запитань), який до того не розглядався у літературі.

4. Запропоновано оцінювати якість результатів застосування процедур імпутації, що працюють на базі розроблених алгоритмів, шляхом порівняння розрахованих балансів відповідей за необробленими даними та тими, що отримані після імпутації, з кількісними даними офіційної статистики.

5. Коефіцієнти кореляції, розраховані для двох видів даних та індексу промислового виробництва для цього виду діяльності, покращуються після використання процедур імпутації.

6. На відміну від імпутації, що здійснюється для груп підприємств з використанням логлінійних моделей, запропоновані процедури дозволяють не тільки уточнювати розрахунки за видом економічної діяльності в цілому, але й заповнювати пропуски у даних за окремими підприємствами, які за різними причинами не відповіли або на всю анкету, тобто не повернули її (глобальний пропуск), або на окреме запитання (локальний пропуск). Тоді як інші процедури, що застосовуються для імпутації якісних даних за групами підприємств, дозволяють тільки заповнити пропуски, що виникли через ненадання відповідей на конкретне запитання тими підприємствами, що взяли участь в опитуванні.

ЛІТЕРАТУРА:

1. *Afifi A.A. and Elashoff R.M.* Missing observations in multivariate statistics: Review of the literature // Journal of American Statistic Association. – 1966. – 61. – Pp. 595-604.

2. *Hartley H.O. and Hocking R.R.* The analysis of incomplete data // Biometrics. – 1971. – 27. – Pp. 783-808.

3. *Orchard T. and Woodbury M.A.* A missing information principle: Theory and applications / Proc. 6th Berkley Symposium on Math. Statist. and Prob. – 1972. – Pp. 697-715.

4. *Dempster A.P., Laird N.M. and Rubin D.B.* Maximum likelihood from incomplete data via the EM algorithm (with discussion) // Journal of Royal Statistic Society. – 1977. – B39. – Pp. 1-38.

5. *Литтл Р.Дж.А., Рубин Д.Б.* Статистический анализ данных с пропусками / Пер. С англ. – М.: Финансы и статистика, 1991. – 336с.

6. *Nie N.H., Hill C.H., Jenkins J.G., Steinbrenner K. and Bent D.H.* SPSS, 2nd ed., McGraw-Hill, New York, 1975. – 356 p.

7. *Dempster A.P. and Rubin D.B.* Overview / Incomplete Data in Sample Surveys, Vol. II: Theory and Bibliography (W.G.Madow, I.Olkin and D.B.Rubin, Eds.). – New York: Academic Press, 1983. – Pp. 3-10.

8. *Сарiогло В.Г.* Сучасні методологічні підходи та алгоритми імпутації відсутніх даних при обробці результатів вибіркових обстежень умов життя населення / Проблеми статистики. Збірник наукових праць. – Вип.2. – 2000. – С.267-271.

9. *Rubin D.B.* Multiple Imputation for Nonresponse in Surveys. - New York: Wiley, 1987. – 421 p.

10. *Fournier J.-M.* Enquête de conjuncture dans l'industrie methodologie. – Paris : INSEE, 1996. – 70 p.

11. *Fanouillet J.C.* La rénovation des enquêtes nationales de conjuncture de 1991 / INSEE Methodes. – N 32. – 1992. – 68 p.

12. *Fansten M.* Introduction à une théorie mathématique de l'opinion / Annales de l'INSEE. – N21. – 1976. – P.134.

13. *Goodman L.A.* Simple Models for the Analysis of Association in Crossclassifications Having Ordered Categories // Journal of American Statistic Association. – 1979. – 74. – Pp. 537-552.

14. *McCullagh P.* Regression models for ordinal data // Journal of Royal Statistic Society. - 1980. – B42. – Pp. 109-142.

15. *Пугачова М.В.* Нові технології обстежень ділової активності. (За методикою Євросоюзу) // Матеріали семінару "Сучасна статистика і проблеми соціального та економічного розвитку України". – К.: Знання, 1997. – С. 15-19.