УДК 330.101:311.2

В. В. Липчук,
доктор економічних наук, професор,
член-кореспондент НААН України,
завідувач кафедри статистики та аналізу,
E-mail : wlipczuk@ukr.net;
О. М. Крупа,
кандидат економічних наук, доцент,
доцент кафедри статистики та аналізу,
E-mail : oksana_krupa@mail.ru;
Львівський національний аграрний університет

Редукція даних у соціально-економічних дослідженнях

Показано, що ускладнення процесу прийняття рішень в умовах невизначеності та труднощі прогнозування динамічного розвитку різноманітних суспільних явищ спричинюють потребу у щоразу більшій кількості первинних даних, які нагромаджуються у великих обсягах. Обґрунтовано необхідність редукції даних як важливого етапу забезпечення достовірності та економічності проведення соціально-економічних досліджень. Розглянуто суть процесу редукції, подано етапи її здійснення та перелічено найбільш типові використовувані методи.

Ключові слова: *соціально-економічні дослідження, редукція даних, суттєві ознаки, способи, етапи, достовірність, методи редукції.*

Питання редукції даних останніми роками досить широко дискутується в науковій літературі, адже її методи мають важливе практичне застосування в економіці, соціально-екологічних дослідженнях, психології, інженерних дисциплінах та інших наукових галузях. Цій проблематиці присвячені праці О. Гончар [1], А. Домаранської [2], О. Єлісеєвої [3] та інших вітчизняних і зарубіжних учених.

В умовах сьогодення кількість даних, які збираються у процесах соціально-економічних досліджень і зберігаються у продуктивних і відносно дешевих системах баз даних вільного доступу, є надто великою і постійно зростає. Водночас ускладнення процесу прийняття рішень в умовах невизначеності та труднощі передбачення динамічного розвитку явищ і процесів спричинюють щоразу більшу потребу у первинних даних, які нагромаджуються у великих обсягах. Зазначимо, що ці дані практично не використовуються у своїй "сирій", необробленій формі. Дослідники, застосовуючи різні техніки перетворення даних, намагаються зробити їх більш корисними, зручними і зрозумілими.

Водночас ускладнення процесу прийняття рішень в умовах невизначеності та труднощі передбачення динамічного розвитку широкого спектра суспільних явищ спричинюють потребу у щоразу більшій кількості первинних даних, які нагромаджуються у великих обсягах.

Проблема редукції даних у вітчизняній статистиці як окремий етап дослідження практично не висвітлюється, хоча в цілому наводяться і практично застосовуються відповідні методи, які розглядаються в багатьох навчальних джерелах, монографіях і окремих публікаціях. Прикладні аспекти розв'язання цієї проблеми потребують відповідного обґрунтування і подальшого поглибленого дослідження

Метою статті є розгляд редукції даних як окремого процесу в соціально-економічних дослідженнях, уточнення її сутності та спроба систематизації способів її здійснення.

У соціально-економічних дослідженнях розрізняють дві послідовні фази: підготовка дослідження і його безпосередня реалізація, що складаються з відповідних етапів (рис. 1, за даними [4]).

Дані, отримані в процесі збирання, є сирими і не придатні для безпосередньої обробки чи прямого застосування. Необхідне їх попереднє приготування до аналізу. У процесі статистичного дослідження достатньо поширеними є ситуації, коли кількість наявних даних починає перевищувати можливості дослідника щодо їх опрацювання і тоді легко заплутатися, загубитися в їх аналізі, а отже, користувачу даних буде неможливо прийняти правильне рішення. Із різних причин, насамперед методологічного характеру, може виявитися, що відібрана сукупність не підпорядковується нормальному закону розподілу, що унеможливлює застосування традиційних статистичних методів. Тоді виникає потреба редукції сукупності. Значущість результатів цього процесу ілюструє рис. 2.

Важливим попереднім етапом аналізу даних є їх попередня обробка з метою усунення або обмеження різного роду недосконалостей і окремих властивостей, які притаманні даним і можуть негативно впливати на отримані результати, а в окремих випадках – навіть їх сфальсифікувати. Насамперед це відбувається тоді, коли даних недостатньо, а також бракує можливостей їх перетво-

рення та застосування різних шкал вимірювання. Досить часто під час прогнозування виникає ситуація, коли для виконання розрахунків зібрано як занадто багато, так і надто мало даних. Деякі дані не стосуються проблеми, що розглядається, а отже, лише знижуватимуть точність прогнозу. Інші можуть відповідати проблемі, але тільки у певний період [5, с. 540].

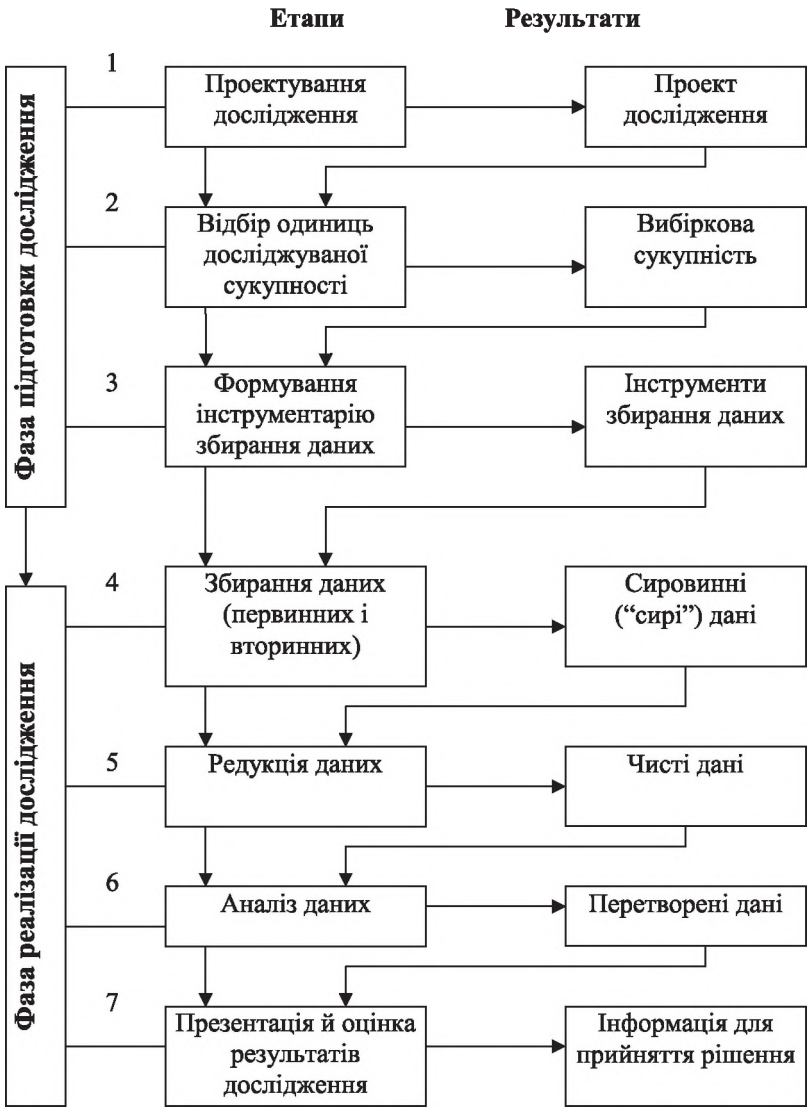

Рис. 1. Місце редукції даних в економічних дослідженнях

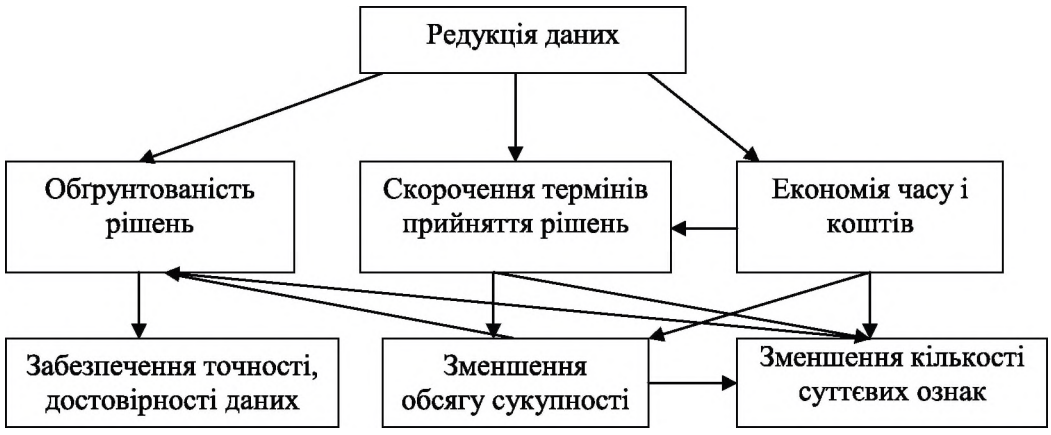

Рис. 2. Значення редукції даних в економічних дослідженнях

Редукція даних є етапом фази реалізації дослідження, який полягає в селекції найбільш суттєвих з погляду мети дослідження даних і дозволяє:

– підвищити ефективність та результативність дослідження;

– скоротити час його проведення;

– елімінувати помилки і надмірні дані з досліджуваної сукупності;

– зменшити вимоги до засобів обрахунків, скоротити зменшити вартість дослідження, завдяки економії на інструментах обробки даних;

– спростити представлення складної проблеми, що розв'язується;

– знайти компроміс між часом, необхідним для дослідження, та його вартістю.

Редукція даних дозволяє скорочувати час, необхідний для проведення аналізу, а відповідно і зменшує вимоги щодо інструментарію (засобів) розрахунків. Це, з одного боку, прискорює процес проведення дослідження, а з другого – покращує його якість. Дуже важливо, щоб способи редукції даних були передбачені при проектуванні дослідження [6, с. 97]. Беззаперечним є той факт, що не всі дані, отримані в процесі збирання, необхідно автоматично редукувати. Однак завжди варто розглядати таку можливість та експериментувати. Аналіз отриманих даних з наступним формулюванням корисних практичних висновків є типовими процесами в прийнятті рішень.

Редукція – це процес, який має на меті приведення сировинних ("сирих") даних до чистих та зменшення кількості атрибутів (ознак) одиниць сукупності, не суттєвих для подальшого аналізу. Це процес приготування одержаних у процесі збирання даних до їх аналізу, селекція дослідником тих даних, які найбільш суттєві з погляду мети дослідження. Таке приготування стосується як формальної, так і технічної сторін. Результатом указаних дій є впорядкування і попереднє представлення даних у описовому, табличному чи графічному вигляді. Особливе значення редукція даних має при анкетних дослідженнях, зокрема через Інтернет, оскільки при цьому використовується достатньо відкритих питань.

Важливого значення набуває усунення з сукупності даних ознак (атрибутів), які малозначущі для подальших класифікацій і аналізу. На практиці суттєві ознаки з позиції розглядуваної проблеми часто апріорі невідомі. Інтуїтивно, нагромадження дослідником якнайбільшого обсягу даних є спробою якнайповніше описати об'єкти за проблемою, що розв'язується. Проте використання надмірної кількості ознак об'єкта може призвести до неочікуваних наслідків, передусім – до неоднозначного тлумачення даних. Тому виникає проблема виявлення й елімінації надмірних, не суттєвих та "засмічених" ознак [7].

Редукція даних в соціально-економічних дослідженнях стосується як первинного дослідження, що передбачає опрацювання нових зібраних даних, так і вторинного, тобто використання інформації уже створених баз даних. Фактично можна вести мову про те, що редукція виконується щодо як даних поточних (on-line), так і даних історичних (off-line).

Таким чином, можна узагальнити, що редукція стосується:

– обмеження кількості елементів досліджуваної сукупності;

– обмеження кількості змінних, що характеризують сукупність;

– обмеження кількості ознак, що характеризують кожний елемент сукупності.

Найважливіше, на нашу думку, не те, що після редукції ми маємо значно менше даних, а те, що вони дають значно більше інформації, є більш змістовними. Багато залежностей і взаємозв'язків стають більш прозорими, наочними.

Редукція даних є процесом адаптації первинних даних до вимог щодо їх аналізу. Всі вихідні дані, а особливо ті, які представлені у вигляді заповнених анкет, часто можуть бути надлишковими. У ході редукції сировинні дані контролюються, рецензуються, редагуються, вводяться в комп'ютер, кодуються, сортуються, обраховуються, табулюються й агрегуються [8]. Отже, редукція даних охоплює: контроль збирання даних (який у вітчизняній методології розглядається як етап процесу збирання даних), редагування даних, класифікацію, зведення і групування даних, їх кодування, передання даних (у зарубіжних джерелах часто вживається термін "трансмісія даних") до пристроїв їх обробки – комп'ютерів. Розглянемо ці етапи.

Контроль збирання даних передбачає контроль ретельності збирача інформації, зокрема тривалості, місця, дати, предмета й інших елементів процесу збирання даних. Метою контролю є отримання впевненості, що весь процес дослідження організований і здійснений у передбачений спосіб, а дані, на основі яких буде проведено аналіз і зроблені висновки, відтворюють реальні факти щодо досліджуваної проблеми. Особливим завданням є виявлення фіктивних (не проведених) обстежень, що може викривити загальну оцінку досліджуваного явища чи процесу. Для цього контролююча особа може щоденно перевіряти підсумки збирання даних за окремими критеріями, використовувати "приховані" (подвійні) питання в опитувальних листах, а також широко застосовувати додаткове (повторне) збирання даних за тими самими респондентами (відбирають групу 10–20% респондентів) [8]. За наявності певних розбіжностей традиційно використовують метод коригуючих коефіцієнтів.

Редагування даних передбачає їх оцінку з погляду можливості подальшого застосування зі з'ясуванням їх точності, надійності, рівня деталізації. Сюди відносять виявлення помилок даних, суперечливих і невідповідних даних, некомплектних, неоднозначних та неадекватних відповідей, а також їх відсутності. Редагування можна вважати формальною і реальною верифікацією, тобто перевіркою зібраних даних з погляду їх ясності й точності на основі логічного аналізу отриманих даних та перевірки правдивості респондентів [9, с. 120].

Результатом цього процесу є отримання перевірених і відібраних даних, придатних з погляду відповідності досліджуваній проблемі. При цьому особа, яка здійснює редагування даних, може прийняти рішення про їх заміну, виключення із досліджуваної сукупності, повторне проведення обстеження і т. д. Відмітимо, що для оцінки наявності помилок у зібраних даних при редагуванні можуть бути використані традиційні ручні й машинні способи їх контролю, зокрема логічний та арифметичний.

З групуванням часто поєднується класифікація (розподіл і підрахунок одиниць з такими самими чи подібними ознаками) або типологія (відомі у вітчизняній статистиці типові групування є результатом виокремлення найбільш типових, характерних та поширених явищ). Остання особливо необхідна за наявності вкрай неоднорідного матеріалу, що не підлягає однозначній класифікації. Зазначимо також, що типологія, на відміну від класифікації, не повинна бути вичерпною або відокремленою.

Класифікація даних передбачає виділення їх окремих типів (груп) за певними критеріями відповідно до мети дослідження. Для цього можуть бути застосовані зведені статистичні формуляри, в яких нагромаджуються дані за певним критерієм. Класифікація полегшує зведення даних, яке може бути простим (арифметичне знаходження підсумків) та складним (розподіл на групи з характеристикою кожної групи системою показників та знаходженням групових підсумків). У первинних дослідженнях класифікація даних закладається в інструментах їх збирання із застосуванням відповідних класифікаційних запитань. Класифікація має відповідати таким умовам [6]:

• роздільність (не може бути жодних сумнівів щодо включення одиниці спостереження до відповідного класу (групи);

• закінченість (усі одиниці сукупності повинні бути охоплені класифікацією);

• однорідність (виокремленні класи явищ повинні бути внутрішньо однорідними).

Розрізняють просте групування даних, якщо воно проводиться за однією ознакою, та складне – за двома і більше ознаками. Методика групувань достатньо добре опрацьована і викладена в навчальній літературі зі статистики. Важливим етапом групування, який дозволяє віднести ті чи інші одиниці сукупності за певними ознаками до відповідної групи є кодування, що являє собою процес групування із призначенням символів (кодів) різним даним. Найчастіше кодування здійснюється з використанням так званої книги кодування, яка розроблюється з огляду на специфіку дослідження і містить відповідні інструкції (символи змінних, тип змінних, значення змінних і т. д.).

Табуляція даних – це представлення даних у вигляді таблиці. Застосування таблиць (одно-, дво- і багатовимірних) значно полегшує аналіз даних. Табуляція може бути простою і складною. Проста табуляція використовується для:

• визначення відсутності відповідей на запитання;

• виявлення відповідей, які суттєво відрізняються від інших;

• визначення виду емпіричного розподілу змінної;

• розрахунку загальних статистичних характеристик.

Складна табуляція є важливим інструментом для вивчення взаємозв'язку між змінними. Яскравим її проявом є кореляційні таблиці.

Проведення редукції даних вручну є рутинним та важким заняттям і займає багато часу. Але оскільки цей етап обробки даних є необхідним, його слід реалізувати із застосуванням комп'ютерної техніки. Цьому сприяла розробка в останні роки як нових статистичних програм обробки даних, так і методів перенесення даних із формулярів до комп'ютерів [6, с. 191].

Нагальним завданням сучасної інформатики, яка нагромаджує, зберігає і надає дані й інформацію, з огляду на зростання їх обсягів стає допомога людині у здійсненні аналізу для прийняття правильних рішень.

Для здійснення редукції машинним способом існує багато різних методів, які достатньо широко висвітлені в [10]. Зазначимо, що у вітчизняній фаховій літературі більша частина цих методів найчастіше вказується за англійською назвою. Серед методів редукції даних зазначимо насамперед групу методів головних компонент (МГК, англійською – Principal Component Analysis, PCA), до яких належать методи Simple PCA, Probabilistic PCA, Kernel PCA та ін. Також поширеними є методи залучення (упровадження, вкладення, англійською – Embedding), серед яких – Locally Linear Embedding, Stochastic Neighbor Embedding, Symmetric Stochastic Neighbor Embedding, t-Distributed Stochastic Neighbor Embedding, Neighborhood Preserving Embedding, Stochastic Proximity Embedding. Основною проблемою є їх застосування у вітчизняних дослідженнях.

Підсумовуючи, відмітимо, що аналіз даних і прийняття правильних рішень потребує селекції і впорядкування зібраних даних, що можна забезпечити використанням сукупності методів їх редукції. Вдало здійснена редукція даних зумовлює ефективність проведених досліджень завдяки забезпеченню точності й достовірності даних, економії коштів і часу. Мета дослідження може бути реалізована тільки за умови правильної організації дослідницького процесу загалом та, зокрема, науково обґрунтованому застосуванні редукції даних. У цьому контексті особливо значущими стають подальші дослідження, присвячені опрацюванню способів контролю збирання даних, їх редагуванню, класифікації, кодуванню та передачі даних до засобів їх обробки. Не менш важливим є технічне опрацювання імплементації сучасних методів редукції у практику вітчизняної статистики.

Список використаних джерел

1. Гончар О. В. Проблеми забезпечення якості статистичної інформації, отриманої з адміністративних джерел / О. В.Гончар // Статистика України. – 2011. – № 4. – С. 4–7.

2. Домаранська А. О. Якість даних в кількісних дослідженнях / А. О. Домаранська // Український соціум. – 2016. – № 4 (39). – С. 150–154.

3. Єлісєєва О. К. Економічна діагностика в управлінні виробничо-економічними системами (статистичний аспект): монографія / О. К Єлісєєва – Дніпропетровськ : Наука і освіта, 2006. – 292 с.

4. Kaczmarczyk S. Badania marketingowe, metody i techniki / Kaczmarczyk S. – Warszawa : PWE, 2011. – 500 s.

5. Коркуна Д. М. Фінансове прогнозування як основа фінансових планів підприємства / Д. М. Коркуна // Вісник Нац. ун-ту "Львів. політехніка". – 2008. – № 628. – С. 539–544.

6. Липчук В. В. Маркетингові дослідження : [навч. посіб.] / В. В. Липчук, Л. В. Погребняк. – Львів : Магнолія 2006, 2012. – 352 с.

7. Dash M. Feature selection for classification / M. Dash, H. Liu // Intelligence Data Analysis. – 1997. – № 1 (3). – S. 131–156.

8. Podstawy prowadzenia badan marketingowych. Etapy procesu badan marketingowych – teoria procesu badawczego [Zasob elektroniczny] / Grupa CRON, Projekt "MarketResearcher – Informatyczny system do tworzenia i zarzadzania badaniami (zewnetrznymi, wewnetrznymi, rynku, opinii)". – Tpyb dostepu :

file:///C:/Users/user/Downloads/Badania_podrecznik_1.pdf

9. Szewczyk M. Podstawy statystyczne badan marketingowych : [skrypt dla studentow] / M. Szewczyk, M. Ciesielska. – Opole : Oficyna Wydawnicza Politechniki Opolskiej, 2010. – 218 s.

10. Van der Maaten L. J. P. Dimensionality Reduction: A Comparative Review / L. J. P. van der Maaten, E. O.Postma, H. J. van der Herik // Journal of Machine Learning Research. – 2009. – № 10. – S. 1-41.

References

1. Honchar, O. V. (2011). Problemy zabezpechennia yakosti statystychnoi informatsii, otrymanoi z administratyvnykh dzherel [Problems of quality assurance of statistical information received from administrative sources]. *Statystyka Ukrainy – Statistics of Ukraine, 4, 4–7* [in Ukranian].

2. Domaranska, A. O.(2016). Yakist danykh v kilkisnykh doslidzhenniakh [Data quality in quantitative studies]. *Ukrainskyi sotsium – Ukrainian society, 4 (39), 150–154* [in Ukranian]..

3. Yelisieieva, O. K. (2006). *Ekonomichna diahnostyka v upravlinni vyrobnycho-ekonomichnymy systemamy (statystychnyi aspekt)[Economic diagnostics in the management of production-economic systems (statistical aspect)].* Dnipropetrovsk: Nauka i osvita [in Ukranian].

4. Kachmarchyk, S. (1997). *Badania marketingowe, metody i techniki [Marketing researches: methods and techniques].* Warszawa: PWE [in Polish].

5. Korkuna, D. M. (2008). Finansove prohnozuvannia yak osnova finansovykh planiv pidpryiemstva [Financial forecasting as the basis of the enterprise's financial plans]. *Visnyk Natsionalnoho Universytetu "Lvivska Politekhnika" – Bulletin of Lviv Politechnic National University, 628, 539–544* [in Ukranian].

6. Lypchuk, V. V., & Pohrebniak, L. V. (2012). *Marketynhovi doslidzhennia [Marketing researches].* Lviv: Mahnoliia 2006 [in Ukranian].

7. Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligence Data Analysis, 1 (3), 131–156* [in English]

8. Podstawy prowadzenia badan marketingowych. Etapy procesu badan marketingowych – teoria procesu badawczego. Basics of marketing research. Stages of the marketing research – the theory of the research process. CRON Group. Retrieved from file:///C:/Users/user/Downloads/Badania_podrecznik_1.pdf [in Polish].

9. Shevchyk, M., & Tsiesielska, M. (2010). *Podstawy statystyczne badanmarketingowych. Skrypt dlia studentow [Fundamentals of statistical marketing research. The manual for students].* Opole: Opole University of Technology Publishing House [in Polish].

10. Van der Maaten, L. J. P., Postma, E. O. & Van der Herik, H. J. (2009). Dimensionality Reduction: A Comparative Review. *Journal of Machine Learning Research, 10, 1–41* [in English]

В. В. Липчук,
доктор экономических наук, профессор,
член-корреспондент НААН Украины,
заведующий кафедрой статистики и анализа;
О. Н. Крупа,
кандидат экономических наук, доцент,
доцент кафедры статистики и анализа;
Львовский национальный аграрный университет

Редукция данных в социально-экономических исследованиях

Показано, что усложнение процесса принятия решений в условиях неопределенности и трудности прогнозирования динамического развития различных общественных явлений обусловливают потребность в растущем количестве первичных данных, которые накапливаются в больших объемах. Обоснована необходимость редукции данных как важного этапа обеспечения достоверности и экономичности проведения социально-экономических исследований. Рассмотрены суть процесса редукции, представлены этапы ее осуществления и перечислены наиболее типичные используемые методы.

Ключевые слова: *социально-экономические исследования, редукция данных, существенные признаки, способы, этапы, достоверность, методы редукции.*

V. V. Lypchuk,
DSc in Economics, Professor,
Corresponding Member of NAAS of Ukraine,
Head of the Department of Statistics and Analysis;
O. M. Krupa,
PhD in Economics, Associate Professor,
Associate Professor of the Department of Statistics and Analysis;
Lviv National Agrarian University

Data Reduction in Socio-Economic Studies

The article is devoted to the problem of data reduction as an important step on the way of providing reliability and efficiency of socio-economic studies. Through the reduction the large amounts of raw data, generated from different sources, become more useful, convenient and clear for use. Meanwhile, the data reduction is not treated as a separate phase of studies in the national statistic practice. The aim of the article is to substantiate the importance of data reduction in economic studies and attempt to systematize and generalize the essence and components of the phase of data reduction as well as ways of their implementation. The study is based on methods of theoretical generalization, abstract and logic, analogy and others.

The essence of data reduction is defined as the process of converting raw data into the pure form and reducing the number of units' attributes (features), which are not significant to further analysis. In fact, this is part of the analysis involving selection of the data that are most important from the viewpoint of the study's goals. The significance of data reduction in economic studies is outlined. It is found that it assures the validity of their results, reduces their time and costs, simplifies the representational complexity of the problem being addressed, eliminates the errors and redundant data from the investigated set, looses the requirements to calculation tools. The data resulting from reduction are much more informative. Many dependencies and relationships become more readable (visual). It is emphasized that reduction applies to the current data (on-line), as well as to historical data (off-line), contained in the already created databases. The phases of data reduction are described. They are: control of data collection, data editing, classification, data construction and grouping, coding and transmission (data transmission to the processing tools – computers). Data reduction techniques and methods most common in the global practice are shown.

Future studies of data reduction problems are expected to focus on potential ways to implement its advanced methods in the domestic practice of statistical science. It will allow for enhancing significantly the speed and efficiency of economic analysis and the reliability of its results.

Key words: *socio-economic studies, data reduction, substantial signs, techniques, stages, credibility, reduction methods.*

Бібліографічний опис для цитування:
Липчук В. В. Редукція даних у соціально-економічних дослідженнях / В. В. Липчук, О. М. Крупа // Статистика України. – 2017. – № 1. – С. 15–20.