Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
https://nv.nltu.edu.ua
https://doi.org/10.15421/40290623
Article received 28.05.2019 p.
Article accepted 27.06.2019 p.
УДК 004.8

ISSN 1994-7836 (print)
ISSN 2519-2477 (online)

@ Correspondence author
O. S. Mishchuk
oleksandra.myroniuk@gmail.com

О. С. Міщук, Р. О. Ткаченко
Національний університет "Львівська політехніка", м. Львів, Україна

МЕТОДИ ОБРОБЛЕННЯ ТА ЗАПОВНЕННЯ ПРОПУЩЕНИХ ПАРАМЕТРІВ У ДАНИХ ЕКОЛОГІЧНОГО МОНІТОРИНГУ

Сьогодні існує багато методів відновлення пропущених параметрів у даних, але для кожної області застосування використовують різні методи заповнення пропусків. У роботі проаналізовано такі методи оброблення пропусків: видалення елементів з пропусками, метод зважування та заповнення пропущених параметрів. Описано механізми утворення пропущених параметрів, за яких ймовірність пропусків для кожного запису набору однакова, за яких ймовірність пропусків визначається на основі іншої наявної інформації без пропусків та за яких дані відсутні залежно від невідомих чинників. Проаналізовано методи заповнення пропущених параметрів у даних екологічного моніторингу, такі як: методи середнього значення, наївного прогнозу та регресійного моделювання. Описано такі методи відновлення пропусків на основі регресійного моделювання: багатошаровий персептрон; Adaptive Boosting; метод опорних векторів; Random Forest та метод лінійної регресії з використанням стохастичного градієнтного спуску. Виконано порівняння найпростіших методів заповнення пропусків та методів відновлення пропусків на основі регресійних моделей. Експериментально доведено, що попередньо розроблений метод заповнення пропусків на основі нейроподібної структури моделі послідовних геометричних перетворень є найефективнішим методом, оскільки показує найточніші результати.

Ключові слова: пропуски в даних; оброблення пропущених елементів; методи заповнення пропусків; регресійне моделювання.

Вступ. З проблемою оброблення пропусків у масивах даних доводиться стикатися під час проведення різноманітних соціологічних, економічних, статистичних, інформаційних та інших досліджень (Maltsev, & Mukharamova, 2011). Традиційними причинами, що призводять до появи пропусків у даних, є неможливість отримання інформації, її спотворення чи навіть приховування. Як наслідок – для виконання аналізу зібраних даних подаються неповні відомості.

Найпростішим рішенням оброблення неповних даних є виключення некомплектних спостережень, що містять пропуски, і подальший аналіз отриманих таким способом "повних" даних. Зрозуміло, що такий підхід призводить до сильної відмінності висновків, зроблених за наявності в даних пропусків і за їх відсутності. Тому перспективнішим є інший шлях – заповнення пропусків перед аналізом масиву даних. Такий підхід має такі переваги: чітке представлення структури даних; обчислення необхідних пропущених значень; впевнена інтерпретація результатів аналізу, оскільки можна спиратися на традиційні характеристики та сумарні значення.

Людська діяльність з кожним днем призводить до збільшення концентрації шкідливих домішок у довкіллі, що своєю чергою збільшує кількість різноманітних захворювань, а інколи, призводить навіть до смерті (Shypulin, 2012). Для контролю викидів встановлено певну межу: якщо виміряний параметр перевищує встановлену межу, вживають певних засобів. Але, як було згадано, не завжди є можливість отримати повні дані виміряних параметрів забруднення повітря.

Тому виникає потреба в аналізуванні наявних та дослідженні нових методів відновлення пропущених значень у масивах даних екологічного моніторингу, для знаходження такого алгоритму, котрий максимально задовольнятиме потреби у збільшенні швидкості, ефективності та точності заповнення пропущених параметрів.

Аналіз останніх досліджень і публікацій. Проблеми оброблення та аналізу пропущених параметрів у даних дослідили такі науковці, як: К. Мальцев, Н. Кузнєцова, Дж. Грехем, Р. Дж. Літтл, Д. Ньюман, С. Ван Бюрен та ін. Наприклад, у роботі Ньюмана (Newman, 2014) можна ознайомитися зі спробами узагальнення основ оброблення пропущених даних у соціальних науках. У роботі (Graham, Olchowski, & Gilreath, 2007) Грехем описує багаторазове заповнення пропусків. Але підбір алгоритмів оброблення пропусків залежить від даних та від області використання, тому виникає потреба в експериментальному дослідженні різних методів заповнення пропущених параметрів для покращення результатів у даних екологічного моніторингу.

Виділення не вирішених раніше частин загальної

Інформація про авторів:
Міщук Олександра Сергіївна, асистент, кафедра інформаційних технологій видавничої справи.
Email: oleksandra.myroniuk@gmail.com; https://orcid.org/0000-0001-6823-985X
Ткаченко Роман Олексійович, д-р техн. наук, професор, завідувач кафедри інформаційних технологій видавничої справи.
Email: roman.tkachenko@gmail.com; https://orcid.org/0000-0002-9802-6799
Цитування за ДСТУ: Міщук О. С., Ткаченко Р. О. Методи оброблення та заповнення пропущених параметрів у даних екологічного моніторингу. Науковий вісник НЛТУ України. 2019, т. 29, № 6. С. 119–122.
Citation APA: Mishchuk, O. S., & Tkachenko, R. O. (2019). Methods of processing and filling of missing parameters in ecological monitoring data. *Scientific Bulletin of UNFU, 29*(6), 119–122. https://doi.org/10.15421/40290623

проблеми. Досить часто дані екологічного моніторингу містять пропуски. Причини появи пропущених параметрів у даних моніторингу забруднення довкілля бувають різними: поломка приладів; несприятливі погодні умови; помилки приладів вимірювання; пошкодження носіїв інформації; призупинення вимірювань під час вихідних днів; виконання мінімальної кількості вимірювань, дозволених державними стандартами. Оскільки йдеться про моніторинг забруднення навколишнього середовища, а саме повітря, без якого людина не може існувати, тому для якісного аналізу даних екологічного контролю важливу роль відіграє відновлення пропусків у таких даних.

Метою роботи є застосування вибраних наявних методів заповнення пропусків на досліджуваних даних екологічного моніторингу, у тому числі з практичною перевіркою відновлювальної здатності найвідоміших алгоритмів; експериментальне використання розробленого неітеративного, нейроподібного фреймворку для заповнення пропусків у даних та порівняння результатів експериментів з відомими методами.

Матеріал і методи дослідження. Вибір методу заповнення пропусків є непростим завданням, і залежить від різних чинників: наявності регулярних компонент і їх особливостей, причин виникнення пропусків у даних і характеру цих пропусків (випадковий чи ні), а також особливостей даних і проведеного дослідження (Karahalios et al., 2012).

Оброблення пропущених значень є досить розвиненою дослідницькою областю з загальноприйнятою термінологією і великою кількістю рішень для різних дисциплін і конкретних досліджень.

Опис методів дослідження. Для того, щоб зрозуміти, як правильно обробити пропуски, необхідно визначити механізми їх формування. Розрізняють такі три механізми формування пропусків:

1. *MCAR* (*Missing Completely At Random*) – механізм формування пропусків, за якого ймовірність пропуску для кожного запису набору однакова. У такому випадку ігнорування чи вилучення записів, що містять пропущені параметри, не веде до спотворення результатів.
2. *MAR* (*Missing At Random*) – механізм формування пропущених паратерів, за якого ймовірність пропуску може бути визначена на основі іншої наявної в наборі даних інформації без пропусків. Зазвичай це не випадково пропущені параметри, а через деякі закономірності, тому вилучення чи заміна пропусків, як і у випадку MCAR, не призводить до істотного спотворення результатів.
3. *MNAR* (*Missing Not At Random*) – механізм формування пропусків, за якого дані відсутні залежно від невідомих чинників. MNAR припускає, що ймовірність пропуску могла б бути описана на основі інших атрибутів, але інформація по цих атрибутах у наборі даних. Як наслідок, ймовірність пропуску неможливо висловити на основі інформації, що міститься в наборі даних (Karahalios et al., 2012).

Існує три основні групи методів оброблення пропущених параметрів у масивах даних (Graham, Olchowski, & Gilreath, 2007). Як видно з рис. 1, це є: вилучення неповних вимірювань, зважування повних спостережень для штучного досягнення запланованого обсягу вибірки та заповнення пропущених параметрів.

Вилучення, тобто видалення неповних виміряних об'єктів, легке в реалізації, але необхідною умовою його використання є приналежність пропущених параметрів до MCAR. Окрім цього, цей метод не є ефективним, бо необхідно, щоб кількість пропусків була невеликою, інакше можуть виникнути сильні зміщення (Karahalios et al., 2012).

Рис. 1. Способи аналізу даних з пропусками

Другий варіант аналізу даних з пропусками містить такі методи зважування: компенсація неправильної ймовірності вибору; компенсація пропусків; створення зваженого розподілу вибірки для ключових змінних, що цікавлять. Зважування здійснюють за допомогою задання початкових ваг, після чого розділяють вибірку на підгрупи й обчислюють зважені рівні відповіді для кожної підгрупи. Потім використовують аналог рівня відповідей підгруп для регуляторів пропусків. І останнім кроком розраховують ваги пропущених параметрів (Van Buuren, 2012).

Третій спосіб оброблення даних з пропусками полягає у заповненні пропущених параметрів. Методи заповнення пропусків у даних екологічного моніторингу можуть містить в собі такі прості алгоритми: заповнення пропущених параметрів за допомогою середнього значення; інтерполяція за сусідніми точками; середнє за n сусідніми точками; медіана N за найближчими значеннями; метод наївного прогнозу; заповнення пропусків з використанням лінійної регресії. Також використовуються й інші методи – метод розрахунку середнього за відповідною датою, метод пошуку відсотка від максимуму або іншої знакової величини та ін. У дослідженні для експериментального порівняння, на даних екологічного моніторингу було використано прості методи середнього значення та наївного прогнозу; також застосовано кілька алгоритмів регресійного моделювання, включно з розробленим нейроподібним фреймворком для заповнення пропусків.

Метод середнього значення. У разі вибору цього методу всі відсутні дані просто замінюються середнім арифметичним значенням для всіх спостережень. Цей метод може не підходити, коли ряд не постійний або коли є великі систематичні коливання у змінних ряду. З іншого боку, повне середнє часто є кращим апріорним (неупередженим) припущенням для відсутніх даних. Цей метод має багато переваг, проте не володіє гнучкістю до несезонних змін і навіть може призвести до труднощів у визначенні лінії тренду, зробити її менш помітною для розпізнавання.

Метод наївного прогнозу. Наївний прогноз полягає в тому, що деякий основний період прогнозованого ряду найкраще описує майбутнє цього ряду. Тому моделі наївного прогнозу, як правило, є простою функцією від значень прогнозованої змінної у близькому минулому. Найпростішою моделлю наївного прогнозу є відповідність припущенню, що "завтра буде як сьогодні":

$$Y_{t+1}=Y_t,$$

де: Y_t – відоме значення, а Y_{t+1} – прогнозоване значення. У роботі застосовано наївний прогноз з використанням найпростішої функції. Основним недоліком наївного прогнозування є дуже низька точність прогнозу.

Відновлення пропусків на основі регресійних моделей. Методи лінійної регресії дають змогу отримати правдоподібно заповнені дані. Однак реальним даним властивий деякий розкид значень, який у разі заповнення пропусків на основі лінійної регресії відсутній. Як наслідок, варіація значень характеристики стає меншою, а кореляція між двома характеристиками штучно посилюється. Тому цей метод заповнення пропусків є ефективнішим, чим нижча варіація значень характеристики, пропуски в якій потрібно заповнити, і чим нижчим є відсоток пропущених параметрів від загальної кількості масиву даних.

Результати дослідження. У дослідженні додатково до найпростіших методів заповнення пропусків середнім значенням та методом наївного прогнозу, також застосовано такі методи регресійного моделювання: багатошаровий персептрон; Random Forest; Adaptive Boosting; метод опорних векторів та метод лінійної регресії з використанням стохастичного градієнтного спуску. Реалізацію описаних методів взято із бібліотек мови програмування Python. Також виконано заповнення пропусків із використанням ШНМ (Izonin et al., 2015; Tkachenko et al., 2017), а саме за допомогою нейроподібної структури на основі моделі послідовних геометричних перетворень, котру описано у (Izonin, & Tkachenko, 2019; Tkachenko, et al., 2018).

Запропоновані методи заповнення пропущених параметрів у масивах даних екологічного моніторингу використано на даних, котрі взято із відкритого репозиторію: UC Irvine Machine Learning Repository (De Vito et al., 2009).

Date	Time	CO(GT)	PT08.S1(C	C6H6(GT)	PT08.S2(N	NOx(GT)	PT08.S3(N	NO2(GT)	PT08.S4(N	PT08.S5(O	T	RH	AH
10.03.2004	18.00.00	2,6	1360	11,9	1046	166	1056	113	1692	1268	13,6	48,9	0,7578
10.03.2004	19.00.00	2	1292	9,4	955	103	1174	92	1559	972	13,3	47,7	0,7255
10.03.2004	20.00.00	2,2	1402	9	939	131	1140	114	1555	1074	11,9	54	0,7502
10.03.2004	21.00.00	2,2	1376	9,2	948	172	1092	122	1584	1203	11	60	0,7867
10.03.2004	22.00.00	1,6	1272	6,5	836	131	1205	116	1490	1110	11,2	59,6	0,7888
10.03.2004	23.00.00	1,2	1197	4,7	750	89	1337	96	1393	949	11,2	59,2	0,7848
11.03.2004	00.00.00	1,2	1185	3,6	690	62	1462	77	1333	733	11,3	56,8	0,7603
11.03.2004	01.00.00	1	1136	3,3	672	62	1453	76	1333	730	10,7	60	0,7702
11.03.2004	02.00.00	0,9	1094	2,3	609	45	1579	60	1276	620	10,7	59,7	0,7648
11.03.2004	03.00.00	0,6	1010	1,7	561	-200	1705	-200	1235	501	10,3	60,2	0,7517
11.03.2004	04.00.00	-200	1011	1,3	527	21	1818	34	1197	445	10,1	60,5	0,7465
11.03.2004	05.00.00	0,7	1066	1,1	512	16	1918	28	1182	422	11	56,2	0,7366
11.03.2004	06.00.00	0,7	1052	1,6	553	34	1738	48	1221	472	10,5	58,1	0,7353
11.03.2004	07.00.00	1,1	1144	3,2	667	98	1490	82	1339	730	10,2	59,6	0,7417
11.03.2004	08.00.00	2	1333	8	900	174	1136	112	1517	1102	10,8	57,4	0,7408
11.03.2004	09.00.00	2,2	1351	9,5	960	129	1079	101	1583	1028	10,5	60,6	0,7691
11.03.2004	10.00.00	1,7	1233	6,3	827	112	1218	98	1446	860	10,8	58,4	0,7552
11.03.2004	11.00.00	1,5	1179	5	762	95	1328	92	1362	671	10,5	57,9	0,7352
11.03.2004	12.00.00	1,6	1236	5,2	774	104	1301	95	1401	664	9,5	66,8	0,7951

Рис. 2. Дані екологічного моніторингу італійського міста з відкритого джерела

Заповнення пропусків відбувалося для кожного параметра окремо, в якому пропущені значення позначені в масиві "-200", що можна побачити на рис. 2. На прикладі параметра "CO" (оксид вуглецю), визначено похибки заповнення пропущених значень, котрі зображено на рис. 3 та 4. Для входів використано всі параметри, окрім "CO", котрий був застосований як вихід. Комірки із заповненими даними використано як виходи для навчання, а для знаходження пропущених значень параметра "CO", комірки з пропусками застосовано як виходи в режимі тестування. Отже, для розділеної вибірки даних на матриці: навчання та тестування; застосовано згадані вище методи заповнення пропусків та знайдено середню абсолютну (MAE) і середню абсолютну відсоткову (MAPE) похибки.

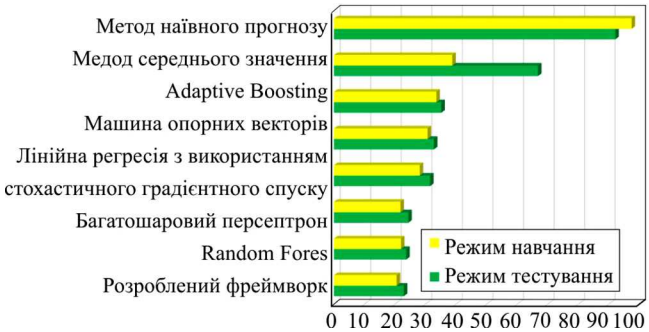

Рис. 3. Результати заповнення пропусків різними методами на основі середньої абсолютної відсоткової похибки (MAPE)

Ефективність розробленого методу визначалася шляхом порівняння результатів його роботи з результатами згаданих відомих простих методів регресійного моделювання та методів заповнення пропусків середнім значенням та методом наївного прогнозу, що зображено на рис. 3.

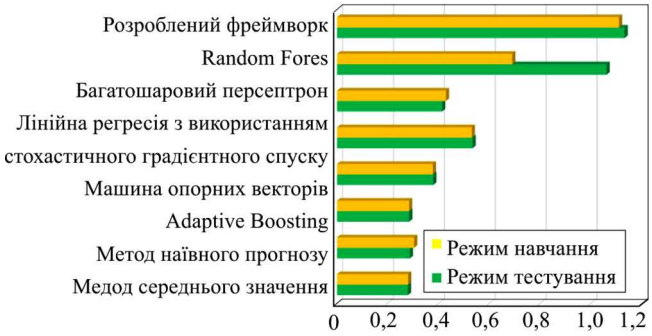

Рис. 4. Результати заповнення пропусків різними методами на основі середньої абсолютної похибки (MAE)

На рис. 3 наведено результати порівняння роботи розробленого фреймворку із найпростішими методами заповнення пропусків: середнім значенням, методом наївного прогнозу та простими методами регресійного моделювання; на основі середньої абсолютної відсоткової похибки. Як видно з рис. 3, найкращі результати, на основі середньої абсолютної відсоткової похибки, отримано саме під час використання розробленого фреймворку. Розроблений алгоритм забезпечує точність розв'язання задачі заповнення пропущених даних у 1,5 раза вище за один з регресійних алгоритмів Adaptive Boosting. Найгірші результати з похибкою більше 60 та 90 % демонструють найпростіші методи заповнення пропусків за допомогою наївного прогнозування та середнього значення відповідно. Хоча методи AdaBoost, машини опорних векторів та лінійної регресії з використанням стохастичного градієнтного спуску резуль-

тують з меншою похибкою, але вона все одно достатньо висока, більше 35, 32 та 31 % відповідно.

На рис. 4 показано результати роботи усіх досліджуваних методів на основі середньої абсолютної похибки. Як видно з цього рисунку, розроблений неітеративний нейроподібний фреймворк демонструє найменшу похибку у режимах як навчання, так і застосування серед усіх методів. У разі використання середньої абсолютної похибки незадовільні результати демонструють методи на основі машини опорних векторів, AdaBoost, наївного прогнозу та метод середнього значення.

Результати на основі середньої абсолютної похибки демонструють таку ж тенденцію. Зокрема, розроблений фреймворк показує похибку RMSE = 0,497323433397, що є найнижчою серед усіх розглянутих методів. Зокрема, у разі визначення середньої абсолютної похибки розроблений нейроподібний фреймворк працює в 1,4 раза ефективніше за SVR і в 1,22 раза краще за AdaBoost алгоритм.

Висновки. Отже, розглянуто та проаналізовано прості методи оброблення та заповнення пропущених параметрів у масивах даних. Хоча застосування простих методів заповнення пропусків може призводити до викривлення статистичних властивостей набору даних (середнє значення, медіана, варіація, кореляція…), такі методи часто використовуються в науковому середовищі, а саме методи регресійного моделювання.

З виконаного дослідження можна зробити висновок про те, що застосування розробленого нейроподібного фреймворку на основі штучних нейронних мереж моделі послідовних геометричних перетворень дає змогу найефективніше виконувати заповнення пропусків, адже забезпечує найточніші результати розв'язання задачі заповнення пропусків у даних екологічного моніторингу на основі похибок MAPE та MAE.

Досліджено, що під час застосування розробленого фреймворку отримуємо найменші похибки при заповненні пропущених параметрів у масивах даних. Наприклад, середньої абсолютної відсоткової похибка становить 20,49 % у режимі навчання, та 22,86 % у тестовому режимі, що найточніше серед всіх застосованих регресійних методів.

Наступні дослідження будуть виконуватися щодо аналізу нових алгоритмів оброблення пропущених параметрів у даних екологічного моніторингу для знаходження мінімальної похибки заповнення пропусків.

Перелік використаних джерел

De Vito, S., Piga, M., Martinotto, L., & Di Francia, G. (2009). CO, NO_2 and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic Bayesian regularization. *Sensors and Actuators B: Chemical, 143*(1), 182–191.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8,* 206–213.

Izonin, I. V., & Tkachenko, R. O. (2019). Komitet neiropodibnykh struktur MPHP z polinomialnym rozshyrenniam vkhodiv dlia zadach Velykykh danykh. In V. S. Ponomarenko (Ed.), *Informatsiina bezpeka ta informatsiini tekhnolohii.* Kharkiv: TOV "DISA PLYuS", 322 p. [In Ukrainian].

Izonin, I. V., Tkachenko, R. O., Peleshko, D. D., & Batiuk, D. A. (2015). Neiromerezhevyi metod zminy rozdilnoi zdatnosti zobrazhen. *Systemy obroblennia informatsii, 9*(134), 30–34. [In Ukrainian].

Karahalios, A., Baglietto, L., Carlin, J. D., English, D. R., & Simpson, J. A. (2012). A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. *BMC Med Res Methodology, 12,* 96 p.

Maltsev, K. A., & Mukharamova, S. S. (2011). *Statystychnyi analiz danykh v ekolohii ta pryrodokorystuvanni.* Kazan: Kazanskyi (Pryvolzhskyi) federalnyi universytet. [In Ukrainian].

Newman, D. (2014). Missing Data: Five Practical Guidelines. *Organizational Research Methods, 17*(4), 372–411.

Shypulin, V. D. (2012). *Osnovni pryntsypy heoinformatsiinykh system.* Kharkiv: KhNAMH, 312 p. [In Ukrainian].

Tkachenko, R. O., Tkachenko, P. R., Izonin, I. V., & Batiuk, D. A. (2017). Metody predvaritelnoi obrabotki izobrazhenii na osnove neiroparadigmy Model geometricheskikh preobrazovanii. *Upravliaiushhie sistemy i mashiny, 1*(267), 59–67. [In Russian].

Tkachenko, R., Cutucu, H., Izonin, I., Doroshenko, A., & Tsymbal, Yu. (2018). Non-iterative Neural-like Predictor for Solar Energy in Libya. In V. Ermolayev, M. C. Suárez-Figueroa, A. Ławrynowicz, R. Palma, V. Yakovyna, H. C. Mayr, M. Nikitchenko, & A. Spivakovsky (Eds), *ICT in Education, Research and Industrial Applications: Proc. 14-th Int. Conf. ICTERI*, May 14–17, 2018. (Vol. 1, pp. 35–45). Kyiv. CEUR-WS.org

Van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman and Hall. CRC, 342 p.

O. S. Mishchuk, R. O. Tkachenko

Lviv Polytechnic National University, Lviv, Ukraine

METHODS OF PROCESSING AND FILLING OF MISSING PARAMETERS IN ECOLOGICAL MONITORING DATA

The variety of sociological, economic, statistical, information and other studies face the problem of processing missing data. Traditional reasons that lead to the emergence of gaps in the data is the inability to obtain information, its distortion or even hiding. In the monitoring environmental pollution data it can be as follows: breakdown of devices; adverse weather conditions; errors of measuring devices; damage to information carriers; suspension of measurements during weekends; implementation of the minimum number of measurements allowed by the state standards. As a result, incomplete information is provided for the analysis of the collected data. Today, there are a large number of methods for recovering missing parameters in the data, but for each application area, different methods are used to fill the missing data. The paper analyzes the following methods for processing missing data: the removal of elements with gaps, the method of weighing and filling missed parameters. The mechanisms of missed parameters appearance are described, in which the probability of gaps for each set of records is the same, in which the probability of gaps is determined on the basis of other available full information and where data is not available depending on unknown factors. There is a need to analyze existing and study new methods for filling missed values in the data sets of environmental monitoring, to find such an algorithm that will maximally satisfy the needs for increasing the speed, efficiency and accuracy of filling out missed parameters. So, authors analyze methods for filling missing parameters in environmental monitoring data such as medium-mean, naive forecast, and regression modeling methods. The article describes the following methods for filling missing data on the basis of regression modeling: multilayer perceptron; Adaptive Boosting; Support vector machine; Random Forest and a linear regression method using stochastic gradient descent. A comparison of the simplest methods of filling missing data and the methods, based on regression models is performed. It has been experimentally proved that the pre-developed method for filling gaps on the basis of the neural-like structure of the model of successive geometric transformations is the most effective method, since it shows the most precise results.

Keywords: missing data; processing of missed values; methods of filling the gaps; regression modeling.