УДК 519.6

## Методы обработки неполных данных геоэкологического мониторинга*

Шарапов Р.В.

В работе проводится анализ методов обработки неполных данных, получаемых при проведении геоэкологического мониторинга. При проведении наблюдений часть данных может отсутствовать вследствие сбоев оборудования, ошибок в проведении исследований, не проведении наблюдений в некоторые периоды и т.д. Кроме того, часть информации может содержать ошибки. Для обнаружения ошибок и заполнения пробелов в наборах данных могут использоваться метод максимального правдоподобия, регрессионный метод, метод главных компонент, пошаговая регрессия, метод многомерной линейной экстраполяции, метод прогностических переменных и т.д. Эти методы хорошо работают для больших наборов данных и известных функций распределения исследуемых величин. При небольших объёмах информации используются эмпирические методы. Для заполнения неполных данных и исправления в них ошибок может использоваться метод моделирования многообразиями малой размерности.

*Ключевые слова:* экзогенные процессы, мониторинг, данные, обработка данных, неполные данные.

## Methods of processing incomplete data for geo-environmental monitoring

Sharapov R.V.

The paper analyzes the methods of incomplete data processing received in the course of geo-environmental monitoring. In the surveying process the part of data can be missed as a result of hardware failures, errors in research conducting, or when observations fail to be made in some periods, etc. Furthermore, the maximum likelihood method, the regression method, the principal component analysis, the stepwise regression, multivariate linear extrapolation method, the method of predictive variables can be used for error detecting and filling in the gaps of data sets. These methods work fine for large data sets and known distribution functions of the values in question. Empirical methods can be used for processing small amounts of information. The method of modeling low-dimensional manifolds is applied for filling in the incomplete data and correcting its errors.

*Keywords:* exogenous processes, monitoring, data, data processing, incomplete data.

### Введение

При проведении геоэкологического мониторинга не всегда получается осуществить весь задуманный объём работ и избежать ошибок. Иногда часть данных наблюдений за тем или иным объектом или процессом может отсутствовать вследствие сбоев оборудования, ошибок в проведении исследований, потери или порчи образцов и т.д. Кроме того, исследование определенных показателей вообще может не проводится в те или иные периоды времени. В полученных при осуществлении мониторинга данных могут содержаться различного рода ошибки, связанные с воздействием помех, погрешностями измерений, сбоями в работе оборудования, человеческим фактором и т.д. С ростом объёмов исследований количество пробелов и ошибок в данных наблюдений только увеличивается.

Геоэкологический мониторинг характеризуется большим числом наблюдаемых объектов и вполне определенным набором исследуемых параметров. Несмотря на отсутствие полноты данных и наличие ошибок, для оценки состояния территории важное значение имеет использование всей доступной информации (что особо актуально, например, при мониторинге экзогенных процессов). Это делает актуальной задачу обработки неполных данных.

***Работа выполнена при поддержке гранта РФФИ № 13-07-97510 р_центр_а.**

Таким образом, при проведении геоэкологического мониторинга возникает необходимость решения следующих задач [1, 2]:

1. Проведение обработки неполных данных мониторинга;
2. Заполнение пробелов в данных статистически достоверными или правдоподобными значениями;
3. Обнаружение и по возможности исправление ошибок в имеющихся данных.

Цель работы – рассмотреть существующие методы обработки данных геоэкологического мониторинга, имеющих пробелы и содержащих ошибки.

**Методы обработки неполных данных**

В ранних работах [3] проводилась оценка максимального правдоподобия неполных (фрагментарных) данных. При отсутствии тех или иных данных параметры или заменялись на среднее значение, или вовсе не учитывались [4]. В дальнейшем были предложены более сложные алгоритмы, основанные на методе наименьших квадратов. В [5, 6] предлагается использовать регрессионный метод. В работе [7] предложено использовать метод главных компонент. В [8] для обработки недостающих данных предлагается использовать пошаговую регрессию. В [9] предлагается использовать метод многомерной линейной экстраполяции. В работе [10] рассматриваются вопросы применения метода прогностических переменных. Работы [11, 12] посвящены оценке ковариационной матрицы на основе неполных данных.

В работе [13] предложен EM-алгоритм для решения общей задачи оценивания значений параметров наблюдения по неполным данным. Работа [14] посвящена повышению эффективности EM-алгоритма.

Подробный обзор статистических методов анализа и обработки неполных данных можно найти в [15].

Изложенные методы хорошо работают для больших наборов данных и известных функций распределения исследуемых величин.

В [16] предложен алгоритм «ZET», в основу которого положены эмпирические методы работы с неполными данными. Алгоритм изучает «похожесть» объектов. На основе наиболее связанной с неизвестным элементом информации строится предсказывающая подтаблица, анализ которой с использованием принципа локальной линейности позволяет спрогнозировать оценку недостающего значения.

В [17] изложен метод транспонированной регрессии. Пусть имеется некоторый набор объектов, каждый из которых характеризуется совокупностью признаков. Тогда можно составить матрицу, в которой строки будут соответствовать объектам, а столбцы – признакам объектов. Рассматривается гипотеза, что значения признака одного объекта могут быть функциями значений этого же признака других объектов, причем эти функции они и те же для всех признаков. Получается транспонированная задача регрессии, которая в отличие от исходной инвариантна к смене шкалы измерений. Недостаточно большое количество признаков по сравнению с числом объектов вынуждает для каждого объекта искать небольшую опорную группу, по признакам которых можно восстановить характеристики данного объекта.

Вводятся следующие выражения для представления вектора свойств (признаков) объекта:

$$\tilde{\boldsymbol{y}} = \sum_{i=1}^{q} \alpha_i \boldsymbol{y}^i, \qquad \sum_{i=1}^{q} \alpha_i = 1$$

где $\tilde{\boldsymbol{y}}$ – вектор свойств объекта с восстановленными недостающими значениями, $\boldsymbol{y}^i$ – вектор свойств $i$-го объекта опорной группы, $\alpha_i$ –коэффициенты разложения, $q$ – мощность опорной группы.

Решение можно представить в виде [17]:

$$\tilde{\boldsymbol{y}} = \boldsymbol{m_y} + \sum_{i=1}^{q} \beta_i (\boldsymbol{y}^i - \boldsymbol{m_y})$$

$$\alpha_i = \beta_i + \frac{1}{q} - \frac{1}{q} \sum_{k=1}^{q} \beta_k$$

где $\beta_i$, $\beta_k$ – коэффициенты разложения опорной группы,

$$\boldsymbol{m_y} = \frac{1}{q} \sum_{i=1}^{q} \boldsymbol{y}^i$$

Для вычисления выражений в [17] предлагается использовать нейронные сети.

В работе [18] приводится метод моделирования неполных данных многообразиями малой размерности. Суть метода состоит в следующем. По аналогии с [17] строится матрица «объект-признак». Ей соответствует матрица $A = (a_{ij})$, где $a_{ij}$ – *j*-ое свойство *i*-го объекта. Каждый объект описывается вектором его параметров $a_i$. Вектор может содержать *k* пробелов (отсутствующих данных). Тогда его можно представить как *k*-мерное линейное многообразие $L_a$, параллельное *k* координатным осям, соответствующим недостающим данным. Производится поиск многообразия *M* малой размерности, наилучшим образом приближающего данные. Для полных данных точность приближения определяется, как расстояние от точки до множества. Для неполных данных используется нижняя грань расстояний между точками *M* и $L_x$. Далее из данных итерационно вычитаются ближайшие к ним точки многообразия *M* до тех пор, пока остатки не приблизятся к нулю.

Авторы [18] предлагают использовать многообразия малой размерности для заполнения пробелов в данных, а также исправления («ремонта») данных, содержащих ошибки.

## Заключение

Таким образом, к настоящему времени разработанно достаточно много методов обработки неполных данных.

Для обнаружения ошибок и заполнения пробелов в наборах данных могут использоваться метод максимального правдоподобия, регрессионный метод, метод главных компонент, пошаговая регрессия, метод многомерной линейной экстраполяции, метод прогностических переменных и т.д. Эти методы хорошо работают для больших наборов данных и известных функций распределения исследуемых величин. При небольших объёмах информации используются эмпирические методы. Для заполнения неполных данных и исправления в них ошибок может использоваться метод моделирования многообразиями малой размерности.

## Литература

1. *Шарапов Р.В.* О согласовании данных мониторинга экзогенных процессов, полученных из разнородных источников // Машиностроение и безопасность жизнедеятельности, 2013, № 4. – С. 43-46.

2. *Шарапов Р.В.* Некоторые вопросы мониторинга экзогенных процессов // Фундаментальные исследования, 2013, № 1-2. – С. 444-447.

3. *Wilks S.S.* Moments and distributions of estimates of population from fragmentary samples // Annals of Mathematical Statistics, 1932, vol.3. – P. 163-195.

4. *Afifi A.A., Elashoff R.M.* Missing observations in multivariate statistics // Journal of the American Statistical Association, 1966, vol. 61. – P. 595-604.

5. *Buck S.F.* A method of estimation of missing values in multivariate data // Journal of the Royal Statistical Society Series B. 1960, vol. 22. – P. 202-206.

6. *Walsh J.E.* Computer-feasible method for handling incomplete data in regression analysis // Journal of ACM, 1961, vol. 18. – P. 201-211.

7. *Gleason T.C., Staelin R.* A proposal for handling missing data // Psychometrika, 1975, vol. 40. – P. 229-252.

8. *Frane G.M.* Some simple procedures for handling missing values in multivariate analysis // Psychometrika, 1976, vol. 41. – P. 409-415.

9. *Растригин Л.А., Пономарев Ю.П.* Экстраполяционные методы проектирования и управления. – М.: Машиностроение, 1986. – 120 с.

10. *Жанатаусов С.У.* Методы прогностических переменных. Машинные методы обнаружения закономерностей – Новосибирск: 1981, вып. 88, Вычислительные системы. – С. 151-155.

11. *Engelman L.* An efficient algorithm for computing covariance matrices from data with missing values // Communications in Statistics Part B., 1982, vol. 11. – P. 113-121.

12. *Huseby J.R., Schwertman N.C., Allen D.M.* Computation of the mean vector and dispersion matrix for incomplete multivariate data // Communications in Statistics Part B., 1980, vol. 9. – P. 301-309.

13. *Dempster A.P., Laird N.M., Rubin D.B.* Maximum likelihood from incomplete data via the EM-algorithm // Journal of the Royal Statistical Society Series B, 1977, vol. 39. – P. 1-38.

14. *Little R.J., Smith P.J.* Editing and imputation for quantitative survey data // Journal of the American Statistical Association, 1987, vol. 82. – P. 58-68.

15. *Little R.J., Rubin D. B.* Statistical Analysis with Missing Data. – Chichester, John Wiley & Sons, 1987. – 278 p.

16. *Загоруйко Н.Г., Ёлкина В.Н., Тимеркаев В.С.* Алгоритм заполнения пропусков в эмпирических таблицах (алгоритм ZET) // Вычислительные системы. Новосибирск, 1975. Вып. 61. Эмпирическое предсказание и распознавание образов. – С. 3-27.

17. *Горбань А.Н., Новоходько Ю.А., Царегородцев В.Г.* Нейросетевая реализация транспонированной задачи линейной регрессии // Нейроинформатика и ее приложения. Тез. докл. IV Всерос. семинара, Красноярск, 5-7 окт. 1996 г. – 1996. – С. 37–39.

18. *Gorban A.N., Rossiev A.A.* Neural network iterative method of principal curves for data with gaps // Journal of Computer and Systems Sciences International. 1999. Т. 38. № 5. С. 825-830.

## References

1. *Sharapov R.V.* O soglasovanii dannyh monitoringa jekzogennyh processov, poluchennyh iz raznorodnyh istochnikov [On the conformity of monitoring data obtained from various sources] // Mashinostroenie i bezopasnost' zhiznedejatel'nosti [Engineering industry and life safety], 2013, № 4. – P.43-46.

2. *Sharapov R.V.* Nekotorye voprosy monitoringa jekzogennyh processov [Some problems of exogenous processes monitoring] // Fundamental'nye issledovanija [Fundamental research], 2013, № 1-2. – P. 444-447

3. *Wilks S.S.* Moments and distributions of estimates of population from fragmentary samples // Annals of Mathematical Statistics, 1932, vol.3. – P. 163-195.

4. *Afifi A.A., Elashoff R.M.* Missing observations in multivariate statistics // Journal of the American Statistical Association, 1966, vol. 61. – P. 595-604.

5. *Buck S.F.* A method of estimation of missing values in multivariate data // Journal of the Royal Statistical Society Series B. 1960, vol. 22. – P. 202-206.

6. *Walsh J.E.* Computer-feasible method for handling incomplete data in regression analysis // Journal of ACM, 1961, vol. 18. – P. 201-211.

7. *Gleason T.C., Staelin R.* A proposal for handling missing data // Psychometrika, 1975, vol. 40. – P. 229-252.

8. *Frane G.M.* Some simple procedures for handling missing values in multivariate analysis // Psychometrika, 1976, vol. 41. – P. 409-415.

9. *Rastrigin L.A., Ponamarev Yu.P.* Jekstrapoljacionnye metody proektirovanija i upravlenija [Extrapolation methods, design and management]. – Moscow: Mashinostroenie, 1986. – 120 p.

10. *Zhanatausov S.U.* Metody prognosticheskih peremennyh. Mashinnye metody obnaruzhenija zakonomernostej [Methods prognostic variables. Machine methods of detection patterns] – Novosibirsk, 1981, vol. 88, Computer systems. – P. 151-155.

11. *Engelman L.* An efficient algorithm for computing covariance matrices from data with missing values // Communications in Statistics Part B., 1982, vol. 11. – P. 113-121.

12. *Huseby J.R., Schwertman N.C., Allen D.M.* Computation of the mean vector and dispersion matrix for incomplete multivariate data // Communications in Statistics Part B., 1980, vol. 9. – P. 301-309.

13. *Dempster A.P., Laird N.M., Rubin D.B.* Maximum likelihood from incomplete data via the EM-algorithm // Journal of the Royal Statistical Society Series B, 1977, vol. 39. – P. 1-38.

14. *Little R.J., Smith P.J.* Editing and imputation for quantitative survey data // Journal of the American Statistical Association, 1987, vol. 82. – P. 58-68.

15. *Little R.J., Rubin D. B.* Statistical Analysis with Missing Data. – Chichester, John Wiley & Sons, 1987. – 278 p.

16. *Zagorujko N.G., Elkina V.N., Timerkaev V.S.* Algoritm zapolnenija propuskov v jempiricheskih tablicah (algoritm ZET) [Algorithm fill the gaps in empirical tables (algorithm ZET)] // Vychislitel'nye sistemy [Computer systems]. Novosibirsk, 1975. vol. 61. Jempiricheskoe predskazanie i raspoznavanie obrazov [Empirical prediction and pattern recognition]. – P. 3-27.

17. *Gorban A.N., Novohodko Yu.A., Tsaregorodtsev V.G.* Nejrosetevaja realizacija transponirovannoj zadachi linejnoj regressii [Neural implementation of the transposed linear regression problem] // Nejroinformatika i ee prilozhenija. Tes. docl. IV Vseross. seminara [Proceedings of Neuroinformatics and its applications], Krasnoyarsk, 5-7 October 1996. – P. 37–39.

18. *Gorban A.N., Rossiev A.A.* Neural network iterative method of principal curves for data with gaps // Journal of Computer and Systems Sciences International. 1999. T. 38. № 5. C. 825-830.

**Статья поступила в редакцию 22 февраля 2014 г.**

*Шарапов Руслан Владимирович* – кандидат технических наук, заведующий кафедрой «Техносферная безопасность» Муромского института (филиала) федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых», г. Муром, Россия. E-mail: info@vanta.ru

*Sharapov Ruslan Vladimirovich* – Ph.D., Murom Institute of Vladimir State University, Murom, Russia. E-mail: info@vanta.ru