УДК 519.6

Приближенный алгоритм моделирования стационарных дискретных случайных процессов с двумерными распределениями последовательных компонент в виде смеси гауссовских распределений*

В.А. Огородников[1,2], М.С. Акентьева[1], Н.А. Каргаполова[1,2]

[1]Институт вычислительной математики и математической геофизики Сибирского отделения Российской академии наук, просп. Акад. Лаврентьева, 6, Новосибирск, 630090

[2]Новосибирский национальный исследовательский государственный университет (НГУ), ул. Пирогова, 2, Новосибирск, 630090

E-mails: ova@osmf.sscc.ru (Огородников В.А.), akenteva@sscc.ru (Акентьева М.С.), nkargapolova@sscc.ru (Каргаполова Н.А.)

Английская версия этой статьи печатается в журнале "Numerical Analysis and Applications" № 2, Vol. 17, 2024.

Огородников В.А., Акентьева М.С., Каргаполова Н.А. Приближенный алгоритм моделирования стационарных дискретных случайных процессов с двумерными распределениями последовательных компонент в виде смеси гауссовских распределений // Сиб. журн. вычисл. математики / РАН. Сиб. отд-ние. — Новосибирск, 2024. — Т. 27, № 2. — С. 211–216.

В работе представлен приближенный алгоритм моделирования стационарного дискретного случайного процесса с одномерными и двумерными распределениями его последовательных компонент в виде смеси двух гауссовских распределений. Алгоритм основан на комбинации метода условных распределений и метода исключения. Приведен пример применения алгоритма для моделирования временных рядов максимальной за сутки температуры воздуха.

DOI: 10.15372/SJNM20240206
EDN: ZIXSFU

Ключевые слова: *стохастическое моделирование, двумерное распределение, смесь нормальных распределений, максимальная температура воздуха.*

Ogorodnikov V.A., Akenteva M.S., Kargapolova N.A. An approximate algorithm for simulating stationary discrete random processes with bivariate distributions of their consecutive components in the form of mixtures of Gaussian distributions // Siberian J. Num. Math. / Sib. Branch of Russ. Acad. of Sci. — Novosibirsk, 2024. — Vol. 27, № 2. — P. 211–216.

The paper presents an approximate algorithm for modeling a stationary discrete random process with marginal and bivariate distributions of its consecutive components in the form of a mixture of two Gaussian distributions. The algorithm is based on a combination of the conditional distribution method and the rejection method. An example of application of the proposed algorithm for simulating time series of daily maximum air temperatures is given.

*Исследование выполнено в рамках государственного задания ИВМиМГ СО РАН (проект № FWNM-2022-0002).

© В.А. Огородников, М.С. Акентьева, Н.А. Каргаполова, 2024

Keywords: *stochastic simulation, bivariate distribution, mixture of Gaussian distributions, maximum daily temperature.*

1. Введение

Одними из наиболее часто моделируемых в стохастических "генераторах погоды" метеорологических характеристик являются минимальная и максимальная за сутки температуры воздуха [1–3]. Согласно [4], в большинстве генераторов погоды, в том числе в USCLIMATE, WXGEN, LARS-WG, CLIMGEN и CLIGEN, использовано предположение о том, что одномерные распределения минимальной и максимальной температур являются нормальными. Однако использование этого предположения не всегда оправдано — часто одномерные распределения оказываются асимметричными [4]. В качестве иллюстрации на рисунке приведена гистограмма одномерного распределения максимальной за сутки температуры в Мамакане (Иркутская область, Россия). Для того, чтобы учесть негауссовость распределений вышеуказанных метеохарактеристик, в работе [5] предложено использовать алгоритм моделирования случайных величин с двумерным косонормальным распределением.

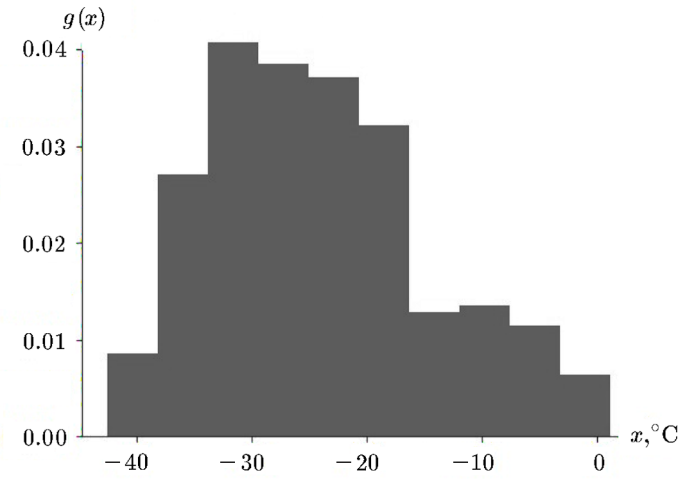

Рис. Гистограмма распределения максимальной за сутки температуры в Мамакане (Иркутская область, Россия), построенная с использованием данных реальных метеонаблюдений в период с 1 по 10 декабря с 1991 по 2021 гг.

Проверка гипотез о виде одномерных и двумерных распределений минимальной и максимальной за сутки температур воздуха на метеостанциях, расположенных на Байкальской природной территории и в прилегающих к ней районах, позволяет предположить, что эти распределения представляют собой смеси двух гауссовских распределений. В связи с этим в данной работе мы предлагаем алгоритм моделирования случайных векторов $\vec{\xi} = (\xi_1, \ldots, \xi_N)$, у которых одномерное распределение компонент ξ_i, $i = 1, 2, \ldots, N$, и двумерное распределение пар (ξ_i, ξ_{i+1}), $i = 1, 2, \ldots, N-1$, суть смеси двух не зависящих от i одномерных и двумерных нормальных распределений соответственно.

2. Алгоритм моделирования

Введем следующие обозначения. Пусть $\vec{x} = (x_1, x_2)^\top$,

$$f(x_1, x_2) = p\, f_1(x_1, x_2) + (1-p)\, f_2(x_1, x_2)$$
$$= p\, \frac{1}{2\pi\sqrt{|\Sigma_1|}}\, e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_1)^\top \Sigma_1^{-1}(\vec{x}-\vec{\mu}_1)} + (1-p)\, \frac{1}{2\pi\sqrt{|\Sigma_2|}}\, e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_2)^\top \Sigma_2^{-1}(\vec{x}-\vec{\mu}_2)}$$

есть плотность двумерного распределения, имеющая вид смеси двух нормальных распределений с параметрами:

$$\vec{\mu}_k = \begin{pmatrix} \mu_{k1} \\ \mu_{k2} \end{pmatrix}, \qquad \Sigma_k = \begin{pmatrix} \sigma_{k1}^2 & c_k \\ c_k & \sigma_{k2}^2 \end{pmatrix}, \qquad k = 1, 2.$$

Здесь p, $0 \leqslant p \leqslant 1$, — вес, определяющий смесь, $|\Sigma_k|$ — определитель ковариационной матрицы Σ_k. Для того, чтобы вектор $\vec{\xi}$ с двумерным распределением $f(x_1, x_2)$ пар (ξ_i, ξ_{i+1}), $i = 1, 2, \ldots, N-1$, был стационарен в широком смысле необходимо выполнение следующих условий:

$$\mu_{k1} = \mu_{k2} = m_k, \qquad \sigma_{k1} = \sigma_{k2} = s_k, \qquad k = 1, 2.$$

В этом случае параметры нормальных распределений принимают вид

$$\vec{\mu}_k = \begin{pmatrix} m_k \\ m_k \end{pmatrix}, \qquad \Sigma_k = \begin{pmatrix} s_k^2 & c_k \\ c_k & s_k^2 \end{pmatrix}, \qquad k = 1, 2.$$

Сформулируем алгоритм моделирования вектора $\vec{\xi}$.

Алгоритм.

1. Моделируем реализацию $(\tilde{\eta}_1, \tilde{\eta}_2)^\top$ случайного вектора $(\eta_1, \eta_2)^\top$ с плотностью двумерного распределения $f(x_1, x_2)$. Полагаем $\xi_1 = \tilde{\eta}_1$, $\xi_2 = \tilde{\eta}_2$.
2. Независимо от ξ_1, ξ_2 моделируем реализации $(\tilde{\eta}_2, \tilde{\eta}_3)^\top$ вектора $(\eta_2, \eta_3)^\top$ с плотностью двумерного распределения $f(x_1, x_2)$ до тех пор, пока не выполнится условие $|\xi_2 - \tilde{\eta}_2| < \varepsilon$, где $\varepsilon > 0$ — предварительно заданное произвольное достаточно малое число. При выполнении условия $|\xi_2 - \tilde{\eta}_2| < \varepsilon$ полагаем $\xi_3 = \tilde{\eta}_3$.
3. Последовательно для всех i, $i < N$, повторяем процедуру из пункта 2 данного алгоритма, моделируя реализации $(\tilde{\eta}_{i-1}, \tilde{\eta}_i)^\top$ векторов $(\eta_{i-1}, \eta_i)^\top$ и полагая $\xi_i = \tilde{\eta}_i$ при выполнении условия $|\xi_{i-1} - \tilde{\eta}_{i-1}| < \varepsilon$.

В данном алгоритме моделирование векторов с плотностью $f(x_1, x_2)$ целесообразно проводить, используя метод дискретной суперпозиции (см. [6, раздел 1.6.2]) вместе с известной техникой компьютерного моделирования двумерного гауссовского случайного вектора с заданными корреляционными характеристиками (см. [6, раздел 1.10.4]). Поскольку аналитическое выражение для плотности $f_{\eta_i}(u \mid \eta_{i-1} = v)$ условного распределения компоненты η_i вектора $(\eta_{i-1}, \eta_i)^\top$ при фиксированном значении компоненты η_{i-1} неизвестно, то в пунктах 2 и 3 алгоритма происходит отбор значений η_{i-1}, удовлетворяющих неравенству $v - \varepsilon < \eta_{i-1} < v + \varepsilon$, на основании допредельной версии $f_{\eta_i}(u \mid v - \varepsilon < \eta_{i-1} < v + \varepsilon) \approx f_{\eta_i}(u \mid \eta_{i-1} = v)$ известного равенства $f_{\eta_i}(u \mid \eta_{i-1} = v) =$

$f_{\eta_i}(u \mid \eta_{i-1} \in dv)$ (см. [7, гл. 4, § 9]), где dv — бесконечно малая окрестность точки v. О точности моделирования и ее зависимости от ε можно судить, например, по различию между моментами распределения с плотностью $f(x_1, x_2)$ и соответствующими выборочными значениями моментов, оцененных по модельным реализациям.

3. Моделирование временных рядов максимальной за сутки температуры воздуха

В этом пункте приведен пример применения вышеприведенного алгоритма для численного моделирования временных рядов $\vec{\xi} = (\xi_1, \ldots, \xi_{10})$ максимальной за сутки температуры воздуха на 10-дневном интервале.

В качестве реальных данных использованы данные наблюдений с 1991 по 2021 гг. на метеостанциях, расположенных на Байкальской природной территории и в прилегающих к ней районах [8]. Оценка параметров плотности двумерного распределения $f(x_1, x_2)$, т. е. параметров p, $\vec{\mu}_k$, Σ_k, $k = 1, 2$, проводилась с помощью EM-алгоритма [9]. Напомним, что EM-алгоритм применяется для получения оценок максимального правдоподобия неизвестных параметров статистических моделей и часто используется для разделения смесей распределений. На каждой итерации EM-алгоритма чередуются два шага: E-шаг и M-шаг. На E-шаге формируется функция $Q(\Theta)$ — математическое ожидание логарифма функции правдоподобия, вычисленное на основе текущих оценок множества параметров Θ. На M-шаге определяются параметры $\Theta_{\max}$, максимизирующие $Q(\Theta)$. В нашем случае $\Theta = \{p, m_1, s_1, c_1, m_2, s_2, c_2\}$.

Входными параметрами предложенной модели являются величины p, m_1, s_1, c_1, m_2, s_2, c_2, оцениваемые по реальным данным. Как было указано выше, о точности моделирования можно судить по различию между моментами распределения $f(x_1, x_2)$, вычисленными уже по определенным значениям p, m_1, s_1, c_1, m_2, s_2, c_2, и соответствующими выборочными значениями моментов, оцененных по модельным реализациям. Иными словами, сравнение моментов теоретического и выборочного двумерных распределений можно использовать для проверки качества модели. В качестве иллюстрации в таблицах 1 и 2 приведены оценки математического ожидания и среднеквадратического отклонения вектора $\vec{\xi}$, полученные по реальным данным и оцененные по модельным траекториям. Для оценок по реальным данным приведены также значения соответствующих статистических погрешностей σ.

Таблица 1. Оценки математического ожидания $E\xi_i$ компонент вектора $\vec{\xi}$ по реальным данным, собранным с 1 по 10 декабря 1991–2021 гг., и по 10^4 модельным траекториям

Метеостанция	Оценка по реальным данным, $E\xi_i \pm \sigma$	Оценка по модельным данным
Червянка	-16.24 ± 1.40	-16.19
Мамакан	-24.04 ± 1.41	-23.96

Таблица 2. Оценки среднеквадратического отклонения $\sqrt{D\xi_i}$ компонент вектора $\vec{\xi}$ по реальным данным, собранным с 1 по 10 декабря 1991–2021 гг., и по 10^4 модельным траекториям

Метеостанция	Оценка по реальным данным, $\sqrt{D\xi_i} \pm \sigma$	Оценка по модельным данным
Червянка	10.61 ± 0.55	10.61
Мамакан	9.58 ± 0.71	9.50

Очевидно, что в общем случае корреляционные функции реального и модельного процессов не совпадают. По построению должны быть близкими только значения коэффициентов корреляции $\mathrm{corr}\,(\xi_i, \xi_{i+1})$. В табл. 3 приведены оценки коэффициентов корреляции $\mathrm{corr}\,(\xi_i, \xi_{i+1})$ по реальным и модельным данным.

Таблица 3. Оценки коэффициентов корреляции $\mathrm{corr}\,(\xi_i, \xi_{i+1})$ по реальным данным, собранным с 1 по 10 декабря 1991–2021 гг., и модельным траекториям

Метеостанция	Оценка по реальным данным, $\mathrm{corr}\,(\xi_i, \xi_{i+1}) \pm \sigma$	Оценка по модельным траекториям
Червянка	0.72 ± 0.05	0.76
Мамакан	0.78 ± 0.04	0.85

Во всех проведенных численных экспериментах разница между оценками математического ожидания, среднеквадратического отклонения и коэффициента корреляции по реальным и модельным данным не превышала соответствующих значений 2σ при $\varepsilon = 0.1$. Такое значение ε позволяет, с одной стороны, с достаточной точностью воспроизводить основные характеристики временного ряда, а с другой — иметь приемлемое время его моделирования.

4. Заключение

В работе представлен приближенный алгоритм моделирования стационарного дискретного случайного процесса с одномерными и двумерными распределениями его последовательных состояний в виде смеси двух гауссовских распределений. Алгоритм основан на комбинации метода дискретной суперпозиции, технологии компьютерного моделирования двумерного гауссовского случайного вектора с заданными корреляционными характеристиками и специального отбора пар соседних состояний с нужной условной плотностью распределения. Приведен пример применения алгоритма для моделирования временных рядов максимальной за сутки температуры воздуха.

Предложенный в работе алгоритм допускает обобщение на случай моделирования векторных рядов с аналогичными свойствами двумерных распределений. В частности, он может быть применен для моделирования совместных временных рядов минимальной и максимальной температур воздуха. В этом случае применение отбора для моделирования необходимых условных реализаций требует существенно больших компьютерных затрат, но, тем не менее, алгоритм может быть реализован за приемлемое время. Для сокращения времени моделирования при численной реализации алгоритма можно также использовать технологии параллельного программирования.

Литература

1. **Kleiber W., Katz R.W., Rajagopalan B.** Daily minimum and maximum temperature simulation over complex terrain // Ann. Appl. Stat. — 2013. — Vol. 7, № 1. — P. 588–612. — DOI: 10.1214/12-AOAS602.

2. **Meng X., Taylor J.W.** Comparing probabilistic forecasts of the daily minimum and maximum temperature // Intern. J. Forecasting. — 2022. — Vol. 38, № 1. — P. 267–281.

3. **Richardson C.W.** Stochastic simulation of daily precipitation, temperature and solar radiation // Water Resour. Res. — 1981. — Vol. 17, № 1. — P. 182–190.

4. **Harmel R.D., Richardson C.W., Hanson C.L., Johnson G.L.** Simulation maximum and minimum daily temperature with the normal distribution // ASAE Annual Meeting. — 2001. — Article № 012240.
5. **Flecher C., Naveau P., Allard D., Brisson N.** A stochastic daily weather generator for skewed data // Water Resour. Res. — 2010. — Vol. 46, № 7. — Article № W07519. — DOI: 10.1029/2009wr008098.
6. **Михайлов Г.А., Войтишек А.В.** Численное статистическое моделирование. Методы Монте-Карло. — М.: Изд. центр "Академия", 2006.
7. **Боровков А.А.** Теория вероятностей. — М.: Наука, 1986.
8. **Булыгина О.Н., Веселов В.М., Разуваев В.Н., Александрова Т.М.** Описание массива срочных данных об основных метеорологических параметрах на станциях России. — http://meteo.ru/data/163-basic-parameters. — (Свидетельство о государственной регистрации базы данных № 2014620549).
9. **Ghojogh B., Ghojogh A., Crowley M., Karray F.** Fitting A Mixture Distribution to Data: Tutorial. — 2019. — arXiv:1901.06708.

Поступила в редакцию 31 января 2024 г.
После исправления 5 февраля 2024 г.
Принята к печати 4 марта 2024 г.

Литература в транслитерации

1. **Kleiber W., Katz R.W., Rajagopalan B.** Daily minimum and maximum temperature simulation over complex terrain // Ann. Appl. Stat. — 2013. — Vol. 7, № 1. — P. 588–612. — DOI: 10.1214/12-AOAS602.
2. **Meng X., Taylor J.W.** Comparing probabilistic forecasts of the daily minimum and maximum temperature // Intern. J. Forecasting. — 2022. — Vol. 38, № 1. — P. 267–281.
3. **Richardson C.W.** Stochastic simulation of daily precipitation, temperature and solar radiation // Water Resour. Res. — 1981. — Vol. 17, № 1. — P. 182–190.
4. **Harmel R.D., Richardson C.W., Hanson C.L., Johnson G.L.** Simulation maximum and minimum daily temperature with the normal distribution // ASAE Annual Meeting. — 2001. — Article № 012240.
5. **Flecher C., Naveau P., Allard D., Brisson N.** A stochastic daily weather generator for skewed data // Water Resour. Res. — 2010. — Vol. 46, № 7. — Article № W07519. — DOI: 10.1029/2009wr008098.
6. **Mikhailov G.A., Voitishek A.V.** Chislennoe statisticheskoe modelirovanie. Metody Monte-Karlo. — M.: Izd. centr "Akademiya", 2006.
7. **Borovkov A.A.** Teoriya veroyatnostei. — M.: Nauka, 1986.
8. **Bulygina O.N., Veselov V.M., Razuvaev V.N., Aleksandrova T.M.** Opisanie massiva srochnykh dannykh ob osnovnykh meteorologicheskikh parametrakh na stanciyakh Rossii. — http://meteo.ru/data/163-basic-parameters. — (Svidetel'stvo o gosudarstvennoi registracii bazy dannykh № 2014620549).
9. **Ghojogh B., Ghojogh A., Crowley M., Karray F.** Fitting A Mixture Distribution to Data: Tutorial. — 2019. — arXiv:1901.06708.