Интеллектуальные системы управления, анализ данных

© 2021 г. Ю.А. ДУБНОВ (yury.dubnov@phystech.edu)
(Институт системного анализа Федерального исследовательского центра "Информатика и управление" РАН, Москва;
Национальный исследовательский университет "Высшая школа экономики", Москва;
Московский физико-технический институт),
В.Ю. ПОЛИЩУК, канд. техн. наук (liquid_metal@mail.ru)
(Институт мониторинга климатических и экологических систем СО РАН;
Национальный исследовательский Томский политехнический университет),
Ю.С. ПОПКОВ, д-р техн. наук (popkov@isa.ru)
(Институт системного анализа Федерального исследовательского центра "Информатика и управление" РАН, Москва;
Брауде Колледж университета Хайфы, Кармиель, Израиль;
Институт проблем управления им. В.А. Трапезникова РАН, Москва),
Ю.М. ПОЛИЩУК, д-р физ.-мат. наук (yupolishchuk@gmail.com),
А.В. МЕЛЬНИКОВ, д-р техн. наук (melnikovav@uriit.ru),
Е.С. СОКОЛ (eugen137@gmail.com)
(Югорский НИИ информационных технологий, Ханты-Мансийск)

МЕТОД ЭНТРОПИЙНО-РАНДОМИЗИРОВАННОГО ВОССТАНОВЛЕНИЯ ПРОПУЩЕННЫХ ДАННЫХ[1]

Статья посвящена проблеме восстановления пр́опусков в коллекциях данных для задач машинного обучения. Предложен новый рандомизированный метод восстановления пропущенных данных, основанный на технологии энтропийно-робастного оценивания и генерации ансамблей случайных величин. Предложенный метод схож с использованием вспомогательной регрессии для восстановления пропущенных значений, но в отличие от последней в случае энтропийного оценивания не накладываются дополнительные ограничения на функцию правдоподобия ошибок в выборке и допустимы малые объемы данных, что становится крайне актуальным в задачах, когда объем данных для обучения ограничен, а пропуски встречаются не систематически. Предложенный метод применяется для восстановления пропущенных данных о площадях термокарстовых озер арктической зоны РФ, измеряемых по спутниковым снимкам.

Ключевые слова: восстановление пропусков, энтропийное оценивание, рандомизированное машинное обучение, термокарстовые озера, Арктика.

DOI: 10.31857/S0005231021040061

[1] Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проекты №№ 19-07-00282, 20-07-00223). Работа выполнена в рамках госбюджетной темы.

1. Введение

В прикладных задачах анализа данных нередко возникает необходимость работы с выборками данных, содержащих пропущенные значения. Пропуски в данных могут быть следствием самых различных причин, от отказа респондентов отвечать на вопросы анкеты до технических сбоев измерительных датчиков или программного обеспечения. Наличие пропусков в данных приводит к искажению результатов анализа и, следовательно, требует дополнительной обработки. Кроме того, большинство современных алгоритмов машинного обучения не поддерживают наличие пропусков в данных, что приводит к необходимости восстановления пропущенных значений.

Проблеме неполноты данных посвящено множество публикаций по статистике [1–4], и в различных сферах прикладного анализа данных, например в медицине [5], образовании [6] и др. [7, 8]. В частности, в [5] приводится один из наиболее полных обзоров современных методов анализа в условиях наличия пропусков в данных, книга содержит подробную теорию и примеры реализации и тестирования методов для клинических данных.

Различают три механизма появления пропусков в данных [2]:

- *Полностью случайные пропуски* (Missing Completely At Random, MCAR) — механизм, при котором вероятность появления пропущенных значений одинакова для всех объектов выборки. Например, если на предприятии не измеряют характеристики некоторых случайно выбранных устройств;
- *Случайные пропуски* (Missing At Random, MAR) — механизм, при котором вероятность появления пропусков может быть выявлена на основе имеющейся в выборке информации. Например, если не измеряются характеристики устройств с определенными критериями по весу, габаритам и пр.;
- *Неслучайные пропуски* (Missing Not At Random, MNAR) — в этом случае наличие пропусков в данных является дополнительной характеристикой объекта и не может игнорироваться при анализе. Например, если оказывается невозможным измерить характеристики уже вышедших из строя устройств.

Существует два основных подхода к анализу неполных данных. Первый подход подразумевает *отбрасывание пропусков*, полное или частичное (Listwise/pairwise deletion), в зависимости от требований алгоритма анализа [9]. Это наиболее простой в реализации способ, но применяемый исключительно для полностью случайных (MCAR) пропусков, когда исключение одного или нескольких объектов не приводит к нарушению репрезентативности выборки. Кроме того, данный способ не подойдет для выборок малых объемов и выборок с высокой долей пропусков.

Второй подход предполагает восстановление пропущенных значений, т.е. вставку (импутирование) некоторых значений на места пропусков. Согласно классификации, приведенной в [1], различают простые и сложные методы импутирования. К простым относятся замещения пропусков специальными значениями, например нулями, средними или медианными значениями по признаку, модой или новой категорией для дискретных и категориальных

признаков [10]. Такие способы восстановления пропущенных значений наиболее распространены на практике благодаря своей простоте реализации, но подходят преимущественно для случайных пропусков (MAR). Также в эту группу методов входит использование вспомогательной регрессии для признаков, метода ближайших соседей (HotDeck [11], kNN [12]) и метода k-средних (k-means [13]) для поиска подходящего значения на место пропуска.

К сложным методам импутирования относятся итеративные алгоритмы, предполагающие оптимизацию некоторого функционала, отражающего точность расчета подставляемых значений. Причем оптимизация функционала может проводиться как глобально, т.е. по всем объектам выборки (метод Бартлетта [4], EM-оценивание [14, 15] и Resampling [4]), так и локально, когда для подбора пропущенного значения используются только близкие объекты (алгоритмы Zet [16, 17] и ZetBraid [18]).

Так или иначе, любые алгоритмы импутирования предполагают использование имеющейся информации для заполнения пропусков, что приводит к искажению статистических характеристик выборки в сторону характеристик только полных наблюдений. Кроме того, замещение пропусков вносит в выборку определенную долю искусственных данных, которые в свою очередь приводят к смещению значимости получаемых на их основе результатов [19].

Для борьбы с данными недостатками в публикациях [20–22] был предложен метод множественного импутирования, предполагающий замещение пропусков несколькими возможными значениями, с каждым из которых будет формироваться отдельный массив данных для обучения и анализа. Множественное импутирование призвано отразить присущую реальным данным неопределенность при подстановке пропущенных значений. Поэтому результаты анализа получившихся массивов данных приводятся в терминах средних значений, дисперсионного разброса и доверительных интервалов. На аналогичной идее усреднения основаны также методы стохастической регрессии для восстановления пропусков и популярная в последнее время технология размножения выборки bootstrap [23].

Таким образом, существующие методы работы с пропущенными данными различаются по качеству восстановления, сложности реализации и критериям применимости. Наиболее интуитивными и простыми в реализации являются отбрасывание пропусков и импутирование специальными значениями, более технологичными и универсальными являются поиск ближайших соседей и восстановление регрессией, а наиболее робастными оказываются стохастические методы, основанные на многократном повторении и усреднении результатов анализа.

В данной статье предложен новый метод восстановления пропущенных данных, основанный на энтропийно-робастном оценивании и генерации ансамблей случайных величин [24, 25]. В результате оценивания восстанавливаются плотности распределения вероятностей как для параметров модели, так и для шумов измерений, что позволяет восстановить пропущенные значения с помощью семплирования случайных величин.

2. Структура процедуры рандомизированного восстановления пропущенных данных

Рассмотрим состав и свойства информационного обеспечения задачи машинного обучения. Обучающая коллекция состоит из данных о функционировании исследуемого объекта O (рис. 1) и его измеренных входа $\mathbf{x}(t) \in \mathrm{R}^m$ и выхода $\mathbf{y}(t) \in \mathrm{R}^n$ объекта O. Переменные $\mathbf{x}(t), \mathbf{y}(t)$ сопровождаются измерительными ошибками $\eta(t) \in \mathrm{R}^m$ и $\xi(t) \in \mathrm{R}^n$ соответственно. Обычно информация о них гипотетическая. Обычно предполагается, что они в обучающей коллекции присутствуют аддитивно:

$$\mathbf{x}(t) = \mathbf{x}^0(t) + \eta(t); \qquad \mathbf{y}(t) = \mathbf{y}^0(t) + \xi(t), \tag{2.1}$$

где $\mathbf{x}^0(t)$, $\mathbf{y}^0(t)$ — "чистые" переменные.

Будем полагать, что измерения входа и выхода осуществляются в одной временной шкале $t = kh$ с шагом h и на одинаковом временном интервале $\mathcal{T} = [0, T]$. Следовательно, реальная обучающая коллекция состоит из матрицы данных о входе и выходе объекта O.

Наборы данных о входе и выходе могут иметь пропуски (рис. 2, интервал $[k^-, k^+]$). Возможны три ситуации пропусков: в выходных, входных, и в выходных и входных данных (необязательно в одни и те же моменты времени).

Основная идея рандомизированного восстановления пропущенных данных состоит в том, чтобы пропуски заменить ансамблем пропущенных данных (АПД) — случайными значениями компонент входа и выхода (см. рис. 2), оптимизированным с учетом имеющихся измерений. При этом используется параметризованная рандомизированная модель (РПМ) объекта, вероятностные характеристики которой — плотности распределение вероятностей ее параметров (ПРВ) — определяются посредством MEE-оценивания (Maximum Entropy Estimation) [26].

Генерация оптимизированных АПД осуществляется путем сэмплирования оптимальных ПРВ [27]. Синтезированный АПД позволяет использовать его непосредственно для решения различных задач машинного обучения, либо определять его численные характеристики: средние, медианные, дисперсион-

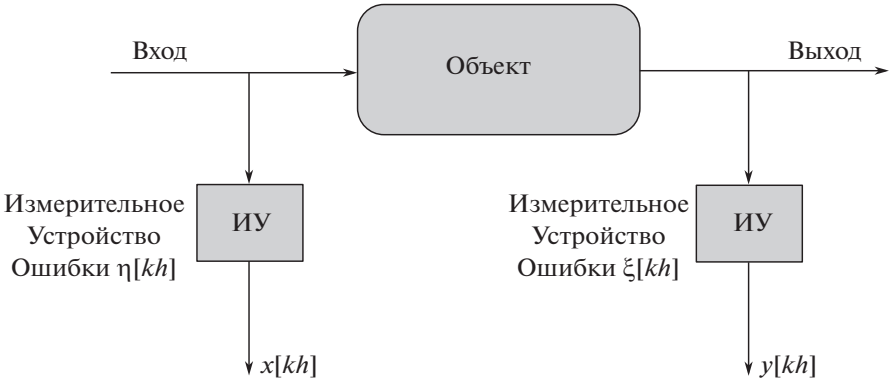

Рис. 1. Блок-схема процедуры измерения данных об объекте.

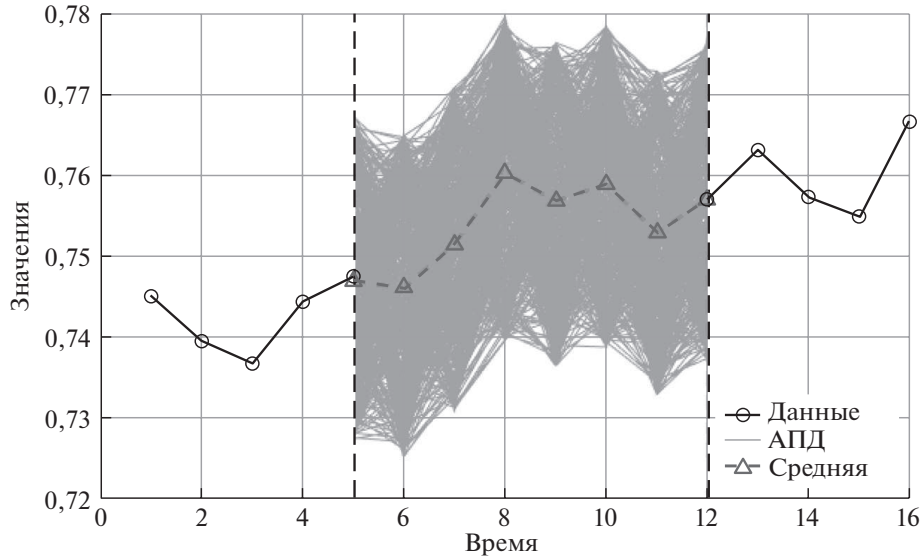

Рис. 2. Пример построения ансамбля пропущенных данных (АПД).

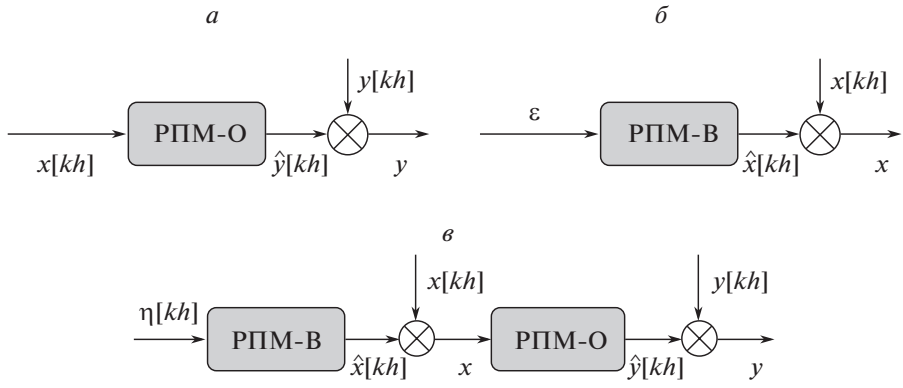

Рис. 3. Блок-схемы моделей РПМ.

ные трубки, интерквантильные множества данных и т.д. Иллюстрация этой идеи представлена на рис. 2.

В ситуации, когда пропуски имеют место в выходных данных, процедура использует входные данные $\mathbf{x}[kh]$, РПМ-О и выходные данные с пропусками $\tilde{\mathbf{y}}[kh]$ для вычисления энтропийно-оптимальных ПРВ (рис. 3,*а*) и генерации ансамбля $\mathcal{Y}$ восстановленных входных данных.

Если имеют место пропуски во входных данных, то для их восстановления используется дополнительная РПМ-В со входом в виде оптимизированной последовательности независимых случайных векторов $\eta[kh]$ и входные данные с пропусками $\tilde{\mathbf{x}}[kh]$. Для определения ПРВ параметров и указанной последовательности используется MEE-оценивание. Они позволяют генери-

ровать ансамбль $\mathcal{E}$ случайных векторов $\eta[kh]$ и ансамбль $\mathcal{X}$ восстановленных входных данных (рис. 3,*б*).

И наконец, если пропуски имеют место и во входных, и в выходных данных, то сначала восстанавливаются ансамбли $\mathcal{E}$ случайных векторов $\eta[kh]$ и входных данных $\mathcal{X}$, которые затем используются для восстановления ансамблей выходных данных $\mathcal{Y}$ (рис. 3,*в*).

Совокупность методов рандомизированного восстановления пропущенных данных будем обозначать RRMD — Random Restored Missing Data.

3. Алгоритмы оптимизации АПД

1. Пропуски в выходных данных.

Пусть на дискретном интервале наблюдения $\tau = [0, N]$ имеется набор входных данных $\mathbf{x} \in \mathrm{R}^m$ без пропусков $\mathbf{x}[0], \ldots, \mathbf{x}[N]$ и набор выходных данных $\mathbf{y} \in \mathrm{R}^n$ с пропусками $\mathbf{y}[0], \ldots, \mathbf{y}[k^- - 1], \mathbf{y}[k^+ + 1], \ldots, \mathbf{y}[N]$. Данные $\mathbf{y}[k^-], \ldots, \mathbf{y}[k^+]$ отсутствуют.

Из набора входных данных образуем матрицу

$$X_{(k-\varrho)} = [\mathbf{x}[k - \varrho], \mathbf{x}[k - \varrho + 1], \ldots, \mathbf{x}[k]] \tag{3.1}$$

с параметром $\rho < N$.

Выходные данные содержат аддитивные ошибки $\xi[k] \in \mathrm{R}^n$ интервального типа:

$$\xi[k] \in \Xi_k = [\xi^-, \xi^+]. \tag{3.2}$$

Вероятностные свойства ошибок характеризуются функциями ПРВ $Q_k(\xi[k])$, $k = \overline{0, N}$, которые предполагаются непрерывно-дифференцируемыми.

В общем случае связь между входом и выходом объекта O описывается рандомизированной параметризованной динамической моделью РДМ-О:

$$\hat{\mathbf{y}}[k] = \mathbb{B}\left(\mathbf{a}, X_{(k-\varrho)}\right), \tag{3.3}$$

где $\mathbb{B}$ — нелинейный функционал со случайными параметрами $\mathbf{a} \in \mathrm{R}^r$ интервального типа:

$$\mathbf{a} \in \mathcal{A} = [\mathbf{a}^-, \mathbf{a}^+]. \tag{3.4}$$

Вероятностные свойства параметров характеризуются функцией ПРВ $P(\mathbf{a})$, которая предполагается непрерывно-дифференцируемой.

Наблюдаемый выход РДМ-О

$$\mathbf{v}[k] = \hat{\mathbf{y}}[k] + \xi[k], \quad k = \overline{0, N}. \tag{3.5}$$

Для определения оптимальных ПРВ воспользуемся алгоритмом рандомизированного машинного обучения (РМО) [25], который в данном случае имеет

вид:

$$\mathcal{H}[P(\mathbf{a}), Q(\xi)] = -\int\limits_{\mathcal{A}} P(\mathbf{a}) \ln P(\mathbf{a}) d\mathbf{a} -$$

$$-\sum_{k=0}^{T} \int\limits_{\Xi_k} Q_k(\xi[k]) \ln Q_k(\xi[k]) d\xi[k] \Rightarrow \max, \tag{3.6}$$

$$\int\limits_{\mathcal{A}} P(\mathbf{a}) d\mathbf{a} = 1; \quad \int\limits_{\Xi_k} Q_k(\xi[k]) d\xi[k] = 1, \ k = \overline{0, N},$$

$$\int\limits_{A} P(\mathbf{a}) \mathbb{B}\left(\mathbf{a}, X_{(k-\varrho)}\right) d\mathbf{a} + \int\limits_{\Xi_k} Q_k(\xi[k]) \xi[k] d\xi[k] = \mathbf{y}[k],$$

$$k \neq k^-, \ldots, k^+; \ k = \overline{0, N}.$$

Учитывая, что ПРВ — непрерывно-дифференцируемые функции и задача (3.6) относится к ляпуновскому классу функциональных оптимизационных задач [28], можно получить ее аналитическое решение, параметризованное множителями Лагранжа $\theta^{(0)}, \ldots, \theta^{k^- - 1}, \theta^{k^+ + 1}, \ldots, \theta^{(N)}$:

$$P^*(\mathbf{a}, \theta) = \frac{\exp\left[-\sum\limits_{k=0}^{N} \langle \theta^{(k)}, \mathbb{B}\left(\mathbf{a}, X_{(k-\varrho)}\right)\rangle\right]}{\mathcal{P}(\theta)}, \tag{3.7}$$

$$Q_k^*(\xi[k], \theta^{(k)}) = \frac{\exp\left(-\langle \theta^{(k)}, \xi[k]\rangle\right)}{\mathcal{Q}_k(\theta^{(k)})}, \qquad k \neq k^-, \ldots, k^+; \quad k = \overline{0, N}.$$

Здесь

$$\mathcal{P}(\theta) = \int\limits_{\mathcal{A}} \exp\left[-\sum_{k=0}^{N} \langle \theta^{(k)}, \mathbb{B}\left(\mathbf{a}, X_{(k-\varrho)}\right)\rangle\right] d\mathbf{a}, \tag{3.8}$$

$$\mathcal{Q}_k(\theta^{(k)}) = \int\limits_{\Xi_k} \exp\left(-\langle \theta^{(k)}, \xi[k]\rangle\right) d\xi[k], \qquad k \neq k^-, \ldots, k^+; \quad k = \overline{0, N}.$$

Множители Лагранжа определяются из системы уравнений

$$\frac{\mathbf{M}_k(\theta)}{\mathcal{P}(\theta)} + \frac{\mathbf{N}_k(\theta^{(k)})}{\mathcal{Q}_k(\theta^{(k)})} = \mathbf{y}[k], \qquad k \neq k^-, \ldots, k^+; \ k = \overline{0, N}, \tag{3.9}$$

где

$$\mathbf{M}_k(\theta) = \int\limits_{\mathcal{A}} \exp\left[-\sum_{k=0}^{T} \langle \theta^{(k)}, \mathbb{B}\left(\mathbf{a}, X_{(k-\varrho)}\right)\rangle\right] \mathbb{B}\left(\mathbf{a}, X_{(k-\varrho)}\right) d\mathbf{a}, \tag{3.10}$$

$$\mathbf{N}_k(\theta^{(k)}) = \int\limits_{\Xi_k} \exp\left(-\langle \theta^{(k)}, \xi[k]\rangle\right) \xi[k] d\xi[k]. \tag{3.11}$$

2. Пропуски во входных данных.

Восстановление пропусков во входных данных представляет собой более сложную задачу, чем то же самое для входных данных. Причина состоит в том, что механизм происхождения входных данных и источник его возбуждения неизвестны. В терминах используемого здесь подхода, неизвестны модель, которая может генерировать указанные входные данные, и возбуждающий ее процесс. Все это повышает уровень неопределенности в данной задаче, что мотивирует применение рандомизированного подхода.

Реализация его состоит в использовании последовательности независимых случайных векторов $\zeta[k] \in \mathrm{R}^q$ $(k = \overline{0, N})$ со специальными вероятностными характеристиками и рандомизированной параметризованной модели, выбор которой определяется ее степенью сложности[2].

Итак, рассмотрим интервал наблюдения $\tau = [0, N]$ и набор входных данных $\mathbf{x} \in \mathrm{R}^m$ с пропусками $\mathbf{x}[0], \ldots, \mathbf{x}[k^- - 1], \mathbf{x}[k^+ + 1], \ldots, \mathbf{x}[T]$. Данные $\mathbf{x}[k^-], \ldots, \mathbf{x}[k^+]$ отсутствуют.

Для восстановления пропущенных входных данных будем полагать, что они являются результатом преобразования последовательности независимых случайных векторов $\zeta[k] \in \mathrm{R}^q$, $k \in \tau$. Будем полагать, что векторы $\zeta[k]$ — интервального типа, т.е.

$$\zeta[k] \in \mathcal{Z}_k = [\zeta^-[k], \zeta^+[k]], \qquad k = \overline{0, N}. \tag{3.12}$$

В силу неопределенности и недостаточности знаний о процессе и его преобразовании будем использовать его рандомизированную параметризованную модель (РПМ-В)

$$\hat{\mathbf{x}}[k] = \mathbb{F}\left(\mathbf{c}, Z_{(k-\rho)}\right), \tag{3.13}$$

где $\mathbb{F}$ — нелинейный векторный функционал со случайными параметрами $\mathbf{c} \in \mathrm{R}^s$ и входом — матрицей, составленной из столбцов векторов $\zeta[k] \in \mathrm{R}^q$, с размером $(q \times \rho)$:

$$Z_{(k-\rho)} = [\zeta[k-\rho], \zeta[k-\rho+1], \ldots, \zeta[k]]\,. \tag{3.14}$$

Введем матрицу

$$Z = \left[Z_{(-\rho)}, Z_{(1-\rho)}, \ldots, Z_{(N-\rho)}\right], \tag{3.15}$$

характеризующую вход РПМ-В (3.13) на интервале $k = \overline{0, N}$. Матрица Z – случайная, интервального типа:

$$Z \in \mathcal{Z} = \bigcup_{k=0}^{N} \mathcal{Z}_k = [Z^-, Z^+]. \tag{3.16}$$

Случайные параметры модели

$$\mathbf{c} \in \mathcal{C} = [\mathbf{c}^-, \mathbf{c}^+]. \tag{3.17}$$

[2] Возможно, потребуется перебор и тестирование разных моделей.

Вероятностные свойства параметров $\mathbf{c}$ и матрицы Z (3.15) характеризуются функцией совместной плотности распределения вероятностей (ПРВ) $W(\mathbf{c}, Z)$, которая определена на множестве

$$\mathcal{F} = \mathcal{C} \bigcup \mathcal{Z} \tag{3.18}$$

и предполагается непрерывно-дифференцируемой.

Пусть имеющиеся входные данные $\mathbf{x}[k]$ содержат ошибки $\eta[k] \in \mathrm{R}^m$ интервального типа:

$$\eta[k] \in \mathcal{E}_k = [\eta^-, \eta^+]. \tag{3.19}$$

Вероятностные свойства ошибок характеризуются функциями ПРВ $G_k(\eta[k])$, $k = \overline{0, N}$, которые предполагаются непрерывно-дифференцируемыми.

Наблюдаемый выход РПМ-В

$$\mathbf{z}[k] = \hat{\mathbf{x}}[k] + \eta[k], \quad k = \overline{0, N}. \tag{3.20}$$

Отсюда видно, что для восстановления пропущенных входных данных используется похожая с РДМ-О модель, но со случайным входом $\zeta[k]$.

Воспользуемся алгоритмом РМО [25] для определения функций ПРВ $W(\mathbf{c}, Z)$ и $G_k(\zeta[k])$, $k = \overline{0, N}$, который в данном случае имеет вид:

$$\begin{gathered}
\mathcal{H}[W(\mathbf{c}, Z), G(\eta)] = -\int\limits_{\mathcal{F}} W(\mathbf{c}, Z) \ln W(\mathbf{c}, Z)\, d\mathbf{c}\, dZ - \\
- \sum_{k \in \mathcal{K}} \int\limits_{\mathcal{E}_k} G_k(\eta[k]) \ln G_k(\eta[k]) d\eta[k] \Rightarrow \max, \\
\int\limits_{\mathcal{F}} W(\mathbf{c}, Z)\, d\mathbf{c}\, dZ = 1; \quad \int\limits_{\mathcal{E}_k} G_k(\eta[k]) d\eta[k] = 1, \\
\int\limits_{\mathcal{F}} W(\mathbf{c}, Z) \mathbb{F}\left(\mathbf{c}, Z_{(k-\rho)}\right) d\mathbf{c}\, dZ + \int\limits_{\mathcal{E}_k} G_k(\eta[k]) \eta[k] d\eta[k] = \mathbf{x}[k], \\
k \in \mathcal{K},
\end{gathered} \tag{3.21}$$

где множество индексов $\mathcal{K} = \overline{0, N}$ не содержит значения $k = k^-, \ldots, k^+$.

Эта задача того же класса, что и (3.6). Ее решение имеет вид

$$\begin{gathered}
W^*(\mathbf{c}, Z, \lambda) = \frac{\exp\left[-\sum\limits_{k \in \mathcal{K}} \langle \lambda^{(k)}, \mathbb{F}\left(\mathbf{c}, Z_{(k-\rho)}\right)\rangle\right]}{\mathcal{W}(\lambda)}, \\
G_k^*(\eta[k], \lambda^{(k)}) = \frac{\exp\left(-\langle \lambda^{(k)}, \eta[k]\rangle\right)}{\mathcal{G}_k(\lambda^{(k)})}, \\
k \in \mathcal{K},
\end{gathered} \tag{3.22}$$

где

$$\begin{gathered}\mathcal{W}(\lambda)=\int\limits_{\mathcal{F}}\exp\left[-\sum_{k\in\mathcal{K}}\langle\lambda^{(k)},\mathbb{F}\left(\mathbf{c},Z_{(k-\rho)}\right)\rangle\right]d\mathbf{c},\\ \mathcal{G}_k(\lambda^{(k)})=\int\limits_{\mathcal{E}_k}\exp\left(-\langle\lambda^{(k)},\eta[k]\rangle\right)d\eta[k].\end{gathered}\tag{3.23}$$

В этих равенствах $\lambda^{(k)}$ — вектор множителей Лагранжа, соответствующий k-му балансовому ограничению. Он определяется из системы уравнений:

$$\frac{\tilde{\mathbf{M}}_k(\lambda)}{\mathcal{W}(\lambda)}+\frac{\tilde{\mathbf{S}}_k(\lambda^{(k)})}{\mathcal{G}_k(\lambda^{(k)})}=\mathbf{x}[k],\quad k\neq k^-,\ldots,k^+;k\in\mathcal{K},\tag{3.24}$$

где

$$\begin{gathered}\tilde{\mathbf{M}}_k(\lambda)=\int\limits_{\mathcal{F}}\exp\left[-\sum_{k\in\mathcal{K}}\langle\lambda^{(k)},\mathbb{F}\left(\mathbf{c},Z_{(k-\rho)}\right)\rangle\right]\mathbb{F}\left(\mathbf{c},Z_{(k-\rho)}\right)d\mathbf{c}dZ,\\ \tilde{\mathbf{S}}_k(\lambda^{(k)})=\int\limits_{\mathcal{E}_k}\exp\left(-\langle\lambda^{(k)},\eta[k]\rangle\right)\eta[k]d\eta[k].\end{gathered}\tag{3.25}$$

3. Пропуски во входных и выходных данных.

В этом случае описанные процедуры применяются последовательно: сначала алгоритм (3.21), а затем алгоритм (3.6).

В результате применения упомянутых алгоритмов отдельно или последовательно получаем

- энтропийно-оптимальные функции ПРВ параметров соответствующих моделей: $P^*(\mathbf{a}), W^*(\mathbf{c})$;
- энтропийно-оптимальные функции ПРВ измерительных шумов: $Q_k^*(\xi[k]), G_k^*(\eta[k])$, $k=\overline{1,N}$;
- энтропийно-оптимальную функцию ПРВ рандомизированной входной последовательности $K_k^*(\zeta[k])$, $k=\overline{1,N}$.

4. Сэмплирование оптимальных функций ПРВ

Сэмплирование предполагает трансформацию функции ПРВ в соответствующую последовательность случайных векторов. В данной процедуре это есть способ реализации ансамблей случайных векторов на участках пропущенных данных.

Общий метод генерации последовательностей случайных векторов с заданной функцией ПРВ изложен в [29]. Он используется для сэмплирования последовательностей соответствующих случайных векторов-параметров РПМ-О, РПМ-В и векторов-шумов измерений и генерирования ансамблей восстановленных данных с помощью метода Монте-Карло [30].

Полезными характеристиками ансамблей являются:

— эмпирические функции ПРВ $\mathcal{P}_k(\hat{\mathbf{v}}_{rst}[k]), \mathcal{W}_k(\hat{\mathbf{z}}_{rst}[k])$ для восстановленных выходных и входных данных соответственно;

— эмпирические функции РВ:

$$\begin{gathered}\mathfrak{P}_k(\hat{\mathbf{v}}^{(i)}_{rst}[k]) = \sum_{j=1}^{i} \mathcal{P}_k(\hat{\mathbf{v}}^{(j)}_{rst}[k]) \quad \text{для выходных данных;} \\ \mathfrak{W}_k(\hat{\mathbf{z}}^{(i)}_{rst}[k]) = \sum_{j=1}^{i} \mathcal{W}_k(\hat{\mathbf{v}}^{(j)}_{rst}[k]) \quad \text{для входных данных.}\end{gathered} \tag{4.1}$$

Востребованными для решения задач машинного обучения являются наборы восстановленных данных, которые являются интегральными характеристиками ансамбля случайных траекторий.

1. Данные, соответствующие максимумам ПРВ параметров и шумов (max-pn):

— восстановленные выходные данные

$$\begin{gathered}\hat{\mathbf{y}}_{rst}[k] = \mathbb{B}(\mathbf{a}^*, X_{(k-\rho)}), \quad \mathbf{a}^* = \arg\max P^*(\mathbf{a}), \\ \xi_{rst}[k] = \arg\max Q^*_k(\xi[k]), \\ \hat{\mathbf{v}}_{rst}[k] = \hat{\mathbf{y}}_{rst}[k] + \xi_{rst}[k], \qquad k = \overline{0, N};\end{gathered} \tag{4.2}$$

— восстановленные входные данные

$$\begin{gathered}\hat{\mathbf{x}}_{rst}[k] = \mathbb{F}(\mathbf{c}^*, Z^*_{(k-\rho)}), \quad \mathbf{c}^* = \arg\max W^*(\mathbf{a}), \\ Z^*_{(k-\rho)} = [\zeta^*[k-\rho], \ldots, \zeta^*[k]], \quad \zeta^*[k] = \arg\max K^*_k(\zeta[k]), \\ \eta_{rst}[k] = \arg\max G^*_k(\eta[k]), \\ \hat{\mathbf{z}}_{rst}[k] = \hat{\mathbf{x}}_{rst}[k] + \eta_{rst}[k], \qquad k = \overline{0, N}.\end{gathered} \tag{4.3}$$

2. Данные, соответствующие максимумам эмпирических ПРВ наблюдаемых траекторий для $k = \overline{0, N}$ (max-ePDF):

— восстановленные выходные данные

$$\check{\mathbf{v}}_{rst}[k] = \arg\max \mathcal{P}_k(\hat{\mathbf{v}}_{rst}[k]), \quad k = \overline{0, N}; \tag{4.4}$$

— восстановленные входные данные

$$\check{\mathbf{z}}_{rst}[k] = \arg\max \mathcal{W}_k(\hat{\mathbf{z}}_{rst}[k]), \quad k = \overline{0, N}. \tag{4.5}$$

3. Данные, соответствующие средней по ансамблю траектории (mean) (M — количество траекторий в ансамбле):

— восстановленные выходные данные

$$\bar{\mathbf{v}}_{rst}[k] = \frac{1}{M} \sum_{i=1}^{M} \hat{\mathbf{v}}^{(i)}_{rst}[k], \quad k = \overline{0, N}; \tag{4.6}$$

— восстановление входных данных

$$\bar{\mathbf{z}}_{rst}[k] = \frac{1}{M}\sum_{i=1}^{M}\hat{\mathbf{z}}_{rst}^{(i)}[k], \quad k = \overline{0, N}. \tag{4.7}$$

4. Данные, соответствующие медианной по ансамблю траектории (med) (M — количество траекторий в ансамбле):

— восстановление выходных данных:

$$\hat{\mathbf{v}}_{rst}^{(j^*)}[k] \Rightarrow \sum_{j=1}^{j^*}\mathfrak{P}_k(\hat{\mathbf{v}}_{rst}^{(j)}[k]) = \sum_{j=j^*+1}^{M}\mathfrak{P}_k(\hat{\mathbf{v}}_{rst}^{(j)}[k]); \tag{4.8}$$

— восстановление входных данных:

$$\hat{\mathbf{z}}_{rst}^{(j^*)}[k] \Rightarrow \sum_{j=1}^{j^*}\mathfrak{W}_k(\hat{\mathbf{z}}_{rst}^{(j)}[k]) = \sum_{j=j^*+1}^{M}\mathfrak{W}_k(\hat{\mathbf{z}}_{rst}^{(j)}[k]). \tag{4.9}$$

5. Восстановление пропусков данных дистанционного зондирования состояния термокарстовых озер арктической зоны

Термокарстовые озера, расположенные в арктической зоне, являются основным природным источником эмиссии метана, который считается вторым по значимости парниковым газом, влияющим на изменение климата Земли [31]. Арктические зоны России и Канады являются основными регионами наиболее динамичных изменений вечной мерзлоты и термокарстовых озер. Поэтому изменения площади озер являются важными для прогнозирования динамики выбросов метана. Для получения данных о площадях озер используются космические снимки Landsat и средства их дешифрирования в геоинформационной системе ArcGIS. Климатические данные (среднегодовая температура и годовая сумма осадков) получены с использованием систем реанализа метеоданных ERA-40, ERA-Interim [32]. Пространственная структура данных состоит из 30 тестовых участков, расположенных в трех зонах вечной мерзлоты Западной Сибири: сплошной (C — *continuous*), прерывистой (D – *discontinuos*) и островной (I — *insular*).

Однако данные о текущем состоянии озер, в частности об их площади и форме, а также климатические данные (например, температура и осадки) регистрируются нерегулярно, а их преобразование от источника измерений до компьютерных атрибутов сопровождается существенными ошибками.

Многочисленными экспериментальными исследованиями с применением методов математической статистики показано, что основными факторами, влияющими на величину площади термокарстовых озер, являются среднегодовая температура воздуха $T[t]$ и годовая сумма осадков $R[t]$ (t — календарный год) [33].

1. Данные.

Данные по температуре $\tilde{T}$, осадкам $\tilde{R}$ и площади озер $\tilde{S}$ в календарной и в дискретной временных шкалах представлены в [34]. В настоящей статье

Таблица 1. Данные площади тестовых участков

Участок $I-5$	
Год	Площадь, га
1973	67,90
1984	68,89
1988	64,27
2001	62,36
2003	62,31
2007	66,96

Участок $D-9$	
Год	Площадь, га
1973	13,88
1987	14,29
1988	15,36
2001	15,40
2002	14,15
2003	14,50
2006	13,88
2007	13,01

Участок $C-24$	
Год	Площадь, га
1977	25,62
1981	22,59
1988	24,24
1999	23,81
2001	22,96
2006	23,76

используются данные для трех участков $I-5$, $D-9$, $C-24$, данные по площади представлены в табл. 1, а данные по температуре и количеству осадков приведены в Приложении в табл. П.1.

Данные охватывают временной интервал [1973–2007], или $k=\overline{1,35}$. Преобразуем их к нормализованной форме, т.е. отобразим их значения на числовой отрезок $[0,1]$:

$$S_r=\frac{\tilde{S}-\tilde{S}_{\min}}{\tilde{S}_{\max}-\tilde{S}_{\min}},\quad T_r=\frac{\tilde{T}-\tilde{T}_{\min}}{\tilde{T}_{\max}-\tilde{T}_{\min}},\quad R_r=\frac{\tilde{R}-\tilde{R}_{\min}}{\tilde{R}_{\max}-\tilde{R}_{\min}}. \tag{5.1}$$

Данные о площади озер на соответствующих участках с пропусками:

- *участок* $I-5$: обучающая коллекция

$$S_{r,LM}^{(I-5)}[k],\qquad k\in\mathcal{K}_{LM}^{(I-5)}=\{1,12,16,29,31,35\};$$

- *участок* $D-9$: обучающая коллекция

$$S_{r,LM}^{(D-9)}[k],\qquad k\in\mathcal{K}_{LM}^{(D-9)}=\{1,16,29,35\};$$

тестовая коллекция

$$S_{r,T}^{(D-9)}[k],\qquad k\in\mathcal{K}_{T}^{(D-9)}=\{15,30,31,34\};$$

- *участок* $C-24$: обучающая коллекция

$$S_{r,LM}^{(C-24)}[k],\qquad k\in\mathcal{K}_{LM}^{(C-24)}=\{5,9,16,27,29,34\}.$$

Данные по площади озер на участках $I-5$, $C-24$ будут использованы для обучения соответствующих моделей и восстановления пропущенных данных, данные по участку $D-9$ — для обучения модели ($\mathcal{K}_{LM}^{(D-9)}$) и тестирования ($\mathcal{K}_{T}^{(D-9)}$) восстановленных данных.

2. Модель.

Восстановление пропущенных данных на трех обозначенных выше зонах и участках осуществляется единой по структуре моделью линейной регрессии вида

$$\hat{S}[k] = \alpha T[k] + \beta R[k] + \xi[k], \quad k = \overline{0, N}, \tag{5.2}$$

где параметры α, β — случайные, независимые, интервального типа:

$$\alpha \in \mathcal{A} = [\alpha^-, \alpha^+], \quad \beta \in \mathcal{B} = [\beta^-, \beta^+]. \tag{5.3}$$

Для принятых тестовых участков границы интервалов одинаковые[3]:

$$\mathcal{A}[0;1], \quad \mathcal{B} = [0;1].$$

Вероятностные свойства параметров характеризуются функциями ПРВ $P(\alpha), F(\beta)$.

Шум также нормализованный и интервальный

$$\xi[k] \in \Xi = [\xi^-, \xi^+] = [-0{,}15; 0{,}15] \tag{5.4}$$

с функциями ПРВ $Q_k(\xi[k]),\ k \in \mathcal{K}$.

3. *Алгоритм* РМО.

Общая форма алгоритма обучения модели имеет вид

$$\mathcal{H} = -\int\limits_{\mathcal{A}} P(\alpha)\ln P(\alpha)d\alpha - \int\limits_{\mathcal{B}} F(\beta)\ln F(\beta)d\beta - \\ -\sum_{k\in\mathcal{K}}\int\limits_{\Xi_k} Q_k(\xi[k])\ln Q_k(\xi[k])d\xi[k] \Rightarrow \max, \tag{5.5}$$

— нормировка

$$\int\limits_{\mathcal{A}} P(\alpha)d\alpha = 1, \quad \int\limits_{\mathcal{B}} F(\beta)d\beta = 1, \quad \int\limits_{\Xi_k} Q_k(\xi[k])d\xi[k] = 1, \quad k \in \mathcal{K}, \tag{5.6}$$

— эмпирические балансы

$$\int\limits_{\mathcal{A}} P(\alpha)\alpha T[k]d\alpha + \int\limits_{\mathcal{B}} F(\beta)\beta R[k]d\beta + \int\limits_{\Xi_k} Q_k(\xi[k])\xi[k]d\xi[k] = S_r[k],\ k \in \mathcal{K}. \tag{5.7}$$

Решение задачи (5.5)–(5.7) имеет вид

$$P^*(\alpha,\theta) = \frac{\exp\left(-\alpha l_r(\theta)\right)}{\mathcal{P}(\theta)}, \quad F^*(\beta,\theta) = \frac{\exp\left(-\beta h_r(\theta)\right)}{\mathcal{F}(\theta)},$$

$$Q_k(\xi[k],\theta_k) = \frac{\exp(-\theta_k\xi[k])}{\mathcal{Q}_k(\theta_k)}, \quad k \in \mathcal{K}, \tag{5.8}$$

где $\theta = \{\theta_k,\ k \in \mathcal{K}\}$ — множители Лагранжа,

[3] Границы интервалов косвенно влияют на качество восстановленных данных, измеряемого принятым функционалом. Для выбора границ можно воспользоваться его численной оптимизацией.

Таблица 2. Расчетные данные по тестовым участкам

Участок $I-5$	
$\mathcal{K}_{LM}^{(I-5)}$	$\theta/10^3$
1	0,0015
12	2,99
16	–0,0278
29	–0,356
31	–4,73
35	3,26

Участок $D-9$	
$\mathcal{K}_{LM}^{(D-9)}$	$\theta/10^3$
1	–2,13
16	4,73
29	0,0321
35	–1,56

Участок $C-24$	
$\mathcal{K}_{LM}^{(C-24)}$	$\theta/10^3$
5	4,73
9	–0,734
16	–0,0102
27	–1,07
29	–0,076
34	–0,0042

— нормировочные коэффициенты

$$\mathcal{P}(\theta)=\int\limits_{\mathcal{A}}\exp\left(-\alpha l_r(\theta)\right)d\alpha;\qquad \mathcal{F}(\theta)=\int\limits_{\mathcal{B}}\exp\left(-\beta h_r(\theta)\right)d\beta;$$

$$\mathcal{Q}_k(\theta_k)=\int\limits_{\Xi}\exp(-\theta_k\xi[k])d\xi[k],\qquad k\in\mathcal{K}, \tag{5.9}$$

— показатели экспонент

$$l_r(\theta)=\sum_{k\in\mathcal{K}}\theta_k T[k],\quad h_r(\theta)=\sum_{k\in\mathcal{K}}\theta_k R[k]. \tag{5.10}$$

Множители Лагранжа определяются из системы уравнений

$$L(\theta)T[k]+K(\theta)R[k]+V_k(\theta_k)=S_r[k],\quad k\in\mathcal{K}, \tag{5.11}$$

где

$$L(\theta)=\frac{\int\limits_{\alpha^-}^{\alpha^+}\alpha\exp\left(-\alpha l_r(\theta)\right)d\alpha}{\mathcal{P}(\theta)},\qquad K(\theta)=\frac{\int\limits_{\beta^-}^{\beta^+}\beta\exp\left(-\beta h_r(\theta)\right)d\beta}{\mathcal{F}(\theta)},$$

$$V_k(\theta_k)=\frac{\int\limits_{\xi^-[k]}^{\xi^+[k]}\xi[k]\exp\left(-\theta_k\xi[k]\right)d\xi[k]}{\mathcal{Q}_k(\theta_k)}. \tag{5.12}$$

В табл. 2 приведены результаты обучения по тестовым участкам $I-5$, $D-9$, $C-24$.

В силу линейности модели (5.2) все энтропийно-оптимальные функции ПРВ — экспоненциальные:

• участок $I-5$:

$$P^*(\alpha)=7{,}28\exp(-7{,}27\alpha),\quad F^*(\beta)=0{,}0448\exp(4{,}65\beta),$$
$$\alpha^{mean}=0{,}137;\quad \beta^{mean}=0{,}795; \tag{5.13}$$

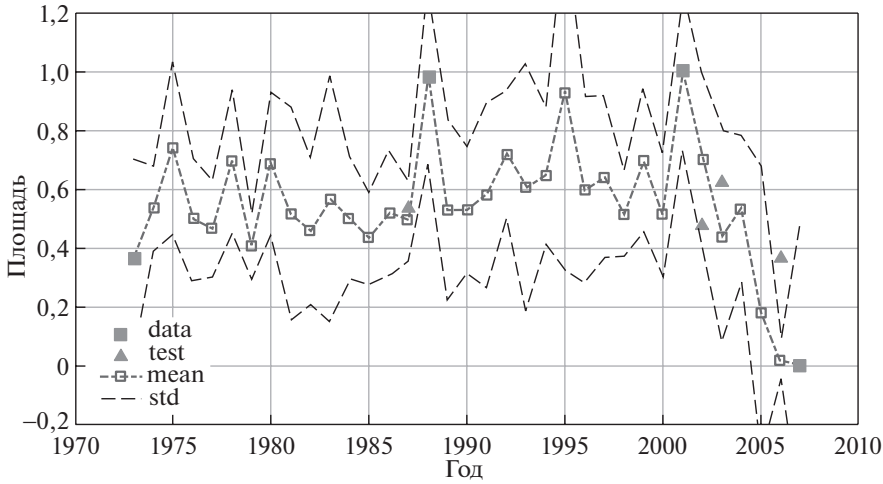

Рис. 4. Восстановление пропущенных данных на участке $D-9$.

• участок $D-9$:

$$P^*(\alpha) = 0{,}476\exp(0{,}543\alpha), \quad F^*(\beta) = 0{,}0462\exp(4{,}60\beta),$$
$$\alpha^{mean} = 0{,}178; \quad \beta^{mean} = 0{,}783; \tag{5.14}$$

• участок $C-24$:

$$P^*(\alpha) = 1{,}84\exp(-1{,}37\alpha), \quad F^*(\beta) = 1{,}14\exp(-0{,}272\beta),$$
$$\alpha^{mean} = 0{,}389; \quad \beta^{mean} = 0{,}477. \tag{5.15}$$

4. Тестирование для участка $D-9$.

Для участка $D-9$ имеются данные $\mathcal{K}_T^{(D-9)}$, которые будем использовать для оценки качества восстановления данных на этом участке.

Имея набор энтропийно-оптимальных ПРВ (5.14), сэмплируем их и, вычисляя для каждого сэмпла траектории $\hat{S}[k]$, формируем ансамбль случайных траекторий на интервале $k \in [1, 35]$ (рис. 4). На этом же интервале показана средняя по ансамблю траектория $mean[k]$, которая принята за траекторию восстановленных данных.

Качество восстановления характеризуется относительной ошибкой

$$\delta = \frac{\sqrt{\sum\limits_{k\in\mathcal{K}_T^{(D-9)}} \left(\hat{S}[k] - S_{r,T}^{(D-9)}[k]\right)^2}}{\sqrt{\sum\limits_{k\in\mathcal{K}_T^{(D-9)}} \hat{S}^2[k]} + \sqrt{\sum\limits_{k\in\mathcal{K}_T^{(D-9)}} \left(S_{r,T}^{(D-9)}[k]\right)^2}}. \tag{5.16}$$

Для участка $D-9$ относительная ошибка $\delta \approx 0{,}23$. Поскольку сформирован ансамбль траекторий восстановленных данных, то по нему можно определить и другие типы траекторий и доверительные области (см. п. 5).

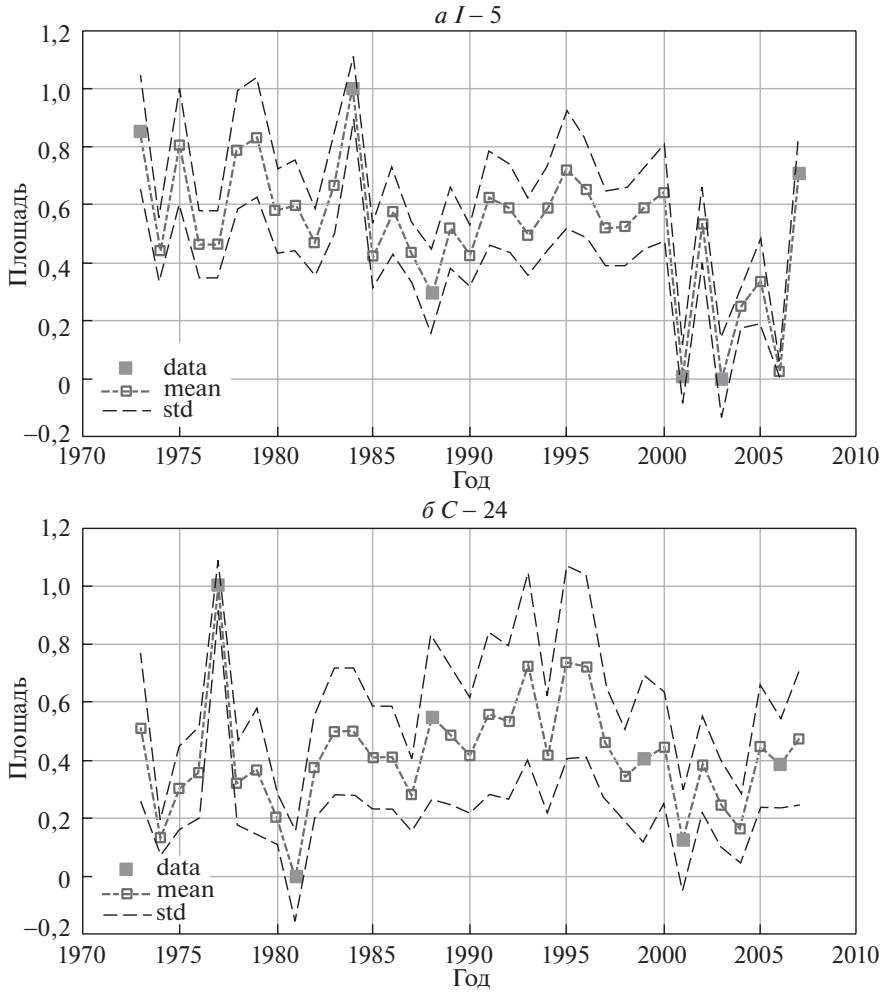

Рис. 5. Восстановление пропущенных данных на участках $I-5$ ($а$) и $C-24$ ($б$).

5. Генерация восстановленных данных.

Для модели (5.2) производилось сэмплирование функций ПРВ параметров согласно (5.13) и (5.15). Для каждого сэмпла вычислялась траектория $\hat{S}[k]$, $k \in \mathcal{K}$, и формировался ансамбль $\hat{\mathcal{S}}$ случайных траекторий.

Полезными вероятностными характеристиками ансамбля являются функции эмпирической плотности $U_k(\hat{S}[k])$ (гистограммы), определенные на множестве $\mathcal{S}_k = [\hat{S}^-[k], \hat{S}^+[k]]$, где $\hat{S}^-[k] = \min \hat{S}[k]$, $\hat{S}^+[k] = \max \hat{S}[k]$. Переменная $k \in \mathcal{K}^{(I-5)}$ для участка $I-5$ и $k \in \mathcal{K}^{(C-24)}$ для участка $C-24$.

Определялись следующие числовые характеристики ансамбля — траектории:

— максимумов гистограмм

$$\max[k] = \max U_k(\hat{S}[k]), \quad k \in \mathcal{T}; \tag{5.17}$$

— медиан

$$\int_{\hat{S}[k]^-}^{\mathrm{med}[k]} U_k(\hat{S}[k])d\hat{S}[k] = \int_{\mathrm{med}[k]}^{\hat{S}[k]^+} U_k(\hat{S}[k])d\hat{S}[k], \quad k \in \mathcal{T}; \tag{5.18}$$

— средних

$$\mathrm{mean}[k] = \int_{\hat{S}[k]^-}^{\hat{S}[k]^+} \hat{S}[k]\, U_k(\hat{S}[k])d\hat{S}[k]. \tag{5.19}$$

На рис. 5,*а* и 5,*б* показаны ансамбли и графики соответствующих траекторий.

6. Заключение

Предложен новый метод восстановления пропущенных данных, состоящий в генерации ансамбля случайных траекторий, формируемых рандомизированной параметризованной моделью, обученной с привлечением имеющихся данных. Для пропусков в выходных данных развит алгоритм рандомизированного машинного обучения, основанный на технике *MEE*-оценивания функций ПРВ параметров модели и шумов измерений. В случае пропусков во входных данных развит метод восстановления, использующий аналогичную рандомизированную параметризованную модель и специальный оптимизированный шум. С помощью техники *MEE*-оценивания определяются функции ПРВ параметров модели и шума.

Предложенный метод сочетает преимущества подходов параметрического оценивания с заданной моделью процесса и непараметрического оценивания при неизвестных характеристиках шумов измерений, что обеспечивает робастность метода, но может в определенных случаях приводить к завышенным ошибкам восстановления. Робастность также обеспечивается и процедурой генерации энтропийно-оптимизированных траекторий путем сэмплирования соответствующих функций ПРВ.

ПРИЛОЖЕНИЕ

Таблица П.1. Данные о температуре и осадках для тестовых участков

Участок $I-5$			Участок $D-9$		Участок $C-24$	
Год	T	R	T	R	T	R
1	2	3	4	5	6	7
1973	–1,00	650,30	–3,50	580,80	–11,00	372,40
1974	–3,50	494,25	–6,00	494,30	–12,00	211,00
1975	–2,00	684,80	–4,00	579,00	–9,50	220,50
1976	–2,50	487,30	–4,50	442,50	–10,00	263,40
1977	–2,50	486,50	–5,00	432,60	–11,00	216,70
1978	–3,50	706,40	–4,50	562,70	–11,00	275,40
1979	–2,00	700,10	–6,00	414,40	–12,50	342,90

Таблица П.1. (окончание)

1	2	3	4	5	6	7
1980	–2,50	556,50	–4,50	556,50	–11,00	216,00
1981	–1,00	534,60	–3,00	427,30	–9,00	209,30
1982	–2,50	489,95	–4,00	409,90	–11,00	303,10
1983	–1,00	578,80	–2,50	448,80	–9,50	318,80
1984	–2,50	465,05	–4,50	444,20	–9,50	319,20
1985	–3,50	482,20	–5,00	412,90	–9,50	274,20
1986	–3,50	577,30	–4,50	456,00	–9,50	274,00
1987	–2,50	467,50	–6,00	467,50	–11,00	253,60
1988	–1,00	501,20	–3,50	390,30	–10,00	390,30
1989	–1,00	487,70	–3,50	444,40	–11,00	357,70
1990	–2,50	462,10	–4,50	462,10	–11,00	323,00
1991	–1,00	551,50	–3,50	473,80	–11,00	396,20
1992	–3,50	585,30	–5,00	585,30	–11,00	381,80
1993	–1,00	471,50	–2,50	471,50	–8,50	402,70
1994	–2,50	562,50	–4,50	532,75	–11,00	324,60
1995	0,50	578,90	–1,00	640,50	–6,00	332,60
1996	–1,00	571,35	–3,50	485,90	–7,50	372,10
1997	–2,50	521,10	–4,00	521,10	–9,50	299,80
1998	–3,50	544,80	–6,00	480,70	–11,00	288,10
1999	–2,50	563,15	–4,50	563,15	–8,50	369,50
2000	–1,00	561,30	–4,50	453,30	–9,50	291,20
2001	–2,00	430,20	–4,00	511,60	–9,50	267,40
2002	–2,06	520,15	–4,02	555,80	–9,82	270,50
2003	–0,48	429,20	–2,92	378,90	–9,05	177,80
2004	–1,80	338,40	–4,13	457,80	–9,78	159,20
2005	0,44	343,90	–1,66	199,50	–8,02	247,70
2006	–2,78	222,00	–5,44	168,30	–9,24	222,00
2007	0,14	325,40	–1,77	285,75	–7,51	246,10

СПИСОК ЛИТЕРАТУРЫ

1. *Загоруйко Н.Г.* Методы распознавания и их применение. М.: Сов. Радио, 1972.
2. *Литтл Р.Дж.А., Рубин Д.Б.* Статистический анализ данных с пропусками. М.: Финансы и статистика, 1990.
3. *Загоруйко Н.Г.* Прикладные методы анализа данных и знаний. Новосибирск: ИМ СО РАН, 1999.
4. *Злоба Е., Яцкив И.* Статистические методы восстановления пропущенных данных // Computer Modeling & New Technologies. 2004. V. 6. P. 55–56.
5. *Molenberghs G., Kenward M.G.* Missing Data in Clinical Studies. Chichester, UK., John Wiley & Sons, 2007. P. 47–50.
6. *Cheema J.* A Review of Missing Data Handling Methods in Education Research // Review of Educational Research. 2014. No. 4. P. 487–508.
7. *Круглов В.В., Абраменкова И.В.* Методы восстановления пропусков в массивах данных // Методы восстановления пропусков в массивах данных. 2005. № 2.
8. *Van Buuren S.* Flexible Imputation of Missing Data. Chapman and Hall/CRC; 1 ed., 2012.
9. *Enders C.* Applied Missing Data Analysis. N.Y.–London, 2010.

10. *Schafer J.L., Schenker N.* Inference with Imputed Conditional Means // Journal of the American Statistical Association. 2000. V. 95. No. 449. P. 144–154.
11. *Mander A., Clayton D.* HotDeck Imputation. Stata Technical Bulletin. 1999. V. 51. P. 32–34.
12. *Batista G.E.A.P.A., Monard M.C.* K-Nearest Neighbour as Imputation Method: Experimental Results. Technical report, ICMC-USP, 2002.
13. *Dan Li, Jitender Deogun, William Spaulding, Bill Shuart.* Towards Missing Data Imputation: A Study of Fuzzy K-means Clustering Method // S. Tsumoto, et al. (Eds.): RSCTC 2004. LNAI 3066. 2004. P. 573–579.
14. *Dempster A.P., Laird N.M., Rubin D.B.* Maximum Likelihood from Incomplete Data via the EM Algorithm // Journal of the Royal Statistical Society. 1977. V. 39. P. 1–38.
15. *Zhou X.-Y., Lim J.S.* EM Algorithm with GMM and Naive Bayesian to Implement Missing Values // Advanced Science and Technology Lett. 2014. V. 46. P. 1–5.
16. *Загоруйко Н.Г., Елкина В.Н., Тимеркаев В.С.* Алгоритм заполнения пропусков в эмпирических таблицах (алгоритм Zet) // Эмпирическое предсказание и распознавание образов. Новосибирск: 1975. Вып. 61: Вычислительные системы. С. 3–27.
17. *Снитюк В.Е.* Эволюционный метод восстановления пропусков в данных. Сборник трудов VI Межд. конф. “Интеллектуальный анализ информации”, Киев: 2006. С. 262–271.
18. Алгоритм ZetBraid // Информационные интеллектуальные системы. 2008. Вып. 40.
19. *Rubin D.B.* Multiple Imputation for Nonresponse in Surveys. N.Y.: Willey, 1987. P. 64–66.
20. *Rubin D.B.* Multiple Imputation After 18+ Years // Journal of the American Statistical Association. 1996. No. 91. P. 473–489.
21. *Lipsitz S.R., Lue Ping Zhao, Molenberghs G.A.* Semiparametric Method of Multiple Imputation // Journal of the Royal Statistical Society. Ser. B (Statistical Methodology). 1998. V. 60. No. 1. P. 127–144.
22. *Horton N.J., Lipsitz S.R.* Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables // The American Statistician. 2001. V. 55. No. 3. P. 244–254.
23. *Efron B.* Missing Data, Imputation, and the Bootstrap // Journal of the American Statistical Association. 1994. V. 89. No. 426. P. 463–475.
24. *Popkov Y.S., Dubnov Y.A., Popkov A.Y.* New Method of Randomized Forecasting Using Entropy-Robust Estimation: Application to the World Population Prediction // Mathematics. 2016. V. 4. Iss. 1. P. 1–16.
25. *Попков Ю.С., Попков А.Ю., Дубнов Ю.А.* Рандомизированное машинное обучение. М.: УРСС, 2019.
26. *Попков Ю.С.* Асимптотическая эффективность оценок максимальной энтропии // ДАН. Математика, Информатика, Процессы управления. 2020. Т. 493. С. 100–103.
27. *Geweke J., Hisashi T.* Note on the Sampling Distribution for the Metropolis-Hastings Output // Joutnal of American Statistical Association. 2003. V. 96. (453). P. 270–281.
28. *Иоффе А.Д., Тихомиров В.М.* Теория экстремальных задач. М.: Наука, 1974.
29. *Дарховский Б.С., Попков Ю.С., Попков А.Ю., Алиев А.С.* Метод генерации случайных векторов с заданной функцией плотности распределения вероятностей // АиТ. 2018. Вып. 9. С. 31–45.

Darkhovsky B.S., Popkov Yu.S., Popkov A.Yu., Aliev A.S. A Method of Generating Random Vectors with a Given Probability Density Function // Autom. Remote Control. 2018. V. 79. P. 1569–1581.

30. *Rubinstein R.Y., Kroese D.P.* Simulation and Monte Carlo Method. John Wiley and Sons, 2007.
31. *Полищук В.Ю., Полищук Ю.М.* Геоимитационное моделирование полей термокарстовых озер в зонах мерзлоты. Ханты-Мансийск: УИП ЮГУ, 2013.
32. *Polishchuk Y.M., Muratov I.N., Polishchuk V.Y.* Remote Research of Spatiotemporal Dynamics of Thermocarst Lakes Fields in Siberian Permafrost / The Arctic: Current Issues and Challenges (Eds. Pokrovsky O.S., Kirpotin S.N., Malov A.I.). N.Y.: Nova Science Pbl., 2020. P. 208–237.
33. *Попков Ю.С., Волкович В., Мельников А.В., Полищук Ю.М.* Методологические вопросы использования рендомизированного машинного обучения для прогнозирования динамики термокарстовых озер Арктики // Вестн. Южно-Уральского гос. ун-та. Сер. “Компьютерные технологии, управление, радиоэлектроника”. 2019. Т. 19. Вып. 4. С. 5–12.
34. Электронный ресурс,
URL: https://cloud.uriit.ru/index.php/s/0DOrxL9RmGqXsV0.

Статья представлена к публикации членом редколлегии А.И. Михальским.

Поступила в редакцию 24.07.2020

После доработки 27.10.2020

Принята к публикации 08.12.2020