

К вопросу об устойчивости восстановления мелкомасштабных региональных полей экстремальных температур статистическими методами*

И. В. НОГОТКОВ, А. И. ЧАВРО, Е. В. ДМИТРИЕВ

Институт вычислительной математики РАН, Москва, Россия
e-mail: ilya@4ka.mipt.ru, chavro@inm.ras.ru, yegor@inm.ras.ru

This paper addresses the problems related to the stability of statistical methods for solving the downscaling problems for the regional fields of various meteorological parameters such as surface temperature, pressure, wind etc. The statistical model, which allows reconstructing more than 75 % of the natural variability of temperature extremes at meteorological stations in Moscow is suggested. In our opinion, there are some interesting difficulties, which are related to the inhomogeneity of data, which establishes a priori statistical information. It is demonstrated that even a little amount of missing data could cause an instability of the solution and reduce its accuracy. In this connection, some frequently used methods of spatial interpolation for reconstruction of the missing values of meteorological parameters are compared. We propose a combined method for filling the missed values, which allows substantial improvement of stability for solutions of the inverse problems.

Введение

Статистические методы широко используются для решения задачи восстановления детальной структуры региональных геофизических полей [1, 2]. Большой интерес представляет задача определения локальных изменений экстремальных значений метеопараметров [3, 4]. В большинстве существующих подходов требуется получить оценки неизвестных статистических данных. Для этого используют данные измерений на метеорологических станциях. Многим исследователям, так или иначе использующим в своей работе методы анализа данных, хорошо известно, что данные такого рода обычно содержат значительное количество пропущенных значений, появление которых может быть вызвано как несовершенством измерительных систем, так и “человеческим фактором”.

Казалось бы, проблема пропусков не так важна и вполне достаточно их построчного исключения из массива данных наблюдений для того, чтобы при дальнейшем анализе “забыть” об их существовании в исходном массиве. Однако устойчивость статистических моделей, используемых для решения данной задачи, определяется не только эффективностью методов оценки параметров и размером априорного ансамбля.

*Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (гранты № 07-05-00328-а и № 06-05-64916-а).

© Институт вычислительных технологий Сибирского отделения Российской академии наук, 2008.

В настоящей работе отмечается, что даже наличие совсем небольшого количества пропусков может быть причиной значительных выбросов в решении. Существующие процедуры заполнения пропусков данных не могут в полной мере решить эту проблему. Предлагаемый в настоящей работе комбинированный метод позволяет не только избежать ошибок, но и заметно повысить точность и устойчивость решения, получаемого с помощью статистической модели.

С развитием моделей краткосрочного прогноза одной из актуальных задач стало восстановление детальной структуры геофизических полей для больших городов. Методика, описанная в работе [5], может быть использована для построения статистической модели восстановления мелкомасштабных полей максимальных и минимальных суточных значений приземной температуры в Московском регионе и г. Москве. Прямая постановка задачи имеет следующий вид:

$$\boldsymbol{\xi} = A\mathbf{f} + \boldsymbol{\nu}, \tag{1}$$

где $\boldsymbol{\xi} \in \mathbf{R}^n$ — известный вектор значений крупномасштабного метеорологического параметра; $\mathbf{f} \in \mathbf{R}^m$ — искомый вектор, определяющий мелкомасштабное поле этого же метеопараметра; A — линейный оператор осреднения, а вектор $\boldsymbol{\nu} \in \mathbf{R}^n$ содержит погрешности такого представления.

Обратная задача, имеющая целью восстановление мелкомасштабного поля по крупномасштабному, является некорректной, поскольку не имеет единственного решения. При этом можно найти ее регуляризованное решение. Например, используя методы, предлагаемые в работах [6, 7], можно найти наилучшую среднеквадратичную оценку $\hat{\mathbf{f}}$ вектора $\mathbf{f} \in \mathbf{R}^k$. В случае, если оператор A неизвестен (см., например, [8]), такое решение имеет вид

$$\hat{\mathbf{f}} = R\boldsymbol{\xi} = \mathbf{C}_{\mathbf{f},\boldsymbol{\xi}}\mathbf{C}_{\boldsymbol{\xi},\boldsymbol{\xi}}^{-1}\boldsymbol{\xi}. \tag{2}$$

Для численного решения обратной задачи (2) необходима информация о вторых моментах исследуемых случайных полей, а именно требуется получить оценки для кросс-ковариационной матрицы случайных полей \mathbf{f} и $\boldsymbol{\xi}$, а также для автоковариационной матрицы крупномасштабного поля и ее последующее обращение. При практической реализации данного метода может возникать описанная выше проблема неустойчивости решения.

1. Оценка статистических данных при наличии пропусков

Очевидно, что если данные наблюдений содержат пропуски, то стандартные формулы для получения несмещенных оценок для вторых моментов неприменимы. Вместо этого часто используются некоторые их модификации [9].

Пусть X и Y — матрицы размерностей $t \times n$ и $t \times m$ соответственно. Столбцы \bar{x}_j и \bar{y}_k этих матриц — временные ряды измерений какого-либо метеопараметра в соответствующих точках пространства.

Обозначим через $U_{\bar{x}_j}$ и $U_{\bar{y}_k}$ множества индексов i , для которых соответствующие элементы $\{\bar{x}_j\}_i$ и $\{\bar{y}_k\}_i$ векторов \bar{x}_j и \bar{y}_k определены, т. е. их значения известны.

Введем множество $U_{jk} = U_{\bar{x}_j} \cap U_{\bar{y}_k}$. Тогда оценку элемента $\{C_{x,y}\}_{jk}$ ковариационной матрицы $C_{x,y}$ можно вычислить по формуле

$$\{C_{x,y}\}_{jk} = \frac{1}{T-1} \sum_{i \in U_{jk}} \{\bar{x}_j\}_i \{\bar{y}_k\}_i, \tag{3}$$

где T — количество элементов множества U_{jk} .

Рассмотрим простой пример случайного поля, описываемого следующей матрицей наблюдений:

$$X = (\bar{x}_1, \bar{x}_2) = \begin{pmatrix} 1 & 1 \\ -1 & -1 \\ 1 & 1 \\ -1 & -1 \\ 1 & 1 \\ -1 & -1 \\ * & 0 \end{pmatrix},$$

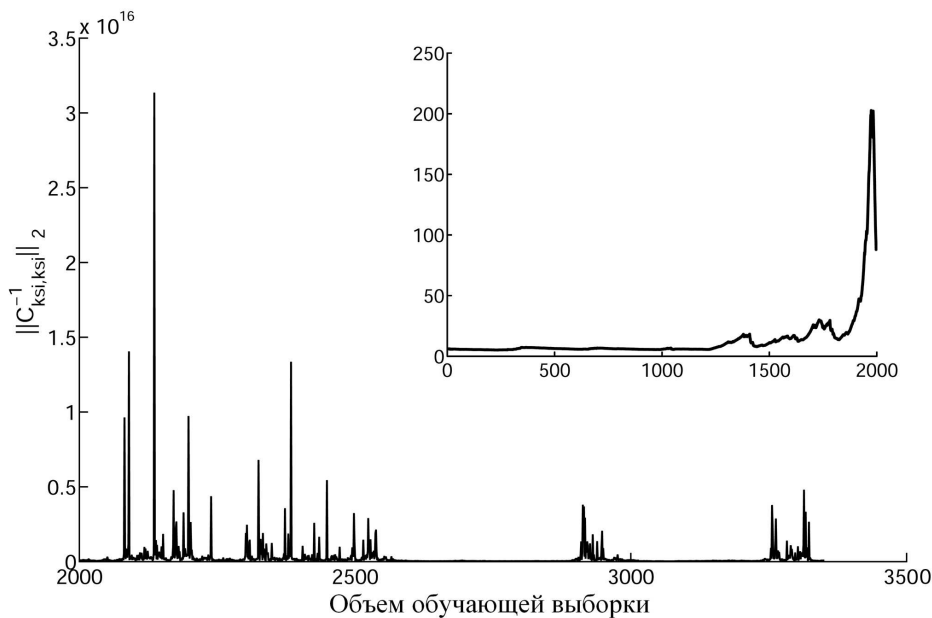
где * означает отсутствие наблюдения.

Вычисляя по формуле (3) автоковариационную матрицу, получим

$$C_{x,x} = \begin{pmatrix} 1.2 & 1.2 \\ 1.2 & 1.0 \end{pmatrix}.$$

Эта матрица имеет два собственных значения: положительное и отрицательное, в данном случае это 2.3 и -0.1 . Причем с ростом числа наблюдений в матрице X последнее собственное значение стремится к нулю, оставаясь отрицательным. Таким образом, у автоковариационной матрицы существует ненулевая вероятность появления крупномасштабного поля отрицательных собственных значений, сколь угодно близких к нулю, что делает ее обращение неустойчивым.

Это подтверждают результаты численных экспериментов для реального крупномасштабного поля приземной температуры в Московском регионе. Поведение второй нормы



Вторая норма обратной автоковариационной матрицы крупномасштабного поля в зависимости от объема выборки, на которой производилось оценивание вторых моментов. До объема около 2000 наблюдений (вставка) норма относительно стабильна. При дальнейшем росте числа наблюдений (основной рисунок) в связи с пропусками в данных наблюдаются очень значительные скачки этой величины

обращенной автоковариационной матрицы $C_{\xi,\xi}^{-1}$ в зависимости от объема выборки, на которой производилось оценивание, показано на рисунке.

Как оказывается, решить проблему неустойчивого обращения ковариационной матрицы можно путем заполнения пропусков в массиве данных, соответствующем крупномасштабному полю.

2. Методы восстановления пропущенных значений

Для заполнения пропущенных значений могут быть использованы самые разные методы временной и пространственной интерполяции. К первым, например, можно отнести такие простейшие процедуры, как подстановка вместо пропущенных наблюдений среднего значения временного ряда или среднего значения из некоторого “окна” вокруг точки с пропущенным наблюдением. С помощью простых методов обычно автоматически обрабатывают временные ряды при их анализе с использованием специальных программных пакетов.

Однако при работе с геофизическими данными более предпочтительны методы пространственной интерполяции, поскольку в этом случае важное значение имеет сохранение пространственных корреляционных структур. Ниже приведено более подробное описание двух методов пространственной интерполяции, которые использовались для заполнения пропусков в данных многолетних наблюдений метеопараметров на метеорологических станциях и которые были протестированы в численных экспериментах при решении задачи детализации крупномасштабного поля приземной температуры с помощью статистической модели.

Метод пространственной интерполяции Крессмана (КР) [10]. Пусть $\mathbf{x}_{\text{obs}} = (x_{\text{obs}}^{(1)}, x_{\text{obs}}^{(2)}, \dots, x_{\text{obs}}^{(n)})$ — вектор наблюдений метеопараметра на n станциях, тогда значение этого параметра в некоторой j -й точке сетки определяется формулой

$$x_{\text{grid}}^{(j)} = \frac{\sum_{i=1}^n w(x_{\text{obs}}^{(i)}, x_{\text{grid}}^{(j)}) x_{\text{obs}}^{(i)}}{\sum_{i=1}^n w(x_{\text{obs}}^{(i)}, x_{\text{grid}}^{(j)})}, \quad (4)$$

где $w(x_{\text{obs}}^{(i)}, x_{\text{grid}}^{(j)}) = \max\left(0, \frac{R^2 - d(x_{\text{obs}}^{(i)}, x_{\text{grid}}^{(j)})}{R^2 + d(x_{\text{obs}}^{(i)}, x_{\text{grid}}^{(j)})}\right)$ — весовая функция, которая равна 1,

если координаты точек пространства, соответствующих $x_{\text{obs}}^{(i)}$ и $x_{\text{grid}}^{(j)}$, совпадают, и $w(x_{\text{obs}}^{(i)}, x_{\text{grid}}^{(j)}) = 0$, если расстояние между этими точками $d(x_{\text{obs}}^{(i)}, x_{\text{grid}}^{(j)}) > R$, где R — заданный радиус действия.

Метод интерполяции бигармоническими сплайнами (ИБС) [11]. Пусть опять $\mathbf{x}_{\text{obs}} = (x_{\text{obs}}^{(1)}, x_{\text{obs}}^{(2)}, \dots, x_{\text{obs}}^{(n)})$ — вектор наблюдений какого-либо физического параметра на метеостанциях, а $\mathbf{x}_{\text{grid}} = (x_{\text{grid}}^{(1)}, x_{\text{grid}}^{(2)}, \dots, x_{\text{grid}}^{(m)})$ — искомый вектор значений в узлах сетки. Предположим, что ни одна станция не совпадает с узлом сетки. Для каждой пары компонент векторов \mathbf{x}_{obs} и \mathbf{x}_{grid} вычислим значение так называемой функции Грина, которая задается выражением

$$g(x_{\text{grid}}^{(j)}, x_{\text{obs}}^{(i)}) = d(x_{\text{grid}}^{(j)}, x_{\text{obs}}^{(i)})^2 (\ln(d(x_{\text{obs}}^{(i)}, x_{\text{grid}}^{(j)})) - 1), \quad i = 1, \dots, n, j = 1, \dots, m, \quad (5)$$

где $d(x_{\text{grid}}^{(j)}, x_{\text{obs}}^{(i)})$ — расстояние между точками пространства, соответствующими j -й и i -й компонентам векторов \mathbf{x}_{grid} и \mathbf{x}_{obs} .

Вычислив попарные значения функции для этих векторов, получим матрицу $G_{\mathbf{x}_{\text{grid}}, \mathbf{x}_{\text{obs}}} = \{g(x_{\text{grid}}^{(j)}, x_{\text{obs}}^{(i)})\}$, $i = 1, \dots, n$, $j = 1, \dots, m$. В j -й строке этой матрицы стоят значения функции Грина от j -й компоненты вектора \mathbf{x}_{grid} и компонент вектора \mathbf{x}_{obs} . Аналогично легко можно получить матрицу Грина $G_{\mathbf{x}_{\text{obs}}, \mathbf{x}_{\text{obs}}} = \{g(x_{\text{obs}}^{(i)}, x_{\text{obs}}^{(j)})\}$, $i, j = 1, \dots, n$, только для вектора \mathbf{x}_{obs} .

Выражение для вычисления искомого вектора значений метеопараметра на сетке имеет следующий вид:

$$\mathbf{x}_{\text{grid}} = G_{\mathbf{x}_{\text{grid}}, \mathbf{x}_{\text{obs}}} G_{\mathbf{x}_{\text{obs}}, \mathbf{x}_{\text{obs}}}^{-1} \mathbf{x}_{\text{obs}}. \quad (6)$$

Таким образом, пропущенное значение наблюдаемого метеопараметра на одной из станций можно заменить его значением, рассчитанным на основе наблюдений на других станциях с помощью выражения (4) или (6). Однако, как показали многочисленные эксперименты, наибольший эффект, по-видимому, дает комбинация двух этих методов, когда для заполнения пропусков на достаточно удаленных станциях (например, станциях в Московской области) используется метод Крессмана, а на станциях, имеющих поблизости соседей (например, в черте г. Москвы), заполнение пропусков производится методом ИБС.

3. Численные эксперименты

В качестве тестовой задачи для проверки и сравнения предложенных методов заполнения пропусков рассматривалась задача детализации мелкомасштабных полей экстремальных приземных температур (максимальной и минимальной за сутки) по их крупномасштабным значениям в Московском регионе. Для численного решения этой задачи использовалась разработанная ранее статистическая модель, которая, как показали более ранние исследования, позволяет восстанавливать максимальные и минимальные температуры на метеостанциях г. Москвы со среднеквадратичными ошибками 1.13 и 1.03 °С соответственно при естественной изменчивости 4.98 и 5.31 °С. Для калибровки

Точность восстановления мелкомасштабного поля максимальной и минимальной температур в Московском регионе в зависимости от применяемого метода заполнения пропусков

Метод заполнения пропусков	Априорная оценка абсолютной ошибки, °С	Априорная оценка относительной ошибки, %	Оценка абсолютной ошибки на тестовой выборке, °С	Оценка относительной ошибки на тестовой выборке, %
Максимальная суточная температура				
—	1.41	28.32	2.01	40.25
ИБС	1.43	28.60	1.68	33.59
КР	1.58	32.71	1.40	29.00
Комбинированный	1.56	32.06	1.54	31.71
Минимальная суточная температура				
—	1.65	30.70	2.53	47.06
ИБС	1.64	30.73	2.05	38.23
КР	1.59	30.93	1.60	31.15
Комбинированный	1.79	35.05	1.46	28.57

этой статистической модели для всего Московского региона использовалась выборка из массива данных наблюдений максимальной и минимальной температур на шестнадцати метеорологических станциях в г. Москве и Московской области за временной период с 01.01.1959 по 08.07.1986. Данные наблюдений предоставлены Гидрометцентром России. В качестве крупномасштабного поля использовались осредненные мелкомасштабные поля, а также данные реанализа NCAR/NCEP, заданные на сетке с разрешением $2.5 \times 2.5^\circ$.

Оценка ковариационных матриц производилась на основе как исходного массива данных, содержащего до 40 % пропусков, так и массивов данных, пропуски в которых были заполнены одним из описанных методов. Погрешности решения обратной задачи в Московском регионе, полученного согласно соотношению (2) и выполненного описанными методами для максимальных и минимальных температур в отдельности, сведены в таблицу.

Как показывают результаты, приведенные в таблице, заполнение пропусков не только приводит к устойчивости численного решения обратной задачи, но и позволяет повысить его точность. Наилучшие результаты удастся получить при заполнении пропусков предложенным комбинированным методом, а также методом Крессмана.

Выводы

1. Построена статистическая модель для восстановления детальной структуры полей максимальных и минимальных суточных значений приземной температуры в Московской области и г. Москве по их крупномасштабным значениям. Данная модель позволяет восстановить максимальные и минимальные температуры на метеостанциях г. Москвы со среднеквадратичными ошибками 1.13 и 1.03 °C, а на метеостанциях всего Московского региона 2.01 и 2.53 °C при естественной изменчивости этих полей 4.98 и 5.31 °C соответственно.

2. Показано, что оценки автоковариационных матриц геофизических полей, заданных массивами данных с пропусками, могут иметь сколь угодно близкие к нулю отрицательные собственные значения. Это, в частности, ведет к неустойчивому решению обратной задачи детализации крупномасштабных метеорологических полей вследствие неустойчивого обращения подобных матриц.

3. Предложен комбинированный метод заполнения пропусков, который позволяет в значительной степени повысить устойчивость решения рассматриваемой задачи по сравнению с существующими методами. При этом удастся не только добиться устойчивости при решении обратных задач подобного рода, но и повысить точность решения благодаря сохранению и воспроизведению пространственной корреляционной структуры исходного случайного поля.

Список литературы

- [1] BIERKENS M., FINKE P., DE WILLIGEN P. Upscaling and downscaling methods for environmental research // *Developments in Plant and Soil Sciences*. 2000. Vol. 88. 204 p.
- [2] MURPHY J. An evaluation of statistical and dynamical techniques for downscaling local climate // *J. Climate*. 1999. Vol. 12. P. 2256–2284.

- [3] SCHUBERT S., HENDERSON-SELLERS A. A statistical model to downscale local daily temperature extremes from synoptic-scale atmospheric circulation patterns in the Australian region // *Climate Dynamics*. 1997. Vol. 13. P. 223–234.
- [4] KOSTOPOULOU E., GIANNAKOPOULOS C., ANAGNOSTOPOULOU C. ET. AL. Simulating maximum and minimum temperature over Greece: a comparison of three downscaling techniques // *Theoretical and Applied Climatology*. 2007. Vol. 90, N 1–2. P. 65–82.
- [5] ДМИТРИЕВ Е.В., ЧАВРО А.И. Восстановление детальной структуры поля температуры в г. Москве по его крупномасштабным значениям // *Наукоемкие технологии*. 2003. Т. 4, № 6. С. 41–49.
- [6] УСПЕНСКИЙ А.Б. Обратные задачи математической физики — анализ и планирование экспериментов // *Мат. методы планирования эксперимента*. Новосибирск: Наука, 1981. С. 199–242.
- [7] ПЫТЬЕВ Ю.П. Математические методы интерпретации эксперимента. М.: Высш. школа, 1989. 351 с.
- [8] ЧАВРО А.И., ДЫМНИКОВ В.П. Методы математической статистики в задачах физики атмосферы. М.: ИВМ РАН, 2000. 210 с.
- [9] ЧАВРО А.И., ДМИТРИЕВ Е.В. Статистическая модель восстановления региональной структуры геофизических полей // *Метеорология и гидрология*. 2002. № 6. С. 39–49.
- [10] CRESSMAN G.P. An operational objective analysis system // *Monthly Weather Review*. 1959. Vol. 87. P. 367–374.
- [11] SANDWELL D.T. Biharmonic spline interpolation of GEOS-3 and SEASAT altimeter data // *Geophys. Research Lett.* 1987. Vol. 2. P. 139–142.

Поступила в редакцию 21 февраля 2008 г.