УДК 551.464:681.32

Е.Ф. Васечкина, В.Д. Ярин

Прогноз временных рядов с применением метода аналогов

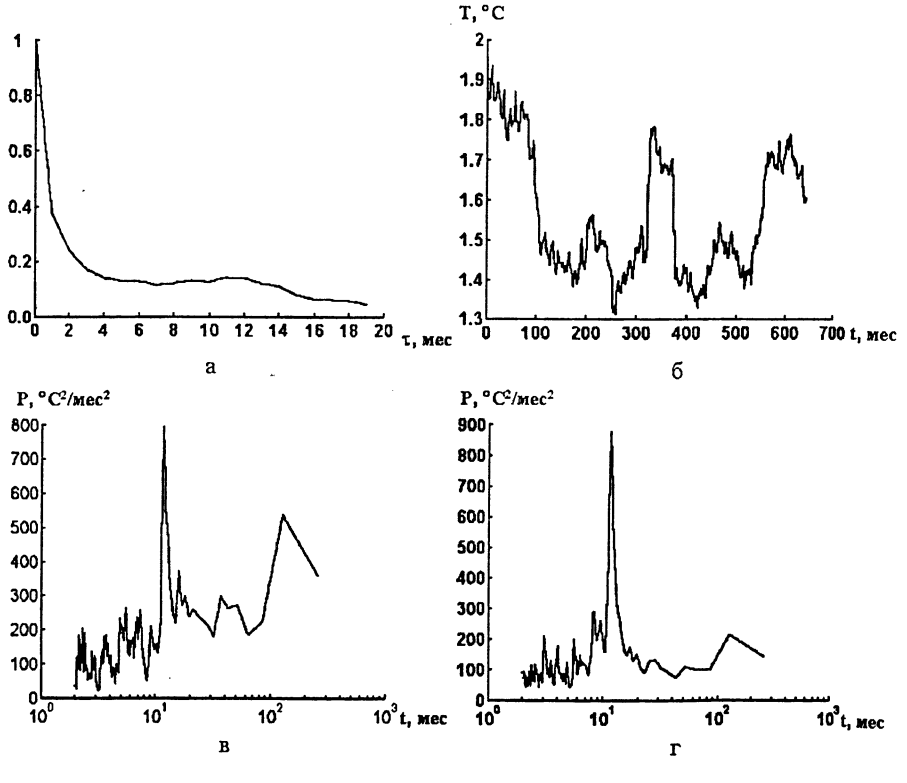
Рассматривается алгоритм прогнозирования нестационарных временных рядов, который основан на методе аналогов. Поскольку для подгонки параметров оптимальной модели прогноза требуется перебор большого количества вариантов, описывается генетический алгоритм, который при этом использовался. Рассмотрены несколько способов построения модели прогноза. По результатам расчетов выбран способ, который дает наименьшую среднеквадратическую ошибку. Определены параметры модели, влияющие на качество прогноза. Предлагаемый метод тестируется на данных реанализа (проект *NCEP/NCAR*) об аномалиях среднемесячной температуры воздуха у поверхности Земли за 58 лет. Результаты прогнозирования сравниваются с оценками, полученными методом линейной регрессии. Показано, что метод аналогов дает удовлетворительные результаты в условиях, когда регрессионные методы приводят к ошибке, равной дисперсии прогнозируемого ряда.

Введение. Несмотря на значительное число публикаций по разработке методов и практическому применению прогнозирования временных рядов, задача получения адекватного прогноза временного ряда остается попрежнему актуальной, поскольку существующие методы часто дают слишком большую ошибку. Это особенно справедливо для нестационарных и слабо коррелированных рядов (т.е. рядов, порождаемых нелинейными процессами). В работах [1, 2] предложен метод восстановления пропусков в данных наблюдений с использованием разложений по системе эмпирических ортогональных функций. При определении коэффициентов разложения в моменты времени, для которых существовали пропуски в данных, использовался генетический алгоритм, подробно описанный в [2]. После разложения поля по собственным функциям ковариационной матрицы для прогноза достаточно рассчитать будущие значения коэффициентов разложения. Однако поведение этих коэффициентов, начиная со второй моды, носит характер «белого шума» и прогнозу практически не поддается. Поэтому в настоящей работе предлагается новый способ прогнозирования временных рядов, основанный на давно известном и широко используемом в различных областях науки, в том числе и в метеорологии, например [3], методе аналогов.

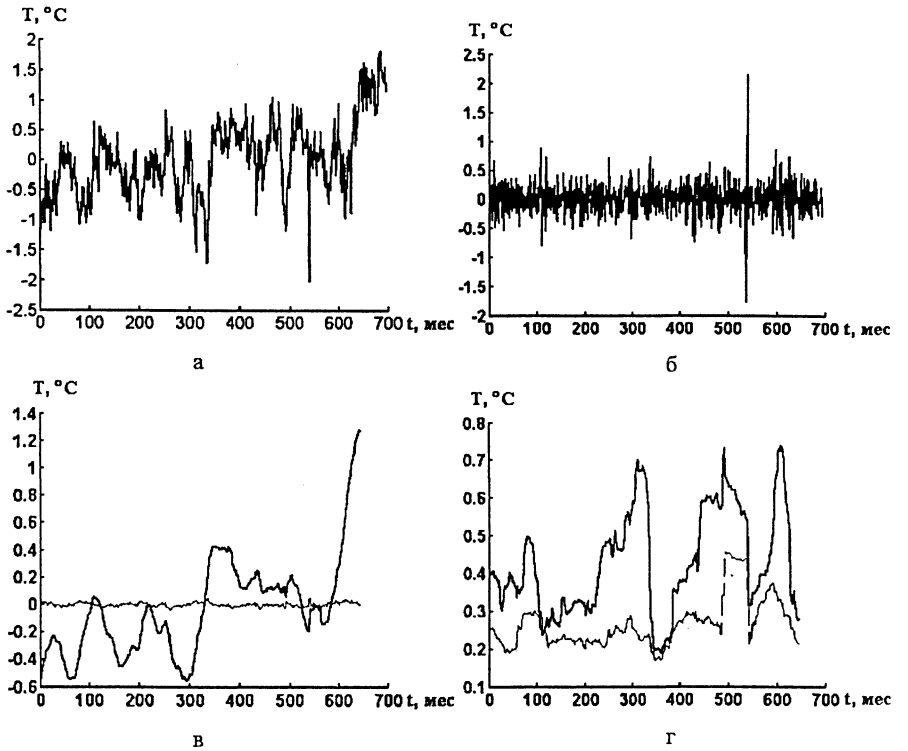
Данные наблюдений. Для изучения возможностей и оценки эффективности предлагаемого метода в расчетах использованы данные реанализа приземной температуры воздуха (проект *NCEP/NCAR*), опубликованные на сайте климатического диагностического центра *NOAA-CIRES* в Колорадо [4]. Данные представляют собой массив среднемесячной температуры воздуха в узлах регулярной сетки 73 × 144 с шагом в 2,5°, охватывающей весь земной шар. Временной диапазон данных составляет 58 лет (с января 1948 г. по январь 2006 г.), т.е. временные ряды в каждом узле сетки содержат 697 отсчетов. Из исходных данных были вычтены неоднородные по пространству многолетние средние значения температуры по каждому месяцу. В дальнейших

расчетах использовался этот массив аномалий. Рассмотрим более подробно статистические характеристики рядов.

На рис.1 приведены автокорреляционная функция, скользящее среднеквадратическое отклонение и спектры временных рядов, полученные на отрезках в 256 отсчетов в начале и в конце полного ряда (697 отсчетов). Эти характеристики рассчитывались по ансамблю, в качестве которого использовались временные ряды в двадцати соседних узлах сетки. Приведенные характеристики свидетельствуют о существенной нестационарности и слабой коррелированности прогнозируемых рядов. На спектре, рассчитанном по начальному отрезку ряда, имеются пики на периодах, соответствующих одному и двум годам, в конце ряда двухгодичного пика нет. На графике среднеквадратического отклонения наблюдаются нерегулярные колебания. Кроме того, как видно из рис. 2, *а*, присутствует квазипериодический тренд. Действительно, прогноз по модели линейной регрессии дает даже на один шаг вперед смещение оценки до ± 0,7 – 0,9°С при среднеквадратической ошибке прогноза, составляющей 75 – 100 % от среднеквадратического отклонения искомого ряда. Поскольку линейный подход в данном случае неприменим, авторами была предпринята попытка использовать для прогноза метод аналогов.



Р и с. 1. Статистические характеристики: временная автокорреляционная функция (*а*), скользящее среднеквадратическое отклонение по 50 отсчетам (*б*), спектральная плотность, рассчитанная на отрезках в начале (*в*) и в конце (*г*) ряда по данным в 20 соседних узлах сетки



Р и с. 2. Временной ряд аномалий среднемесячной температуры в одном из узлов сети – *а;* первая производная по времени от предыдущего ряда – *б;* скользящее среднее исходного ряда (сплошная линия) и его производной по времени (пунктир) – *в;* скользящее среднеквадратическое отклонение исходного ряда (сплошная линия) и его производной по времени (пунктир) – *г.* Скользящее осреднение выполнено по 50 отсчетам

Метод аналогов. Суть метода очень проста. В его основе лежит предположение, что если некоторый отрезок ряда длиной n в некотором смысле «похож» на другой отрезок такой же длины, то с некоторой уменьшающейся вероятностью значения $n+1, n+2, \ldots$ (продолжения этого отрезка) также «похожи» в этом же смысле на соответствующие значения другого отрезка. То есть необходимо подобрать из прошлых значений ряда близкие к текущей ситуации и на основе известного поведения в прошлых ситуациях построить прогноз развития текущих событий. В применении к прогнозу временных рядов этот метод использовался следующим образом.

Рассмотрим ряд аномалий температуры в некотором узле сетки $X(t_i)$, $i = 1, \ldots, N$, где N – длина ряда. Представим этот ряд в виде таблицы A, каждая строка которой содержит D предыдущих и последующих значений $X(t_i)$ относительно выбранной точки t_k:

$$X(t_k - n), \quad X(t_k - n + 1), \ldots, X(t_k - 1), \quad X(t_k), \quad X(t_k + 1), \ldots, X(t_k + S), \qquad (1)$$

где S – заблаговременность прогноза; n – длина учитываемой предыстории ряда ($n = D - S - 1$); $k = 1 \ldots K$, здесь K – число строк в таблице ($K = N - D + 1$). Размеры полученной таблицы $K \times D$.

Будем рассматривать строку построенной таблицы как некий паттерн возможной динамики ряда аномалий от момента времени ($t_k - n$) к моменту ($t_k + S$). Обозначив один из моментов времени t_m как условно текущий, можно попытаться получить прогноз будущих значений ряда $X(t_m + S)$, используя другие строки матрицы A как возможные варианты. Естественно ожидать, что строки, более «похожие» на искомую строку m, окажутся более информативными при построении прогноза. Критерий «похожести» можно выбирать по-разному, самым простым способом представляется сумма квадратов отклонений между значениями двух различных строк в матрице A. Эту сумму необходимо считать только в левой части таблицы, где расположены предыдущие значения $X(t)$ относительно $X(t_k)$, поскольку будущие значения искомой строки m нам неизвестны.

Детальное изучение рядов аномалий и полученных путем описанного выше преобразования таблиц паттернов показало, что самый простой путь не является оптимальным. Первое, что необходимо учесть, это нестационарность. Ряды аномалий температуры имели меняющиеся во времени скользящее среднее и среднеквадратическое отклонение и даже могли содержать в себе некоторый тренд, как, например, ряд, представленный на рис. 2. Поиск же аналогов искомой строки – паттерна требовал стационарности ряда относительно среднего, поскольку иначе похожие по форме паттерны могли бы отличаться существенно по абсолютной величине значений, вследствие чего выбор аналогов был бы необоснованно ограничен. Поэтому ряды аномалий при построении таблицы A были заменены на их первые производные по времени, что существенно уменьшило изменчивость скользящих средних, колебания среднеквадратических отклонений при этом уменьшили свою амплитуду (рис. 2, *б*,– *г*). Стандартная процедура приведения ряда к стационарному виду путем деления на переменное среднеквадратическое отклонение не применялась, поскольку предварительные расчеты показали, что в этом случае число возможных аналогов было неоправданно завышено, т.к. выбросы значений аномалий содержат необходимую информацию для отсечения ложных аналогов.

Практические эксперименты показали, что эффективность прогноза зависит главным образом от пяти основных факторов:

– длины используемой предыстории – «памяти» ряда;

– выбора критерия оценки близости аналогов к искомому паттерну;

– того, насколько хорошо удастся подобрать аналоги (имеются ли в истории ряда похожие паттерны);

– способа построения прогностической модели на основе ближайших аналогов;

– числа аналогов, которые так или иначе используются при получении прогноза.

Начнем с выбора критерия близости аналогов к искомому паттерну. Эмпирически установлено, что наилучшим вариантом является комбинация

двух величин: взвешенной суммы квадратов отклонений аналога от искомого паттерна и разности соответствующих значений второй производной по времени, определяющей близость форм двух строк – паттернов. Обозначим через a_1 первую (искомую) строку таблицы, для которой нужно получить прогноз, а все прочие строки – через a_i. Тогда выражение для функции близости строки a_i к искомой строке a_1 можно записать следующим образом:

$$Q_i = \sum_{j=-n}^{j=0} w_{ij}(a_{ij} - a_{1j})^2 + \frac{C}{n}\sum_{j=-n+1}^{j=0}\left|(a_{1j} - a_{1j-1}) - (a_{ij} - a_{ij-1})\right|. \qquad (2)$$

Так как исходный ряд заменен на свою первую производную по времени, второй член выражения (2) соответствует средней абсолютной величине разности вторых производных. Весовая функция подбирается таким образом, чтобы значения ряда с меньшим запаздыванием имели бóльшую значимость при расчете критерия. В простейшем случае она имеет вид:

$$w_j = \frac{j+n+1}{\sum_{i=1}^{n+1} i}, \qquad j = -n,...,0, \qquad \sum_{j=-n}^{0} w_j = 0. \qquad (3)$$

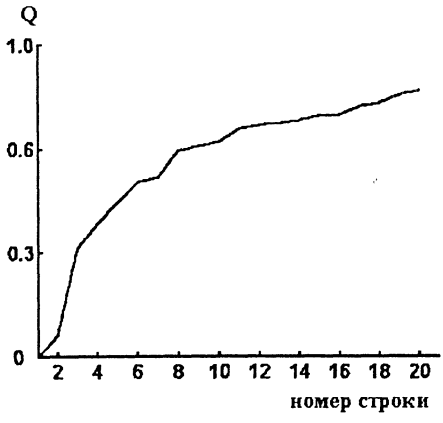
Величина коэффициента C определяет вклад каждой из частей комбинированного критерия и зависит от конкретного ряда. В дальнейшем мы вернемся к обсуждению получения ее оценки.

Можно предложить разные способы построения прогностической модели на основе информации, содержащейся в ближайших строках-аналогах. Авторами были рассмотрены и оценены следующие варианты.

1. Построение модели линейной авторегрессии, аргументами в которой являются данные предыстории, и использование ее для получения прогноза искомой строки a_1.

2. Получение оценки ковариационной матрицы по данным предыстории $a_i(t_k - n)...a_i(t_k)$ с последующим расчетом эмпирических ортогональных функций (ЭОФ) и подбором коэффициентов разложения для искомой строки a_1. Подбор осуществляется с помощью генетического алгоритма на данных $a_1(t_k - n)...a_1(t_k)$ (алгоритм восстановления пропусков в данных, подробно изложенный в [2]). Тогда свертка найденных коэффициентов с набором ЭОФ позволяет получить полную строку a_1, включающую и правую часть – будущие значения $a_1(t_k + 1)...a_1(t_k + S)$. Практическое применение этого способа показало, что на данных предыстории искомый паттерн подбирается с очень малой ошибкой, не превышающей, как правило, 10^{-4}, однако правая часть паттерна часто может иметь куда большую ошибку.

3. Получение взвешенного среднего будущих значений ближайших строк-аналогов $a_i(t_k + 1)...a_i(t_k + S)$ и использование этого среднего в качестве прогноза. Весовая функция строится исходя из оценки критерия близости строки-аналога к искомому паттерну. Чем меньше отличия аналога, тем больший вес придается его значениям при построении прогноза.

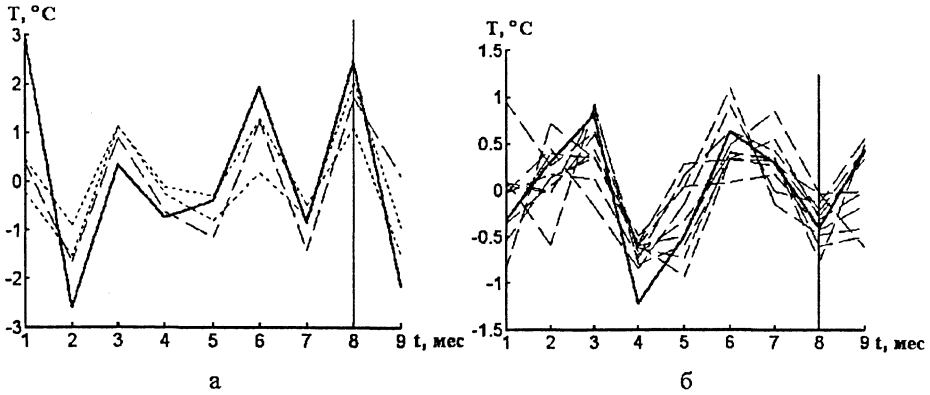


Р и с. 3. Пример зависимости нормированной функции близости аналогов к искомому паттерну Q от номера строки-аналога. Аналоги отсортированы по мере возрастания их отличий от искомого паттерна

Естественно, что качество этих способов построения модели во многом зависит от количества используемых строк-аналогов. Для оценки коэффициентов авторегрессионной модели, имеющей хорошие обобщающие свойства, и для расчета более или менее устойчивой системы ЭОФ, которая бы могла описывать и другие паттерны, отличные от использованных при ее оценивании, необходимо достаточное число строк-аналогов. Однако при имеющейся длине ряда функция Q растет очень быстро, и действительно «похожих» аналогов оказывается недостаточно для первых двух способов построения прогностической модели. Этим, вероятно, можно объяснить тот факт, что третий, наиболее грубый, алгоритм оказывается в среднем более устойчивым способом прогноза. Рис. 3 иллюстрирует типичное поведение нормированной функции близости аналогов к искомому паттерну Q в зависимости от номера строки в ранжированной по Q таблице.

На рис. 4 показаны два варианта подбора аналогов, причем количество аналогов, использованных для получения прогноза, в каждом случае является оптимальным, т.е. получено путем подбора для конкретного ряда и определенной точки прогноза (временного отсчета) t_k. Способ такого подбора будет обсуждаться ниже. Легко видеть, что в первом случае (рис. 4, *а*) аналоги довольно близки к искомому паттерну и ближайший аналог довольно точно предсказывает будущее значение. Во втором случае (рис. 4, *б*) количество аналогов больше и на шаге $t_k + 1$ их поведение неопределенно, кривые расходятся, образуя веер возможных направлений развития процесса. Аналоги, наиболее хорошо совпадающие на шаге $t_k + 1$ с истинным рядом, не являются в этом примере самыми близкими по введенному критерию (2).

Одним из главных факторов эффективности прогнозирования является выбор подходящей длины используемой предыстории ряда n. Этот параметр наряду с коэффициентом C и количеством ближайших строк-аналогов, участвующих в построении модели, необходимо подбирать при каждом прогнозировании. Эмпирически установлено, что оптимальное соотношение этих параметров меняется при движении вдоль временного ряда, следовательно, подгонять их нужно как можно ближе к точке, где необходимо получить прогноз. Подгонка параметров осуществлялась с помощью генетического алгоритма (ГА).



Р и с. 4. Примеры подбора аналогов для двух разных рядов аномалий (сплошная линия – искомый паттерн, штриховые – ближайшие аналоги). Количество аналогов для обоих вариантов *а* и *б* получено подбором модели с помощью ГА. Прогноз осуществляется в обоих случаях на один шаг вперед, положение момента времени t_k обозначено прямой вертикальной линией

Алгоритм построения прогностической модели. ГА является универсальным алгоритмом поиска решения задачи в пространстве ее всевозможных решений. Он относится к классу так называемых эволюционных алгоритмов, в основе которых лежит идея поиска решения задачи путем имитации «биологической» эволюции [5]. Алгоритм генерирует и поддерживает «популяцию» потенциальных решений исследуемой проблемы – «индивидуумов». Некоторые из этих решений используются для создания новых решений путем применения к ним специальных операторов, имитирующих процессы реальной биологической эволюции. Это операторы селекции, скрещивания (кроссовера) и мутации. Таким образом, в алгоритме реализуется дарвиновская триада движущих сил эволюции: наследственность, изменчивость, отбор. Потенциальные решения для оператора скрещивания отбираются на основе оценки их функции качества согласно введенным критериям, определяемым конкретной задачей. Решения, обладающие лучшим качеством в смысле соответствия введенным критериям, имеют бóльшие шансы «дать потомков» и пройти селекционный отбор в новое поколение. Новые поколения популяции потенциальных решений рассчитываются последовательно до выполнения алгоритмом некоторого условия, обычно связанного с глубиной найденного экстремума функции качества.

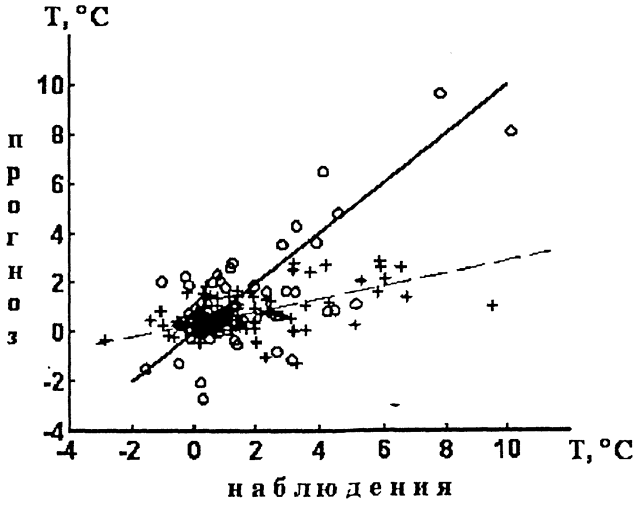
В нашем случае индивидуум популяции представлял собой комбинацию трех параметров, однозначно определяющих прогностическую модель: ширину таблицы паттернов D (или длину используемой предыстории ряда n), коэффициент соотношения двух частей комбинированного критерия C и количество используемых при построении прогностической модели строк-аналогов M. Эти параметры, записанные последовательно в двоичном виде в строку, представляли собой «генотип» экземпляра популяции. При этом учитывался допустимый интервал их изменчивости, определенный эмпирически. Для D он составлял 6 – 13 при $S = 1$, для C 0 – 0,7 с дискретностью 0,1, для M – 3 – 18. Соответственно для приведения диапазона изменений параметров

к удобному для записи числа в двоичном виде вводились корректирующие коэффициенты. Например, пусть $D = 10$, $C = 0{,}5$, $M = 9$. Величина $D - 6$ преобразуется в двоичную строку из трех символов «0 0 1», коэффициент $10C = 5$ в «1 0 1», а $M - 3 = 6$ в строку из четырех двоичных единиц «0 1 1 0». Тогда полная строка генотипа будет включать в себя 10 двоичных разрядов – «0 0 1 1 0 1 0 1 1 0». Диапазоны допустимых изменений необходимо было сделать как можно у́же, чтобы использовать для кодировки параметров как можно меньше двоичных разрядов. Это связано с затратами времени на поиск оптимального решения. Выбранные диапазоны соответствовали третьему способу построения прогностической модели, результаты применения которого и изложены ниже.

Практически алгоритм подбора параметров модели работал следующим образом. С использованием датчика случайных чисел формировалась начальная популяция удвоенного объема $2P$, где P – мощность популяции. Каждый экземпляр популяции представлял собой двоичную строку, состоящую из 10 разрядов. Выбор P зависит от длины генотипа L, при $L = 10$ P принималась равной 30. Оценивалось качество каждого из потенциальных решений. Для этого в конце ряда выделялся небольшой отрезок (5 – 7 точек), на котором с помощью модели, определенной набором параметров, записанных в генотипе, строился прогноз. Средняя ошибка прогноза давала оценку функции качества индивидуума популяции. Таблица экземпляров популяции ранжировалась по возрастанию средней ошибки модели, т.е. по убыванию функции качества решений. Далее к верхней половине таблицы объемом P (лучшие потенциальные решения) применялся оператор одноточечного кроссовера, действие которого заключалось в том, что отобранные для скрещивания два экземпляра популяции в случайной точке обменивались своими частями: к началу первого добавлялся конец второго и наоборот. Затем к каждому новому экземпляру популяции применялся оператор мутации, который с некоторой малой вероятностью мог изменить значение любого бита строки генотипа на противоположное. Оба новых экземпляра оценивались, и в том случае, если новые экземпляры имели лучшее качество, тогда они заменяли собой родительские. В результате применения этих операторов популяция вновь увеличивалась вдвое, и к ней применялся оператор селекции. Из $2P$ «родителей» и «потомков» отбирались P экземпляров, которые формировали новое поколение популяции. Отбор производился с вероятностью, пропорциональной функции качества индивидуума. Лучшее полученное решение сохранялось в новом поколении вне конкуренции. Для предотвращения преждевременного выравнивания популяции по функции качества осуществлялось слежение за ее градиентом. При уменьшении этого градиента до определенного малого значения (вырождение популяции) производилась процедура «элиминации», при которой все экземпляры популяции, за исключением нескольких лучших, заменялись на новые, сгенерированные с помощью датчика случайных чисел. Процесс последовательного применения операторов кроссовера, мутации и селекции к новым поколениям популяции потенциальных решений повторялся вплоть до достижения устойчивого минимума функции качества (один и тот же экземпляр популяции должен был не менее 10 раз появиться в первой строке таблицы). Лучшее полученное решение использовалось затем для по-

строения собственно прогноза временного ряда на S шагов вперед. Было замечено, что чем ближе отрезок, на котором осуществлялась подгонка модели, к моменту t_k, тем меньшую ошибку дает модель.

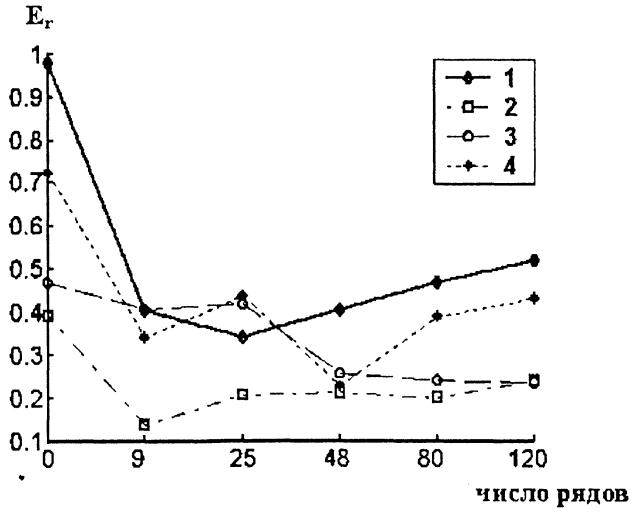
Результаты расчетов. Метод оценивался по относительной среднеквадратической ошибке E_r и смещению прогнозов временных рядов аномалий температуры воздуха в 20 случайным образом отобранных узлах сетки. Для каждого ряда рассчитывались прогнозы в пяти последовательных точках с заблаговременностью от 1 до 3 месяцев. На рис. 5 показаны результаты сопоставления прогностических и истинных значений аномалий температуры. Необходимо отметить, что качество прогноза для различных рядов отличалось весьма существенно. Прогнозы, полученные для точек, ближайших к отрезкам, на которых подгонялись модели, в 70 % случаев имели среднеквадратическую ошибку не выше 1/3 среднеквадратического отклонения соответствующего временного ряда. Существовали, однако, ряды, для которых ошибка была намного выше (сопоставима с изменчивостью ряда). Вероятно, это объясняется тем, что для рассматриваемых рядов на отрезке построения прогноза не находилось достаточно близких по форме аналогов. Длина рядов оказывается здесь ключевым фактором, определяющим качество прогнозирования.



Р и с. 5. Результаты сопоставления прогностических и истинных значений аномалий температуры, полученные для 20 временных рядов, случайным образом выбранных из массива аномалий. Прогноз рассчитывался для пяти точек каждого ряда с заблаговременностью 1 мес. Кружочками обозначены точки, соответствующие прогнозам, полученным методом аналогов, крестиками – методом линейной регрессии. Прямая линия соединяет точки с нулевой ошибкой прогноза

В связи с этим была предпринята попытка расширить пространство выбора аналогов за счет добавления временных рядов в соседних узлах сетки. Действительно, коэффициент пространственной корреляции уменьшается достаточно медленно и для соседних точек (на расстоянии 2 – 5 Δx) составля-

ет ~ 0,98 – 0,75, поэтому можно ожидать, что временные колебания аномалий в соседних узлах сетки будут иметь похожий характер. Были выполнены расчеты, в которых к таблице аналогов (1) исходного ряда добавлялись таблицы, построенные таким же образом для соседних узлов, отстоящих от исходного на $\pm h$, $h = 1...5$. Исходный ряд при этом оказывается в центре квадрата с шириной $2h + 1$. Количество дополнительных временных рядов зависит от h как $(2h+1)^2 - 1$. В ходе численных экспериментов было установлено, что оптимальное число h для разных точек различно и этот параметр также нуждается в подгонке. На рис. 6 показана зависимость относительной среднеквадратической ошибки прогноза для нескольких рядов (выбранных случайным образом) от количества дополнительно привлеченных соседних узлов. Как видим, в первом случае оптимальным числом является 25 ($h = 2$), во втором – 9 ($h = 1$), в третьем ошибка плавно снижается и, хотя не проходит минимума, меняется незначительно – от $h = 4$ до $h = 5$. В четвертом случае оптимальное число дополнительных рядов 48 ($h = 3$). Очевидно, что вследствие нестационарности рядов подгонка параметра h должна осуществляться на отрезке, максимально приближенном к точке построения реального прогноза.



Р и с. 6. Зависимость относительной среднеквадратической ошибки прогноза от количества дополнительно привлеченных рядов, которые используются для расширения пространства выбора аналогов

В среднем смещение прогноза было весьма незначительным – 0,05°С, относительная среднеквадратическая ошибка составила ~ 0,51 при заблаговременности прогноза 1 месяц; 0,62 при $S = 2$ мес и 0,77 при $S = 3$ мес. Соответственно прогнозы, рассчитанные методом линейной регрессии для этих же рядов, имели ошибку 0,86 при $S = 1$ мес (среднее смещение 0,87°С), 0,87 при $S = 2$ мес (смещение 1,14°С) и 0,89 при $S = 3$ мес (смещение 1,33°С). Штриховая линия на рис. 5 представляет собой линейную аппроксимацию абсолютных ошибок прогноза аномалий температуры по модели линейной регрессии,

построенную методом наименьших квадратов. Аналогичная прямая для метода аналогов практически совпадает с линией нулевой ошибки прогноза, которой соответствует линия, проходящая через начало координат под углом 45°. Действительно, как уже отмечалось, средняя абсолютная ошибка метода аналогов близка к нулю, в то время как регрессионный прогноз дает смещенную оценку, тем бóльшую, чем дальше прогнозируемые значения отстоят от среднего значения ряда, равного в данном случае нулю. Этот факт также подчеркивает то обстоятельство, что прогноз по модели линейной регрессии для некоррелированных рядов дает оценку, близкую к среднему значению ряда. В то же время метод аналогов лишен этих недостатков.

Заключение. Расчеты, выполненные с помощью предлагаемого метода, показали, что метод аналогов дает возможность получать прогнозы с удовлетворительной точностью в тех случаях, когда регрессионные методы не дают удовлетворительных результатов. Поскольку построение оптимальной прогностической модели с подгонкой нескольких параметров требует перебора большого количества вариантов, необходимо применение генетического алгоритма.

СПИСОК ЛИТЕРАТУРЫ

1. *Васечкина Е.Ф., Ярин В.Д.* Генетический алгоритм в задаче реконструкции гидрометеорологических полей // Системы контроля окружающей среды-2001. – Севастополь: ЭКОСИ-Гидрофизика, 2002. – С. 141 – 145.
2. *Васечкина Е.Ф., Ярин В.Д.* Использование генетического алгоритма в задаче восстановления пропущенных данных // Морской гидрофизический журнал. – 2002. – №4. – С. 30 – 39.
3. *H.M. Van Den Dool.* A new look at weather forecasting through analogues // Month. Wea. Rev. – 1989. – 117, Oct. – P. 2230 – 2247.
4. *NCEP* Reanalysis data. NOAA-CIRES ESRL/PSD Climate Diagnostics branch. – Boulder, Colorado, USA, http://www.cdc.noaa.gov/
5. *Beasley D., Bull D.R., Martin R.R.* An Overview of Genetic Algorithms: Part I, Fundamentals. – University Computing. – 1993. – 15(2). – P. 58 – 69.

Морской гидрофизический институт НАН Украины,
Севастополь

Материал поступил
в редакцию 20.03.06
После доработки 10.05.06

ABSTRACT Algorithm for a non-stationary time series forecast based on the analog method is considered. Fitting of the parameters of the optimum forecast model requires scrutiny of numerous variants; therefore the genetic algorithm used for this purpose is described. Several alternatives of the model construction are considered. The calculation results provide the method that gives a minimum root-mean-square error. The model parameters influencing the forecast quality are determined. The proposed method is tested using the reanalysis data (NCEP/NCAR project) on the anomalies of the monthly average air temperature near the Earth surface for 58 years. The forecasting results are compared with the estimates obtained by the linear regression method. It is shown that the analog method yields satisfactory results in the conditions when the regression methods give the error equal to the dispersion of the predicted series.