АЛГОРИТМ ВРЕМЕННОЙ ЭКСТРАПОЛЯЦИИ ДВУМЕРНЫХ ПОЛЕЙ

Ярин В.Д., Васечкина Е.Ф.

Морской гидрофизический институт
НАН Украины
г. Севастополь, ул. Капитанская, 2
E-mail: evasvyar@bios.iuf.net

Предлагается новый способ прогноза временных рядов, основанный на поиске аналогов в предыстории процесса. Алгоритм может использоваться также для временной экстраполяции двумерных полей, если для них может быть построена система эмпирических ортогональных функций. Применение алгоритма иллюстрируется прогнозом поля температуры поверхности земного шара.

В последнее время накоплены значительные массивы архивных данных наблюдений по различным гидрометеорологическим характеристикам. Эти данные используются для верификации как эмпирических регрессионных моделей, так и моделей, основанных на уравнениях термогидродинамики. Для эффективной проверки работоспособности моделей необходима последовательность значений физических полей в узлах некоторой пространственно-временной сетки. В архивных данных по различным причинам имеются пропуски как по времени, так и по пространству. Вследствие этого задача оптимального восстановления пропущенных данных наблюдений является актуальной. Из существующих подходов и методов решения этой проблемы особо можно выделить подход, основанный на разложении полей по базису эмпирических ортогональных функций (ЭОФ), поскольку ряды разложения по этим функциям обладают максимальной скоростью сходимости, кроме того, попутно решается задача компактного адекватного малопараметрического представления архивных данных. Последнее обстоятельство является немаловажным в связи с тем, что при интенсивном использовании спутниковых методов наблюдений объём архивов быстро растет. Применение метода разложения по базису ЭОФ в геофизических исследованиях восходит к работам [1, 2]. Различным аспектам использования метода, вопросам вычисления системы ЭОФ посвящены работы [2 - 4]. В [5] приведена обширная библиография работ по рассматриваемой теме, опубликованных в последние годы. Особенностью применения ЭОФ в задаче восстановления пропусков в измерениях является тот факт, что в этом случае коэффициенты разложения нельзя рассчитать точно, а приближенная оценка получается за счет сокращения общего числа опорных функций (ЭОФ). В результате восстанавливаемое поле в зависимости от различных факторов может получиться в значительной степени искаженным. В работах [6, 7] была показана эффективность применения генетического алгоритма для расчета коэффициентов разложения поля в сравнении с другими методами при большом количестве пропусков. При восстановлении профилей гидрохимических характеристик, измеренных в Черном море и содержащих большое количество пропущенных горизонтов [6], а также полей среднемесячного приземного давления в Северном полушарии за период 1891 – 1990 гг. [7], рассматриваемый метод позволил осуществить интерполяцию и экстраполяцию по пространственным переменным с приемлемой точностью.

Ряд работ [8, 9] был посвящен прогнозу коэффициентов разложения по времени, поскольку успешное решение такой задачи позволяет легко получать прогноз двумерного поля. Однако временные ряды коэффициентов являются плохо коррелированными, близки к белому шуму и плохо поддаются прогнозированию. Поскольку малые отклонения коэффициентов приводят к большим искажениям результирующего поля, такой подход оказывается мало пригодным для временной экстраполяции полей.

В настоящей работе предлагается иное решение этой задачи, основывающееся на возможности восстановления поля по очень малому числу точек относительно числа используемых ЭОФ. Кроме того, акцент делается на случае «малой выборки», когда число реализаций, имеющихся для построения системы ЭОФ – m, значительно меньше числа точек на карте – n. При этом ковариационная матрица имеет большую размерность ($n \times n$), но ее ранг будет не выше m, т.е. не менее ($n - m$) ее собственных чисел будут равны нулю. Тогда вместо ковариационной матрицы $\overline{\mathbf{FF}^*}$, где $\mathbf{F}$ – выборочная матрица размерностью ($n \times m$),

следует рассматривать матрицу $\overline{\mathbf{F}^*\mathbf{F}}$, размер которой $(m \times m)$ [2]. Собственные числа этих матриц совпадают, а между векторами существует простое соотношение

$$\mathbf{U} = \mathbf{F}\mathbf{V}\mathbf{D}^{1/2} \qquad (1)$$

где $\mathbf{U}$ – собственные вектора ковариационной матрицы $\overline{\mathbf{F}\mathbf{F}^*}$, $\mathbf{V}$ – собственные вектора матрицы $\overline{\mathbf{F}^*\mathbf{F}}$, $\mathbf{D}$ – диагональная матрица собственных чисел матрицы $\overline{\mathbf{F}\mathbf{F}^*}$. Из (1) следует представление исходной матрицы в виде

$$\mathbf{F} = \mathbf{U}\mathbf{D}^{1/2}\mathbf{V}, \qquad (2)$$

которое удобно использовать для расчета собственных функций в том случае, когда обычный способ неприменим из-за большой размерности полей в выборке (так называемый SVD – метод). Оценка ковариационной матрицы $\overline{\mathbf{F}\mathbf{F}^*}$ по выборке объема $m < n$ обладает слабой устойчивостью. При добавлении новой реализации к выборке появляется новое собственное число, вообще говоря, не равное нулю, и изменяются другие собственные числа. Причем относительно больше изменяются малые числа. Соответствующие малым числам собственные функции также менее устойчивы, чем первые функции, соответствующие большим числам. Поэтому не имеет смысла при аппроксимации поля учитывать собственные функции с большими номерами, содержащие много шума. Вопрос о том, сколько функций следует учитывать, зависит от конкретной задачи, но для обеспечения устойчивости их число должно быть намного меньше объема выборки. Экстраполяция поля во времени предполагает перенесение статистических свойств имеющейся выборки на новую реализацию, поэтому желательно, чтобы этот перенос был оправдан.

После разложения поля по собственным функциям ковариационной матрицы для прогноза поля достаточно спрогнозировать будущие значения коэффициентов разложения. Однако, поведение этих коэффициентов, начиная со второй моды, носит характер «белого шума» и прогнозу практически не поддается. Поэтому более целесообразно прогнозировать будущие значения самого поля в некоторых точках, а затем использовать метод интерполяции, подробно описанный ранее в [7]. Алгоритм прогноза должен обеспечивать высокую точность, поскольку даже небольшие ошибки в значениях поля повлекут за собой ошибки в аппроксимации коэффициентов разложения, что может в итоге привести к большим искажениям результирующего поля. Поэтому для построения прогностических уравнений нами была предложена новая модификация давно известного метода «аналогов». Суть ее заключается в следующем. Временной ряд некоторой переменной $f(x_k)$ преобразовывается в таблицу $\mathbf{A}$ размерностью $d \times m - d + 1$, где d определяется числом запаздывающих аргументов в прогностическом уравнении (*Lag*) и заблаговременностью прогноза (*Step*), m – длина выборки. *Lag* определяет длину паттерна, к которому подбираются наиболее близкие по определенному критерию аналоги из предыстории процесса. В качестве критерия предлагается использовать среднее абсолютных величин разностей производных по времени искомого паттерна и строк-аналогов

$$q = \frac{1}{d - Step - 1} \sum_{k=2}^{d-Step} \left| \Delta p_k - \Delta a_k \right| \qquad (3)$$

где p_k и a_k – значения переменной в строках искомого паттерна и оцениваемого аналога соответственно.

Каждой строке матрицы $\mathbf{A}$ ставится в соответствие некоторое значение критерия q, после чего таблица ранжируется по этому значению. Для построения уравнения регрессии не имеет смысла использовать всю получившуюся таблицу, поскольку, как правило, только первые ее строки действительно близки по выбранной мере к искомому паттерну. Поэтому следует применять метод, позволяющий на короткой выборке получать адекватное регрессионное уравнение. Нами применялся метод построения полиномиальной нейронной сети, предложенный в [10]. Этот алгоритм основан на известном Методе группового учета аргументов, который хорошо себя зарекомендовал при работе с короткими выборками. Величина d, а также объем выборки, используемой при построении модели, являются эмпирическими параметрами и существен-

но влияют на качество получаемого прогноза. Эти параметры необходимо тщательно подбирать для каждой конкретной задачи.

Опыт показал, что на результативность прогноза большое влияние оказывает правильный подбор аналогов, т.е. строк для формирования тренировочной и контрольной последовательностей в алгоритме построения нейронной сети. Для уточнения подбора предлагается помимо критерия (3) учитывать знак приращения в строках-аналогах $a_d - a_{d\text{-}Step}$, аналогичного прогнозируемому приращению $p_d - p_{d\text{-}Step}$. Практически, для каждой строки-аналога определялся знак указанного приращения, после чего производился подсчет числа положительных и отрицательных знаков в верхней части таблицы. В таблице оставлялись только строки, знак приращения которых совпадал со знаком большей части строк. Эта операция существенно уменьшала результирующую ошибку прогноза, если только знак приращения предсказывался верно. В численных экспериментах с реализациями случайных переменных близких к белому шуму, таких как биржевые ряды, было установлено, что знак приращения в среднем определяется верно примерно в 65-70% случаев. Для временных рядов аномалий температуры, например, этот процент выше – 80-85%.

С целью тестирования метода были выполнены эксперименты по временной экстраполяции поля среднемесячной температуры поверхности земного шара. Использовались данные реанализа за 50 лет (проект NCEP/NCAR) на сетке 144 × 73. Количество точек в реализации поля, таким образом, равнялось 10512, число реализаций – 612. Прогнозировались значения поля температуры поверхности земного шара в случайно выбранных точках, после чего поле восстанавливалось с помощью системы ЭОФ, рассчитанных ранее SVD методом. Коэффициенты разложения определялись с помощью генетического алгоритма и методом МНК. В разложении поля использовались первые 10 ЭОФ, описывающих 97% изменчивости. Точность прогноза температуры при заблаговременности от одного до пяти месяцев составляла в среднем 0,32-0,34°С при стандартном отклонении реализаций 6,13°С, т.е. относительная ошибка прогноза равнялась примерно 5%. Наилучшие уравнения линейной регрессии, для сравнения, давали ошибку от 8 до 21% в зависимости от заблаговременности прогноза. Качество прогноза на 5 шагов вперед иллюстрирует рис.1.

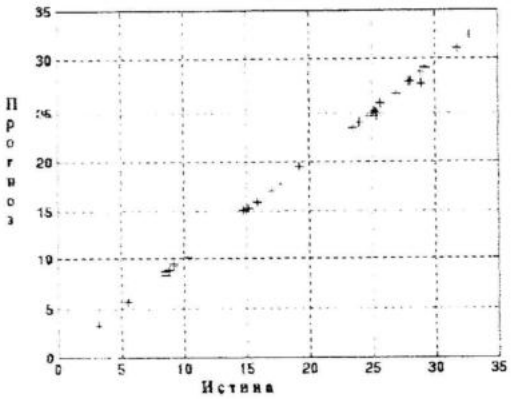
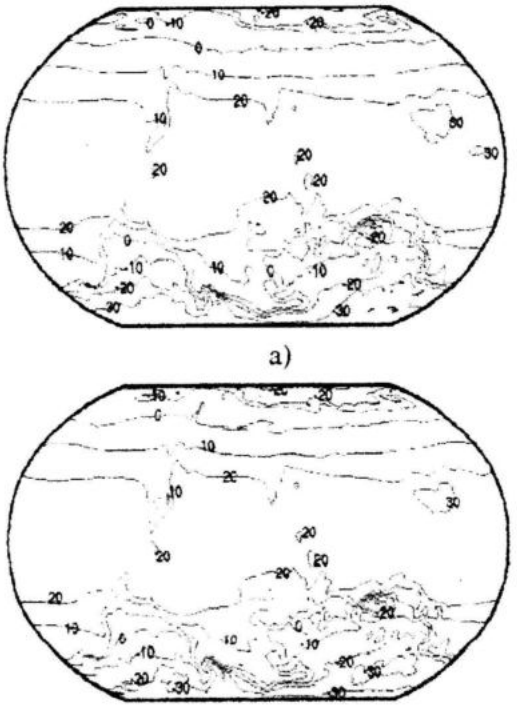


Рис. 1 - Результаты прогноза температуры поверхности на 5 шагов вперед в 30 случайно выбранных узлах сетки.



а)

б)

Рис. 2 - Результат прогноза поля поверхностной температуры по 30 узлам сетки: а) истинное поле, б) прогностическое.

Результирующее поле температуры, реконструированное по 30 точкам (из 10512 узлов сетки) в сравнении с истинным полем, показано на рис. 2. Здесь под «истинным» понимается поле, построенное по 10 собственным функциям. Средняя относительная среднеквадратичная ошибка реконструкции поля по 30 точкам составляет

19%, по 100 точкам – 13%. При восстановлении поля по 30 точкам использовался генетический алгоритм, подробное описание его приведено в [6]. В процессе генерации все новых поколений популяции потенциальных решений задачи, постепенно улучшается их функция качества *F*. Цель работы алгоритма – минимизировать функционал, задающий критерий качества. На рис. 3 показана динамика минимального значения *F(n)*, достигаемого в последовательных поколениях популяции.

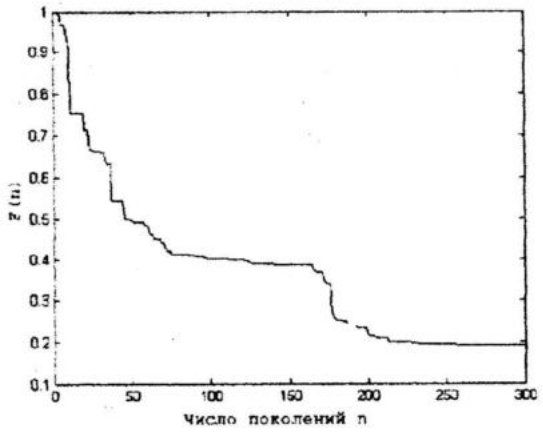


Рис. 3 – Поведение функции ошибки в процессе генерации последовательных поколений популяции потенциальных решений задачи.

Выводы. Тестирование показало, что предлагаемый метод является эффективным инструментом для восстановления пропусков в больших массивах натурных данных, а также для временного прогноза полей. Кроме того, преобразование полей с помощью ЭОФ для последующего хранения и использования позволяет естественным образом производить фильтрацию шума и существенно экономить память. Разработанный метод может практически применяться для оперативной обработки данных, а именно, восстановления полей различных характеристик по фрагментарным изображениям, получаемым со спутника. Также его применение возможно для интерполяции и экстраполяции данных дрифтеров. Для получения оценки ковариационной матрицы возможно использовать различные архивные данные, в том числе данные реанализа.

ЛИТЕРАТУРА

1. Lorenz E. Empirical orthogonal functions and statistical weather prediction, Tech. Rep. 1, Statistical Forecasting Project, 1956, Department of Meteorology, Massachusetts Institute of Technology, Cambridge, Massachusetts.
2. Багров Н.А. Естественные составляющие малых выборок при большом числе параметров. Метеорология и гидрология, 1978, № 12, С. 5-14.
3. Kelly K.A. The influence of Winds and topography on the sea surface temperature patterns over northern California slope. J of Geophys. Res., 1985, Vol. 98, № C6, pp 11783-11798.
4. Пичугин Ю.А., Покровский О.М. Анализ пространственно-временной структуры поля H_{500} в Северном полушарии. Метеорология и гидрология, 1990, № 3, сс.38-46.
5. Everson R., Cornillon P., Sirovich L. & Webber A. An empirical eigenfunction analysis of sea surface temperatures in the Western North Atlantic, J. of Phys.Oceanogr., 1997, Vol. 27, № 3, pp. 468-479.
6. Васечкина Е.Ф., Ярин В.Д. Использование генетического алгоритма в задаче восстановления пропущенных данных. Морской гидрофизический журнал, 2002, № 4, С. 30-39.
7. Васечкина Е.Ф., Ярин В.Д. Генетический алгоритм в задаче реконструкции гидрометеорологических полей. Материалы Международного научно-технического семинара «Системы контроля окружающей среды - 2001», Севастополь, 2002, С. 141-145.
8. Тимченко И.Е., Ярин В.Д., Бамба З. Прогноз температуры поверхности океана на основе спутниковых измерений. НИЦ Конакри-Рогбане, Конакри, 1991, 42с.
9. Репинская Р.П., Бабич Я.Б. Аппроксимация рядами эмпирических ортогональных функций северополушарных полей облачности по спутниковым данным. Исследование Земли из космоса, 1999, № 6, сс. 8-15.
10. Васечкина Е.Ф., Ярин В.Д. Использование эволюционных методов в задаче моделирования экосистем. Морской гидрофизический журнал, 2001, № 3, С. 65-74.