

ПОИСК СХОДСТВА В МЕТЕОРОЛОГИЧЕСКИХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ МЕТОДА СИНГУЛЯРНОГО СПЕКТРАЛЬНОГО АНАЛИЗА

Л. И. Леви, д-р техн. наук;

А. В. Чичикалов, аспирант,

Луганский национальный аграрный университет, г. Луганск

В работе представлен алгоритм выбора ближайшего по структурным признакам временного ряда на основе сингулярного спектрального анализа.

Ключевые слова: *обработка данных, временной ряд, сингулярный спектральный анализ.*

В роботі подається алгоритм відбору найближчого за структурними ознаками часового ряду на основі сингулярного спектрального аналізу.

Ключові слова: *обробка даних, часовий ряд, сингулярний спектральний аналіз.*

ВВЕДЕНИЕ

Интенсивное внедрение современных технологий в сельскохозяйственную отрасль обеспечивается за счет развития компьютерной техники, программного обеспечения, систем управления. Чтобы контролировать сложный процесс или объект, информационно-управляющая система должна вести мониторинг основных параметров и обрабатывать собранную информацию. Рассматривая процесс орошения и связанные с ним задачи планирования водораспределения, следует отметить наличие больших объемов агрометеорологической информации. На основании этих данных строятся прогнозы и принимаются управляющие решения.

Извлечение полезной информации из массива данных является актуальной задачей во многих сферах деятельности. В орошаемом земледелии это непосредственно связано с рациональным использованием водных ресурсов и увеличением урожайности. Например, анализ предыдущего опыта существенно упрощает решение многих задач. Чем ближе по начальным условиям подобран аналог решаемой задачи, тем точнее найденное решение. Такой подход чаще применяют там, где наблюдаются сезонные (периодические) изменения входных величин.

Так как при планировании поливных работ учитывается множество климатических факторов, то общее решение требует проведения достаточно сложных расчетов и моделирования. Однако, чтобы получить начальные сведения и возможное направление для решения исходной задачи, достаточно использовать базу данных и алгоритмы поиска ближайшего по исходным (метеорологическим) условиям года-аналога. При этом результаты прошлого опыта можно расценивать как первое приближение в решении текущей задачи. В основе такого анализа лежат классификация и сравнение временных рядов.

Одним из методов исследования временного ряда является сингулярный спектральный анализ (ССА). Этот вид обработки данных обладает широкими возможностями по структурному анализу, прогнозированию и сглаживанию временных рядов. Фактически, ССА дает разложение исходной временной последовательности в виде набора составляющих. При последующем изучении свойств и характеристик компонент разложения определяются наиболее существенные и важные составляющие, такие как тренд или шум. Эти свойства ССА нашли свое применение в задачах фильтрации в качестве инструмента для

исключения грубых ошибок в измерениях, а также изоляции резких выбросов сигнала.

АНАЛИЗ ЛИТЕРАТУРЫ

Авторами работ [1, 2] выполнен практический анализ и представлены результаты моделирования параметров электрических сетей по методу ССА. За последнее время появилось несколько модификаций метода [3], адаптированных под конкретные нужды исследователей. Некоторые прикладные аспекты обработки временных рядов освещены в статье [4], где рассмотрен робастный подход к исследованию экологических данных. Теоретические основы метода изложены в [5].

ПОСТАНОВКА ПРОБЛЕМЫ

Общий подход при сравнении временных рядов, как правило, не может дать удовлетворительных результатов, так как при этом не учитываются структурные составляющие ряда и влияние различных компонент на степень сходства. Цель данной работы - построение алгоритма сравнения временных рядов по структурным признакам для выбора года-аналога из массива метеорологических данных с применением метода сингулярного спектрального анализа.

ОСНОВНАЯ ЧАСТЬ

Пусть имеется ряд наблюдений, зафиксированных через равные интервалы времени:

$$Y(t) = \{y_1, y_2 \dots y_N\}. \quad (1)$$

Согласно базового алгоритма ССА, одномерный ряд разворачивается в многомерный путем составления траекторной матрицы. Для формирования матрицы задается размер скользящего окна L и выполняется операция циклического сдвига окна на один шаг вдоль исходного временного ряда (1). Полученная выборка векторов сохраняется, как траекторная матрица (2) с размером $L \times (N - L + 1)$. Если диапазон значений ряда имеет большой разброс, то матрица дополнительно центрируется и нормируется по столбцам.

$$A = \begin{bmatrix} y_1 & y_2 & y_3 & \dots & y_{(N-L+1)} \\ y_2 & y_3 & y_4 & \dots & y_{(N-L+2)} \\ y_3 & y_4 & y_5 & \dots & y_{(N-L+3)} \\ \dots & \dots & \dots & \dots & \dots \\ y_L & y_{L+1} & y_{L+2} & \dots & y_N \end{bmatrix} \quad (2)$$

Матричная форма записи временного ряда дает новое геометрическое представление данных в виде набора из L точек в K -мерном пространстве, где $K = N - L + 1$. На следующем шаге вычисляется матрица связи R между векторами:

$$R = A \cdot A^T = \begin{bmatrix} r_{11} & \dots & r_{1L} \\ \dots & \dots & \dots \\ r_{L1} & \dots & r_{LL} \end{bmatrix} \quad (3)$$

Дальнейшие преобразования связаны с сингулярным разложением матрицы (3). По стандартным правилам находим собственные числа и собственные вектора матрицы R :

$$R = U\Lambda V^T, \quad (4)$$

где столбцы матриц U и V - собственные вектора RR^T и R^TR соответственно, а $\Lambda = \text{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_L})$ - диагональная матрица сингулярных чисел. Величины собственных чисел характеризуют долю объясняемой дисперсии θ (5) и являются основным критерием при выборе главных компонент.

$$\theta = \frac{\lambda_i}{\sum_i \lambda_i} \cdot 100. \quad (5)$$

Ортогональную матрицу собственных векторов можно представить как набор переходных функций фильтров, настроенных на выделение соответствующей компоненты из временного ряда. После прохождения исходной последовательности через такой блок фильтров ряд распадется на аналогичное количество спектральных составляющих. Чтобы определить выход каждого фильтра, выполняется дискретная операция свертки сигнала с коэффициентами импульсной характеристики. Учитывая то, что траекторная матрица имеет ганкелеву форму, результаты фильтрации выразим через произведение матриц:

$$S = R' \cdot P, \quad (6)$$

где S - матрица разложения, столбцы которой содержат компоненты исходного ряда; P - матрица собственных векторов.

Необходимо отметить, что фильтры настраиваются самим процессом и не требуют дополнительных знаний о структуре временного ряда, что является несомненным преимуществом при исследовании стохастических процессов.

По матрице S выполняются оценка и сравнение структурных компонентов ряда по выбранной метрике. Варьируя количеством анализируемых компонент, оценивается степень сходства рядов по интересующим структурным составляющим.

Для анализа использовались среднесуточные значения температуры воздуха за 4 года (рис. 1). Показания температуры за 2007 год служили образцом, по которому выполнялся поиск ближайшего ряда-аналога.

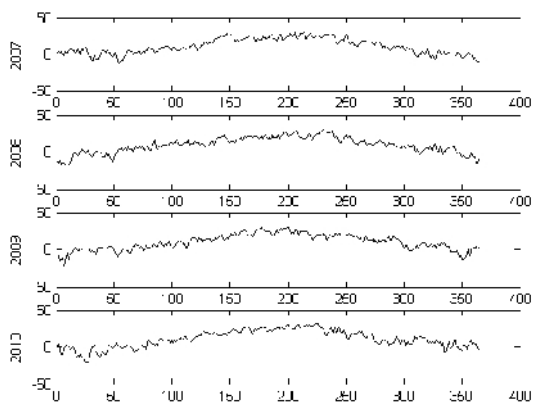


Рисунок 1 - Значения суточных температур за 2007-2010 гг.

Так как наибольшей информативностью обладают именно первые собственные числа, то соответствующие им фильтры формируют “скелет” временного ряда. Сравнительная характеристика рядов для набора из 25 первых компонент показана на рис. 2.

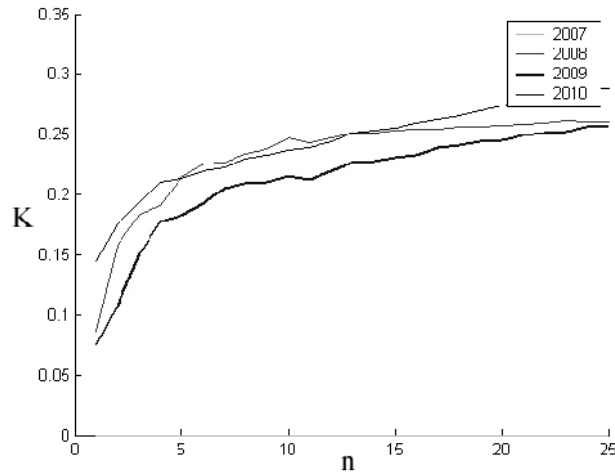


Рисунок 2 - Зависимость коэффициента несоответствия от числа используемых компонент (n)

В качестве меры близости рядов использовался коэффициент несоответствия Тейла, определяемый по формуле (7)

$$K = \sqrt{\frac{\sum_i (Yp_i - Yt_i)^2}{\sum_i Yp_i^2 + \sum_i Yt_i^2}} \quad (7)$$

где Yt - эталонный ряд; Yp - подбираемый аналог.

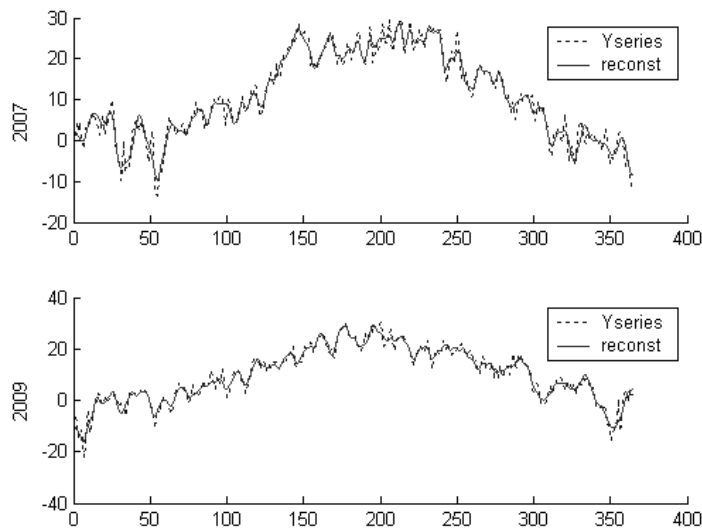


Рисунок 3 - Исходный год и ближайший год-аналог с восстановленными значениями по 25 компонентам

Алгоритм сравнения состоит из следующих этапов:

1. Формирование траекторных матриц и матриц связи для исследуемых временных рядов.
2. Сингулярное разложение матрицы связи. Вычисление собственных чисел и собственных векторов.
3. Выбор числа анализируемых компонент по значениям собственных чисел.
4. Синтез ряда по выбранным компонентам.
5. Оценка подобия рядов по критерию (7).

Из рис. 2 видно, что добавление компонент по-разному влияет на показатель сходства рядов. Поэтому сравнение временных рядов с использованием разложения позволяет одновременно оценивать влияние каждой составляющей на степень близости рядов. Поиск соответствия можно вести как по средним значениям или тренду, так и по быстро меняющимся или шумовым компонентам ряда.

ВЫВОДЫ

Простое применение норм (например, евклидовой) для сравнения временных рядов не всегда дает адекватный результат. Рассмотренный в статье алгоритм обходит это ограничение за счет разложения ряда на составляющие по методу ССА с последующим анализом и синтезом по выбранным компонентам. Появляется возможность целенаправленного подбора рядов по определенным признакам, что делает алгоритм более устойчивым к шумам и избирательным по отношению к структуре ряда.

SUMMARY

SIMILARITY SEARCH IN METEOROLOGICAL DATA USING SINGULAR SPECTRUM ANALYSIS

*L. I. Levi, A. V. Chyhykalov,
Lugansk National Agrarian University*

In this work the algorithm for selection of similar time series by structure sign is considered. It is based on Singular Spectrum Analysis.

Key words: data mining, time series, singular spectrum analysis

СПИСОК ЛИТЕРАТУРЫ

1. Мартко Е. О. Сингулярный спектральный анализ как метод моделирования электрической нагрузки / Е. О. Мартко, И. В. Белицын // Ползуновский вестник. – 2009. – № 4. – С. 76–85.
2. Гнатюк В. И. Прогнозирование электропотребления методом анализа главных компонент / В. И. Гнатюк, Д. В. Луценко, Л. И. Двойрис, Д. В. Антоенков, П. Ю. Дюндик // Электрика. – 2007. – № 3. – С. 41–46.
3. Любчик Л. М. Диагностика структурных изменений временных рядов на основе метода ядерного сингулярно-спектрального анализа / Л. М. Любчик, А. А. Мирошниченко // Восточно-Европейский журнал передовых технологий. – 2008. – № 4/5 (34). – С. 17–20.
4. Jackson Donald A. Robust principal component analysis and outlier detection with ecological data / Donald A. Jackson, Yong Chen // *Envirometrics*. – 2004. – № 15. – P. 129–139.
5. Голяндина Н. Э. Метод «Гусеница-SSA»: анализ временных рядов: учеб. пособие / Н. Э. Голяндина. – СПб. : ВВМ, 2004. – 76 с.

Поступила в редакцию 28 октября 2011 г.