УДК 004.2

УНИФИКАЦИЯ ВРЕМЕННЫХ РЯДОВ ГИДРОМЕТЕОРОЛОГИЧЕСКИХ ДАННЫХ С ПОМОЩЬЮ ФИЛЬТРА КАЛМАНА

А. А. Шульган
Брестский государственный технический университет, Брест;
Научный руководитель: Костюк Д. А., к.т.н., доцент

По мере появления все большего числа автоматических источников гидрологических и гидрометеорологических данных на речных системах (средств измерения уровня воды, количества осадков, температуры воздуха и др.), для решения задач, связанных с моделированием гидрологических явлений и прогнозированием их динамики, становится доступно больше временных рядов, в результате чего должны повышаться точность и оперативность как расчета текущей ситуации, так и прогноза. Однако при использовании большого числа потоков данных возникают дополнительные проблемы.

На практике оказывается, что одни временные ряды отбираются с более высокой частотой, чем другие. Иногда данные поступают практически в режиме реального времени, в плоть до разрешения в секундах, и характерной проблемой таких высокочастотных данных является шум. Кроме того, сложность параллельного использования нескольких показателей заключается в том, что они могут давать противоречивые сигналы, и не существует согласованного способа агрегирования статистических данных для получения однозначного ответа.

Необходимость построения многомерной модели возникает в двух основных ситуациях:

1. высокочастотные и низкочастотные данные могут отбираться для нескольких переменных, и в результате требуется получить оценки для ненаблюдаемых высокочастотных значений, используя всю доступную (агрегированную
и дезагрегированную) информацию;
2. важный информативный параметр системы оценивается с низкой частотой, но формирующие его показатели отбираются на более высокой частоте. Чтобы оценить эволюцию целевой переменной, нужно: (а) рассчитать высокочастотный показатель ее колебаний и (б) спрогнозировать ее очередное значение в низкочастотном временном ряду, используя в обоих случаях информацию, получаемую на основе высокочастотных значений.

Одна из ситуаций, в которых возникает такая задача, известна как «дезагрегация дождевых осадков» (rainfall disaggregation), когда высокочастотные данные об осадках требуется получать на основе низкочастотных данных [1].

В основе линейных моделей, предполагающих представление в виде пространства состояний, может лежать использование максимального гауссовского подобия в форме разложения ошибки прогнозирования с использованием фильтра Калмана для вычисления требуемых условных моментов. В [2] представлены

аналитические выражения для функции подобия, ее производных и соответствующей матрицы данных. Уравнение наблюдений в данной модели легко модифицировать для учета данных, наблюдаемых с ограничениями агрегации [2, 3].

Фильтр Калмана является разновидностью рекурсивных фильтров. Для вычисления оценки состояния системы на текущий такт работы ему необходима оценка состояния (в виде оценки состояния системы и оценки погрешности определения этого состояния) на предыдущем такте работы и измерения на текущем такте. Данное свойство отличает его от пакетных фильтров, требующих в текущий такт работы знание истории измерений и/или оценок. Далее под записью $\hat{x}_{n|m}$ будем понимать оценку истинного вектора x в момент n с учетом измерений с момента начала работы и по момент m включительно.

Состояние фильтра задается двумя переменными:

- $\hat{x}_{k|k}$ – апостериорная оценка состояния объекта в момент k, полученная по результатам наблюдений вплоть до момента k включительно;
- $P_{k|k}$ – апостериорная ковариационная матрица ошибок, задающая оценку точности полученной оценки вектора состояния и включающая в себя оценку дисперсий погрешности вычисленного состояния и ковариации, показывающие выявленные взаимосвязи между параметрами состояния системы.

Каждая итерация фильтра Калмана делится на две фазы: экстраполяция (прогноз) и коррекция. Во время экстраполяции фильтр получает предварительную оценку состояния системы $\hat{x}_{k|k-1}$ на текущий шаг, на основе итоговой оценки состояния с предыдущего шага (либо предварительную оценку на следующий шаг на основе итоговой оценки, текущего шага, в зависимости от интерпретации). Эту предварительную оценку также называют априорной оценкой состояния, так как для её получения не используются наблюдения соответствующего шага. В фазе коррекции априорная экстраполяция дополняется соответствующими текущими измерениями для коррекции оценки. Скорректированная оценка также называется апостериорной оценкой состояния, либо просто оценкой вектора состояния $\hat{x}_k$. Обычно эти две фазы чередуются: экстраполяция производится по результатам коррекции до следующего наблюдения, а коррекция производится совместно с доступными на следующем шаге наблюдениями, и т. д. Однако если по некоторой причине наблюдение оказалось недоступным, то этап коррекции может быть пропущен и выполнена экстраполяция на основе нескорректированной оценки (априорной экстраполяции). Аналогично, если независимые измерения доступны только в отдельные такты работы, всё равно возможны коррекции (обычно с использованием другой матрицы наблюдений H_k).

При построении высокочастотной модели обычно требуется аппроксимировать ненаблюдаемые значения на основе информации, имеющейся в выборке. Эти значения интерполируются или экстраполируются в зависимости от того, находятся ли они внутри или вне выборки. В рамках модели пространства состояний эту задачу может выполнять алгоритм сглаживания с фиксированным интервалом [4].

Предположительно между целевой переменной и отдельными показателями существует высокочастотная связь:

$$y_t = x_t^T \beta + \varepsilon_t;\ \varepsilon_t = \frac{1}{(1-\varphi_1 B)(1-\varphi_2 B)} a_t \quad (1)$$

где y_t обозначает целевую переменную, x_t - вектор показателей, $a_t \sim iid(0, \sigma^2_a)$, а B обозначает оператор обратной связи, такой что для любой последовательности w_t: $B^k w_t = w_{t-k}$, $k = 0, \pm 1, \pm 2, \ldots$

При этом предполагается статическая линейная зависимость между показателем (причина) и целевой переменной (эффект), разные порядки интегрирования для переменных (в некоторых случаях подразумевается интеграция между показателем
и целевой переменной), а также модель не учитывает сезонные факторы, поэтому сезонность должна быть либо убрана заранее, либо являться общей особенностью y_t и x_t, так, чтобы линейная комбинация $y_t - x_t^T \beta$ не имела сезонной структуры.

Для проверки эффективности унификации временных рядов гидрометеорологических данных с помощью фильтра Калмана была использована Python-библиотека «pykalman», которая содержит реализацию KF в виде сглаживающего фильтра.

И KF, и его сглаживающая реализация, традиционно используются с уже заданными параметрами. В случае библиотеки pykalman, класс KalmanFilter может быть инициализирован любым подмножеством обычных параметров модели и использоваться без подгонки. Для всех неопределенных параметров задаются значения по умолчанию.

Реализация со сглаживанием может включать «будущие» измерения, а также прошлые при одинаковых вычислительных затратах $O(Td^3)$, где T – количество временных шагов, а d – размерность пространства состояний.

В дополнение класс KalmanFilter реализует алгоритм максимизации ожидания (EM). Этот итерационный алгоритм – способ максимизировать вероятность наблюдаемых измерений.

В реальных гидрометеорологических измерениях случается временный выход из строя одного из датчиков – например, выход из строя микроволнового сканирующего радиометра-поляриметра SSM/I в 2014 году, использовавшегося для определения количества осадков, накопленных в снежном покрове, на основе спутниковых измерений радиояркостных температур [5]. Интерпретация пространства состояний с применением KF является популярным подходом к восстановлению отсутствующих высокочастотных наблюдений.

Если обозначить через t единицу времени временного ряда с более низкой частотой, $t=1, \ldots, T$, и использовать для связи частот понятие частотной смеси m (высокочастотная переменная наблюдается m раз в интервале от t до $t-1$), то единица времени для высокочастотных данных τ может быть записана как $\tau=1, \ldots mT$. Наличие двух частот в одной модели временного ряда представлено записью $x_{t-i/m}$.

Пусть Y_τ - вектор размерности $N \times 1$ стационарного временного ряда с наблюдениями для моментов $\tau = 1, \ldots, mT$. Предполагается, что ряд имеет нулевые средние значения. Переменные в факторной модели представляются в виде суммы двух взаимно-ортогональных компонент, общего и специфического. Общий компонент формирует небольшое количество факторов, общих для всех переменных модели. Специфический компонент формируется воздействиями, специфичными для различных видов переменных. Факторную модель можно описать следующим образом:

$$Y_\tau = \Lambda F_\tau + \epsilon_\tau \tag{2}$$

где F_τ имеет размерность $(r \times 1)$ и представляет собой вектор факторов, а матрица Λ размерности $(N \times r)$ содержит факторные нагрузки. Специфический компонент представлен вектором ξ_τ.

Небольшое количество факторов может объяснить большую часть дисперсии данных. Факторы могут быть оценены с использованием подхода главных компонентов.

Агрегация по времени высокочастотных переменных в низкочастотные встречается в прикладных задачах достаточно часто, хотя и приводит к потере информации.

Простейшим случаем является агрегация в рамках одиночного временного ряда. Собственно «агрегация» может выполняться по-разному в зависимости от практического значения переменных. На роль представителя последовательности высокочастотных значений может выбираться первое, второе, либо последнее из них. Подобное «объединение» характерно для ситуаций, когда расхождение частот не очень велико, а выбор делается из прагматических соображений, без существенного теоретического обоснования.

Стандартным способом агрегации безусловно является усреднение по периоду:

$$x_t = \frac{1}{m} \sum_{i=1}^{m} x_{t-i/m} \tag{3}$$

Наконец, для величин с накоплением, значения просто складываются:

$$x_t = \sum_{i=1}^{m} x_{t-i/m} \tag{4}$$

В случае несколько временных рядов, при наличии большого количества переменных возникает рост неопределенности, который может приводить к снижению точности прогноза. Одним из способов преодоления этой проблемы является сжатие всех доступных временных рядов в меньшее число переменных. Построение таких композитных индикаторов может служить целям прогнозирования или оценки текущего состояния системы.

Построение композитных индикаторов предполагает, что временные ряды имеют одинаковую частоту. В случае данных смешанной частотности, аналогичного результата можно добиться опять же приведением их к форме пространства состояний и применением фильтра Калмана, средствами простой аппроксимации

или, в более сложном варианте, использованием нелинейной модели и, соответственно, нелинейной фильтрации [6].

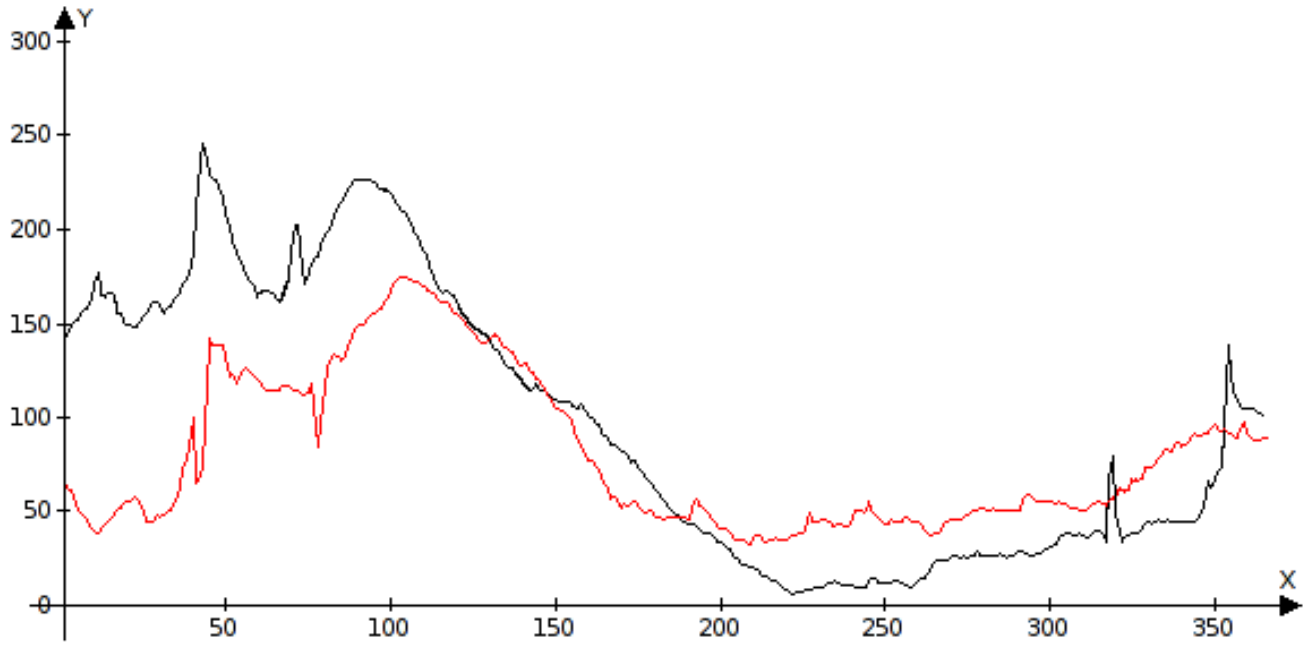

Рис. 1 – Изменение расхода воды в контрольной точке речного русла

Отметим, применение KF является для медленно изменяющихся гидрологических показателей, расход воды в речном русле (рис. 1), и наименее эффективным — для сложно прогнозируемых высокочастотных временных рядов, связанных с измерениями в реальном времени, включающими в себя суммированный статистический шум локального происхождения.

Список цитированных источников

1. Onof C. et al. Spatial-temporal rainfall fields: modelling and statistical aspects // Hydrology and Earth System Sciences, v. 4, iss. 4, 2000. – p. 581-601.

2. Terceiro J. Estimation of Dynamic Econometric Models with Errors in Variables – Springer-Verlag, Berlin, 1990 – 132 p.

3. Ansley C.F., Kohn R. Exact likelihood of vector autoregressive-moving average process with missing or aggregated data – Biometrika, V. 70, No. 1, 1983 – p. 275–278

4. Anderson B.D.O., Moore J.B. Optimal Filtering. – Prentice-Hall, 1979. – 357 p.

5. Волчек А.А., Костюк Д.А.,Петров Д.О., Шешко Н.Н. Метод прогнозирования половодий на основе многофакторного нейросетевого анализа // Вестник БрГТУ. – 2018.– №5(113): Физика, математика, информатика. – С. 74–76.

6. Proietti T., Moauro F. Dynamic factor analysis with non-linear temporal aggregation constraints // Journal of the Royal Statistical Society Series C, No. 55(2), 2006 - P. 281–300.