Главная » Оптика атмосферы и океана » 2019 (25) » Труды

XXV Международный Симпозиум “Оптика атмосферы и океана. Физика атмосферы”

30 июня - 5 июля 2019 года, Новосибирск

≡

NOVOSIBIRSK 2019
ATMOSPHERIC and OCEAN OPTICS. ATMOSPHERIC PHYSICS

Труды

Оптика атмосферы и океана. Физика атмосферы: Материалы XXV Международного симпозиума.

Электронный ресурс

Томск: Издательство ИОА СО РАН.

ISBN 978-5-94458-176-1

Содержание

2002 (9)
2001 (8)
2000 (7)
1999 (6)

МОЛЕКУЛЯРНАЯ СПЕКТРОСКОПИЯ ВЫСОКОГО РАЗРЕШЕНИЯ

ИМПУЛЬСНЫЕ ЛАЗЕРЫ НА ПЕРЕХОДАХ АТОМОВ И МОЛЕКУЛ

РАБОЧАЯ ГРУППА "АЭРОЗОЛИ СИБИРИ"

РАСПРОСТРАНЕНИЕ РАДИОВОЛН

КОНФЕРЕНЦИЯ МОЛОДЫХ УЧЕНЫХ "МАТЕРИАЛОВЕДЕНИЕ, ТЕХНОЛОГИИ И ЭКОЛОГИЯ"

XXV МЕЖДУНАРОДНАЯ

КОНФЕРЕНЦИЯ D

ФИЗИКА ТРОПОСФЕРЫ

ОБНАРУЖЕНИЕ И ВЫДЕЛЕНИЕ АНОМАЛЬНЫХ НАБЛЮДЕНИЙ ПРИ АКУСТИЧЕСКОМ ЗОНДИРОВАНИИ СКОРОСТИ ВЕТРА

Симахин В.А.[1], Черепанов О.С. [1], Шаманаева Л.Г.[2,3]

[1]Курганский государственный университет

[2]Институт оптики атмосферы им. В.Е. Зуева СО РАН;

[3]Национальный исследовательский Томский государственный университет

E-mail: sva_full@mail.ru; sima@iao.ru

Ключевые слова: акустическое зондирование, скорость ветра, робастный непараметрический алгоритм маятникового усечения.

Статистический анализ мини-содарных измерений высотных профилей трех компонентов скорости ветра в слое 5–200 м показывает, что данная задача относится к классу робастных непараметрических задач математической статистики. В докладе для обработки данных мини-содарных измерений скорости ветра предложен новый последовательный непараметрический алгоритм маятникового усечения для обнаружения и выделения аномальных наблюдений в выборке. С использованием данного алгоритма найдены выборочные моменты высотных профилей скорости ветра и проведено их сравнение с классическими выборочными моментами.

1. Введение

Мини-содары позволяют исследовать тонкую структуру атмосферного пограничного слоя (АПС). Однако обработка реальных результатов мини-содарных измерений скорости ветра в АПС [1] выявила ряд вопросов, связанных с проблемой *big data*. Большой объём результатов измерений, наличие разнообразных выбросов, трудность подбора параметрических моделей (непараметричность задачи) исключают *ручную подгонку* под известные параметрические модели и требуют применения робастных непараметрических методов статистики [1, 2]. Экспериментаторы давно знакомы с проблемой появления в опытных данных аномальных наблюдений (выбросов), которые могут существенно исказить результат. Типичный приём робастной статистики, который применяется при обработке данных в этой ситуации – обнаружить и удалить выбросы, причём в условиях априорной неопределенности об их количестве и местонахождения. Проблема обнаружения и выделения выбросов уже давно привлекает внимание исследователей и актуальна как с теоретической, так и с практической точки зрения [3–5]. Предложен ряд параметрических критериев, из которых выделим критерии Граббса [4] и их обобщения [3]. Дальнейшие исследования показали, что данные критерии являются неустойчивыми при отклонениях от нормального распределения [5], что и вызывает определенную долю скепсиса при их использовании. В данном докладе на основе эмпирической функции влияния и чувствительности [2] предлагается итеративная непараметрическая процедура, позволяющая ранжировать выборочные значения претендентов на выброс. Для формального обоснования данной процедуры требуется предположение о непрерывности и стационарности второго порядка функции чувствительности [2]. Фактически алгоритм производит маятниковое усечение выборочных значений на основе упорядочивания эмпирических функций влияния. На основе данного алгоритма удобно строить адаптивные робастные оценки, основанные на операциях усечения выборки, минуя анализ симметричности и затянутости хвостов распределения [2].

2. Процедура обнаружения и выделения выбросов

2.1 Алгоритм маятникового усечения. Пусть $\vec{x}_N = \{x_1,...x_N\}$ – выборка н.о.р. из неизвестного распределения $F(x) = (1-\varepsilon)G(x) + \varepsilon H(x)$ – модель выбросов Тьюки, где $G(x)$ – основное априорное распределение, $H(x)$ – распределение выбросов, ε – доля выбросов, $k = [N \cdot \varepsilon]$ – количество выбросов в выборке. Пусть $F(x), G(x), H(x)$ – абсолютно непрерывные унимодальные распределения с плотностями $f(x), g(x), h(x)$, соответственно. Стандартная задача обнаружения и выделения k удаленных от центра распределения $F(x)$ выбросов сводится к задаче проверки гипотез:

$$H_0 : k = 0\ (F = G),$$
$$H_1 : k \neq 0\ (F = (1-\varepsilon)G + \varepsilon H).$$

В качестве меры аномальности, возьмем модуль отклонения наблюдения от среднего значения $T = \int |x - EX| dF(x)$. Введем переменный объём выборки $\vec{x}_n = \{x_1,...x_n\}$, $n = N, N-1,...,[N/2]$. В соответствии с мерой аномальности, рассмотрим преобразование выборочных наблюдений к виду

$$T_i(x_i) = (x_i - \bar{T}_n(\vec{x}_n)),\ \bar{T}_n(\vec{x}_n) = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad (1)$$

$$t_i(n) = |T_i(x_i)|. \qquad (2)$$

Упорядочим величины $t_i(n) = |T_i(x_i)|$, $t_{(1)}(n) < t_{(2)}(n) < ... < t_{(n)}(n)$. Рассмотрим последовательную процедуру определения претендентов на выброс. Выбросы по мере аномальности T представляют крайние порядковые статистики $t_{(N)}(n),...,t_{(N-k+1)}(n)$. Наблюдение x_{i_0} соответствующее $t_{(n)}(n)$ $(x_{i_0} = \arg\max |T_i(x_i)|)$ является претендентом на выброс, поэтому удаляем его из выборки $\vec{x}_n = \{x_1,...x_n\}$. В результате получаем выборку $\vec{x}_{n-1}$ объёма $(n-1)$. Данную процедуру выявления претендентов на выброс повторяем для $n = N, N-1,...,[N/2]$. Выборочные наблюдения, удаленные таким образом, являются лишь претендентами на выброс, и поэтому для определения, какие из претендентов являются выбросами, необходима дополнительная процедура вынесения решения. Введем статистику

$$L_n = \frac{S_n}{S_N}, \tag{3}$$

где $S_n = \sum_{i=1}^{n} (T_i(x_i))^2$, $n = N, N-1, \ldots, [{}^{N}\!/_{2}]$. (4)

Так как $S_n = S_{n-1} + (t_{(n)}(n))^2$, $S_N = const(N)$, то $S_{n-1} < S_n$ и, следовательно, статистика $0 < L_n \le 1$ является монотонно убывающей функцией от *n*. Найдем среднее значение ES_N, ES_n, $E(t_{(n)}(n))^2$, $EL_n = \frac{ES_n}{ES_N} + 0(N^{-1})$:

$$E\frac{1}{N}S_N = \int (t - ET_N)^2 d[(1-\varepsilon)G(t) + \varepsilon H(t)] = (1 - \frac{k}{N})\sigma_1^2 + \frac{k}{N}\sigma_2^2, \tag{5}$$

$$E\frac{1}{n}S_n = \int (t - ET_n)^2 d[(1-\varepsilon)G(t) + \varepsilon H(t)] =$$

$$= \begin{cases} \frac{1}{n}(N-k)\sigma_1^2 + (n-N+k)\sigma_2^2, & n = N, N-1, \ldots, N-k+1, \\ \sigma_1^2, & n = (N-k), \ldots, 1 \end{cases} \tag{6}$$

$$EL_n \approx \frac{ES_n}{ES_N} = \begin{cases} \frac{N}{n} \times \frac{(N-k)\sigma_1^2 + (n-N+k)\sigma_2^2}{(N-k)\sigma_1^2 + k\sigma_2^2}, & n = N, N-1, \ldots, N-k+1, \\ \frac{N\sigma_1^2}{(N-k)\sigma_1^2 + k\sigma_2^2}, & n = (N-k), \ldots, 1 \end{cases} \tag{7}$$

$$Et_n^2 = \int (t)^2 d[(1-\varepsilon)G(t) + \varepsilon H(t)] = \begin{cases} \sigma_1^2 + \sigma_2^2, & n = N, N-1, \ldots, N-k+1, \\ \sigma_1^2, & n = (N-k), \ldots, 1 \end{cases} \tag{8}$$

где $\sigma_1^2 = \int (t - Et)^2 dG(t), \sigma_2^2 = \int (t - Et)^2 dH(t)$. Рассмотрим первые разности L_n

$$\Delta_n^1 = L_n - L_{n-1} = \frac{(t_{(n)}(n))^2}{S_N}. \tag{9}$$

Найдем среднее значение $E\Delta_n^1(n)$ разности

$$E\Delta_n^1(l) \approx \frac{E(t_{(n)}(n))^2}{ES_N} = \left[(1 - \frac{k}{N})\sigma_1^2 + \frac{k}{N}\sigma_2^2\right]^{-1} \begin{cases} \sigma_1^2 + \sigma_2^2, & n = N, N-1, \ldots, N-k+1, \\ \sigma_1^2, & n = (N-k), \ldots, 1 \end{cases} \tag{10}$$

Как следует из (10), первые разности $E\Delta_n^1(n)$ в случае присутствия *k* выбросов ($n = N, N-1, \ldots, N-k+1$) в среднем постоянны на уровне $D \cdot (\sigma_1^2 + \sigma_2^2)$, а при отсутствии выбросов ($n = (N-k), (N-k-1), \ldots, [{}^{N}\!/_{2}]$) в среднем постоянны на уровне $D \cdot (\sigma_1^2)$, $D = const(N)$. В точке $n = N - k$ происходит скачок функции $E\Delta_n^1(n)$ в среднем на величину $\delta = \sigma_2^2$. Рассмотрим вторые разности $\Delta_n^2(n) = \Delta_n^1(n) - \Delta_{n-1}^1(n)$. Вторые разности в среднем будут равны нулю, а в точке $n=N-k$ происходит дельта-образный всплеск функции $E\Delta_n^2(n)$. Отмеченные особенности поведения статистик L_n, Δ_n^1, Δ_n^2 позволяют построить последовательную процедуру для обнаружения и выделения выбросов, которая обобщает алгоритм маятникового усечения. Для выборки $\vec{x}_N=\{x_1,...x_N\}$ и $n=N,N-1,...,[N/2]$

1. Вычисляем $\bar{T}_n(\vec{x}_n)=\bar{x}_n=\frac{1}{n}\sum_{i=1}^{n}x_i$;
2. Вычисляем $T_i(x_i)=(x_i-\bar{T}_n(\vec{x}_n))$;
3. Упорядочим величины $t_i(n)=|T_i(x_i)|$, $t_{(1)}(n)<t_{(2)}(n)<...<t_{(n)}(n)$;
4. Вычисляем $S_n=\frac{1}{n-1}\sum_{j=1}^{n}(T_i(x_i))^2$;
5. Вычисляем $L_n=\frac{S_n}{S_N}$;
6. Находим первые разности $\Delta_n^1=L_n-L_{n-1}$;
7. Находим вторые разности $\Delta_n^2(n)=\Delta_n^1(n)-\Delta_{n-1}^1(n)$;
8. Наблюдение x_{i_0} соответствующее $t_{(n)}(n)$ удаляем из выборки;
9. Цикл по $n=N,N-1,...,[N/2]$ с пункта 1 по пункт 9.

Отметим, что алгоритм маятникового усечения является непараметрическим, т. е. результат его работы не зависит от вида распределения и автоматически определяет, с какой стороны от центра находится претендент на выброс.

2.2. В качестве меры аномальности и преобразования (1) $T_i(x_i)$ можно использовать функционалы вида $T=\int\varphi(x,\theta)dF(x)$, $T_i(x_i)=\varphi(x_i,\theta_N)-\bar{T}_n(\vec{x}_n,\theta_N)$, $\bar{T}_n(\vec{x}_n)=\frac{1}{n}\sum_{i=1}^{n}\varphi(x_i,\theta_N)$, где $\varphi(x,\theta)$ – непрерывная функция с ограниченной вариацией, θ – параметр, и θ_N – оценка параметра θ.

2.3. Для проверки работоспособности алгоритма маятникового усечения был проведен эксперимент путем моделирования на ЭВМ для симметричных и асимметричных распределений с разной степенью затянутости хвостов.

2.4. Полученные оценки использовались для обработки данных допплеровского мини-содара AV4000. Рабочая частота содара 4900 Гц, длительность импульса излучения 60 мс, период повторения импульсов 4 с. Излучение последовательно посылалось в трех направлениях – вертикально вверх и под углами 14° к вертикали в двух взаимно ортогональных плоскостях. Анализировались данные измерений трех компонентов скорости ветра в 40 высотных стробах вертикальной протяженностью 5 м в диапазоне высот 5–200 м. С целью анализа суточных вариаций первых четырех моментов скорости ветра в АПС, обрабатывались результаты утренних (с 07:00 до 07:10 местного времени), дневных (с 14:00 до 14:10), вечерних (с 19:00 до 19:10), и ночных (с 00:00 до 00:10) измерений. Обрабатывались серии из $N = 150$ профилей, что обеспечивало усреднение за 10-минутный период измерения. Рисунки 1–3 иллюстрируют высотные профили первых четырех моментов компонентов скорости ветра в утренние часы, включая их средние значения V_i, в м/с (*а*), дисперсии σ_i^2, в м^2/с^2 (*б*), и коэффициенты асимметрии $K_{i\,\mathrm{sc}}$ (*в*) и эксцесса $K_{i\,\mathrm{kurt}}$ (*г*), где $i = x, y, z$. В докладе анализируются особенности суточного хода первых четырех моментов компонентов скорости ветра.

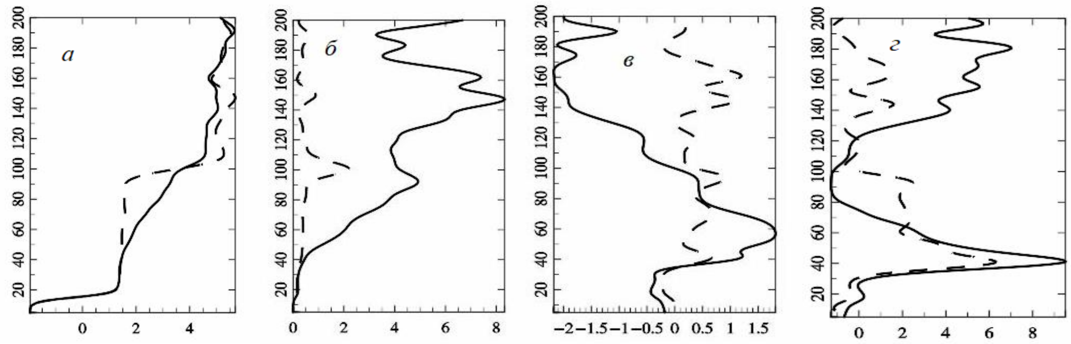

Рисунок 1 – Высотные профили четырех моментов V_x компонента скорости ветра по результатам мини-содарных измерений в утренние часы (СО – сплошные кривые и ММУ– пунктир).

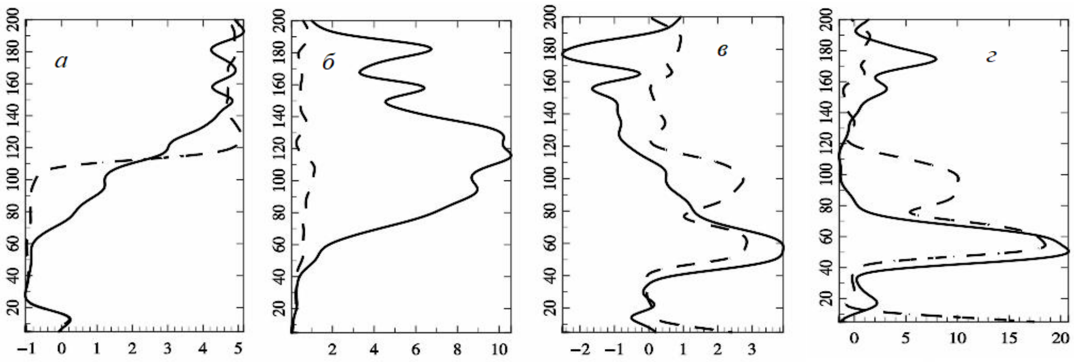

Рисунок 2 – Высотные профили четырех моментов V_y компонента скорости ветра.

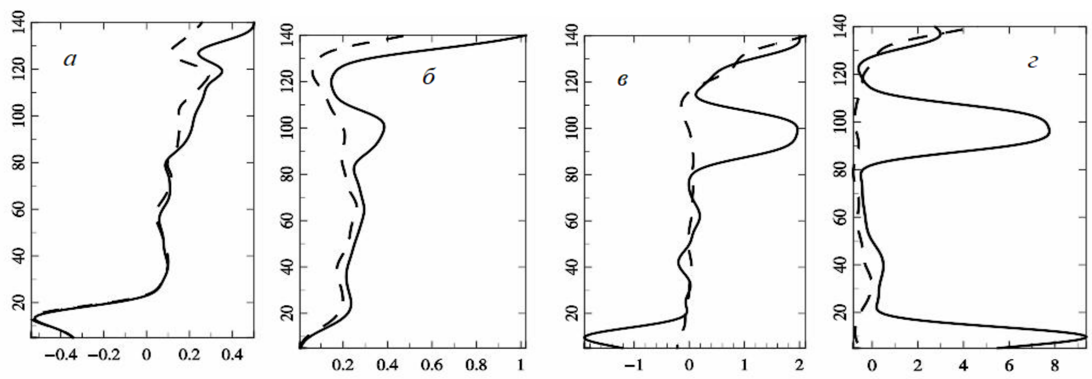

Рисунок 3 – Высотные профили четырех моментов V_z компонента скорости ветра.

1. *Симахин В.А., Черепанов О.С., Шаманаева Л.Г.* Пространственно-временная динамика скорости ветра по результатам мини-содарных измерений // Известия вузов. Физика. 2015. Т. 58. № 12. С. 176–181.
2. *Шуленин В.П.* Робастные методы математической статистики. Томск: Изд-во НТЛ, 2016. 260 с.
3. *R. Muthukrishnan and G. Poonkuzhali.* A Comprehensive Survey on Outlier Detection Methods // American-Eurasian Journal of Scientific Research. 2017. 12 (3). P. 161–171.
4. *Grubbs F.E.* Sample criteria for testing outlying observations // Annals of Mathematical Statistics. 1950. V. 21. P. 27–58.
5. *Орлов А. И.* Неустойчивость параметрических методов отбраковки резко выделяющихся наблюдений //Заводская лаборатория. 1992. № 7. С. 40-42.