КОНФЕРЕНЦИЯ D

ФИЗИКА ТРОПОСФЕРЫ

ЦЕНЗУРИРОВАНИЕ ВЫБОРОК СОДАРНЫХ ИЗМЕРЕНИЙ СКОРОСТИ ВЕТРА С ИСПОЛЬЗОВАНИЕМ МЕТОДА МАЯТНИКОВОГО УСЕЧЕНИЯ

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 содарных измерений высотных профилей трех компонентов скорости ветра в атмосферном пограничном слое (АПС) показывает, что выборки состоят из неоднородных наблюдений с неизвестными распределениями. В докладе для обнаружения и выделения аномальных наблюдений в выборке предложен непараметрический алгоритм маятникого усечения (АМУ), который позволяет не только обнаруживать, но и выделять аномальные наблюдения. Проведено исследование АМУ на модельных примерах. На основе АМУ, производилось цензурирование выборок содарных измерений трех компонентов скорости ветра в АПС, вычислялись их автокорреляционные и структурные функции, проведено их сравнение с классическими выборочными оценками.

1. Введение

Применение мини-содаров для исследования структуры атмосферного пограничного слоя (АПС) выявило 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 – провести цензурирование и использовать усеченные робастные процедуры обработки экспериментальных данных. Вся сложность синтеза усеченных робастных процедур связана с тем обстоятельством, что задачу приходится решать в условиях априорной неопределенности о распределении и доле выбросов [1, 3]. Изначально задачи обнаружения и выделения выбросов рассматривались для экстремальных выбросов в случае одномерных наблюдений [4–6], где был предложен ряд параметрических критериев [4, 5, 6]. Дальнейшие исследования показали, что данные критерии являются неустойчивыми при отклонениях от нормального распределения [6]. В многомерном случае, важнейшее направление исследований связано с задачами корреляционного и регрессионного анализа, определения степени связи и вида зависимости между переменными. В работах [3–7]

показано, что классическая оценка выборочного коэффициента корреляции не является робастной оценкой, а выбросы с ортогональным коэффициентом корреляции по отношению к корреляционной зависимости основной группы могут существенно изменить выборочный коэффициент корреляции. В данном случае под выбросом понимается наблюдение, *удаленное* по мере зависимости от основной группы наблюдений

В данной работе рассмотрен и исследован последовательный непараметрический алгоритм маятникового усечения (АМУ) [8] для обнаружения и выделения выбросов, нарушающих корреляционную структуру двумерного распределения. Проведено моделирование и исследование данного алгоритма для моделей распределений Тьюки и различных моделей выбросов. На основе АМУ, производилось цензурирование выборок содарных измерений трех компонентов скорости ветра и вычислялись их автокорреляционные и структурные функции. Проведено их сравнение с классическими выборочными оценками.

2. АМУ для коэффициента корреляции

Рассмотрим модификацию АМУ [8] для обнаружения и выделения выбросов, изменяющих корреляционную структуру распределения. Пусть $\vec{z}_N = (x_1, y_1), ..., (x_N, y_N)$ – выборка из двумерного распределения $F(\vec{z}) = (1-\varepsilon)G(\vec{z}, \rho) + \varepsilon H(\vec{z}, \rho_1)$, где $G(\vec{z}, \rho)$ – основное априорное распределение с коэффициентом корреляции ρ, $H(\vec{z}, \rho_1)$ – распределение выбросов с коэффициентом корреляции ρ_1, ε – доля выбросов. Введем переменный объём выборки $n = N, N-1, ..., [N/2]$. Рассмотрим статистику $T_i(\vec{z}_i) = (x_i - \bar{x}_n)(y_i - \bar{y}_n) - \bar{T}_n(\vec{z}_n)$, где $\bar{T}_n(\vec{z}_n) = \sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)$, $\bar{x}_n = \sum_{i=1}^{n} x_i$, $\bar{y}_n = \sum_{i=1}^{n} y_i$, $\vec{z}_i = (x_i, y_i)$.

Упорядочим величины $t_i(n) = |T_i(\vec{z}_i)|$, $t_{(1)}(n) < t_{(2)}(n) ... < t_{(n)}(n)$. Обозначим через $k = [\varepsilon N]$ количество выбросов в выборке $\vec{z}_N$. Рассмотрим последовательную процедуру определения *претендентов на выброс*. Выбросы по коэффициенту корреляции представляют крайние порядковые статистики $t_{(N-k+1)}(n), ..., t_{(n)}(n)$. Наблюдение $\vec{z}_j$ $\left(\vec{z}_j = \arg\max |T_i(\vec{z}_i)|\right)$, соответствующее $t_{(n)}(n)$, является претендентом на выброс, поэтому удаляем его из выборки $\vec{z}_n = (x_1, y_1), ..., (x_n, y_n)$. В результате получаем выборку $\vec{z}_{n-1}$ объёма $(n-1)$.

Процедуру выявления претендентов на выброс повторяем с $n = N, N-1, ..., [N/2]$. Для определения, какие из претендентов являются выбросами, введем статистику $L_n = \frac{S_n}{S_N}$,

$S_n = \frac{1}{n}\sum_{i=1}^{n}(T_i - \bar{T}_n)^2$, $n = N, N-1, ..., [N/2]$. Тогда $S_n = S_{n-1} + (t_{(n)}(n))^2$, $S_N = const(N)$, то есть, $S_{n-1} < S_n$ и следовательно, статистика $0 < L_n \le 1$ является монотонно убывающей функцией от n. Рассмотрим первые разности L_n, $\Delta_n^1 = L_n - L_{n-1} = \frac{(t_{(n)}(n))^2}{S_N}$. Можно показать [8], что если в выборке присутствует k выбросов, то первые разности $\Delta_n^1(n)$ испытывают в точке $n = N - k$ скачок в среднем на величину δ. Вторые разности $\Delta_n^2 = \Delta_n^1 - \Delta_{n-1}^1$ 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 последовательный алгоритм для обнаружения и выделения выбросов в корреляционной зависимости, который обобщает АМУ [9].

1. Вычисляем средние $\bar{X}_{N-l} = \frac{1}{N-l}\sum_{i=1}^{N-l} X_i$, $\bar{Y}_{N-l} = \frac{1}{N-l}\sum_{i=1}^{N-l} Y_i$
2. Вычисляем $T_i = (x_i - \bar{X}_{N-l})(y_i - \bar{Y}_{N-l})$, $i = 1, ...N-l$
3. Вычисляем среднее $\bar{T}_{N-l} = \frac{1}{N-l}\sum_{i=1}^{N-l} T_i$
4. Вычисляем $S_{N-l} = \frac{1}{N-l}\sum_{i=1}^{N-l}(T_i - \bar{T}_{N-l})^2$
5. Вычисляем $L_l = \frac{S_{N-l}}{S_N}$
6. Находим максимальное T_i, наблюдение $z_i = (x_i, y_i)$ удаляем из выборки
7. Находим первые разности $\Delta_l^1 = L_{l+1} - L_l$
8. Находим вторые разности $\Delta_l^2 = \Delta_{l+1}^1 - \Delta_l^1$

7. Цикл с пункта 1 по пункт 7, по $l = 0$ до $N/2$.

Для исследования работы АМУ, был проведен численный эксперимент. В качестве модели выбросов, рассматривалась модель Тьюки двумерного нормального распределения $F(\vec{z}) = (1-\varepsilon)G(\vec{z}) + \varepsilon H(\vec{z})$, $G(\vec{z}, \rho_1) = \Phi(\mu_1^{(1)} : \mu_2^{(1)} : (\sigma_1^{(1)})^2 : (\sigma_2^{(1)})^2 : \rho_1)$, $H(\vec{z}, \rho_2) = \Phi(\mu_1^{(2)} : \mu_2^{(2)} : (\sigma_1^{(2)})^2 : (\sigma_2^{(2)})^2 : \rho_2)$, $\Phi(\mu_1^{(i)} : \mu_2^{(i)} : (\sigma_1^{(i)})^2 : (\sigma_2^{(i)})^2 : \rho_i)$ со средними значениями $EX = \mu_1^{(i)}$, $EY = \mu_2^{(i)}$, дисперсиями $DX = (\sigma_1^{(i)})^2$, $DY = (\sigma_2^{(i)})^2$, коэффициентом корреляции ρ_i, ε – доля выбросов. В эксперименте основная выборка генерировалась из распределения $G(\vec{z}, \rho_1) = \Phi(0:0:1:0,2:0,9)$ и доля выбросов (10%) $\varepsilon = 0,1$.

Эксперимент: N = 20; $\varepsilon = 0,1$ (N = 20 = 18 + 2 выброса); $G(\vec{z}, \rho_1) = \Phi(0:0:1:0,2:0,9)$; $H(\vec{z}, \rho_2) = \Phi(0:0:1:0,2:-0,9)$. Выборочный коэффициент корреляции без выбросов

$R_B = 0.93$. Выборочный коэффициент корреляции с выбросами R_B = 0.42. Критерий независимости на основе статистики $T_{набл} = R_B \cdot \sqrt{N-2} \Big/ \sqrt{1-R_B^2}$ при уровне значимости $\alpha = 0,01$ и критическом значении $T_{крит} = 2,88$ показывает: с выбросами $T_{набл} = 2.04 < T_{крит} = 2,88$ – нулевая гипотеза принимается; без выбросов $T_{набл} = 7.61 > T_{крит} = 2,88$ – нулевая гипотеза отвергается. Выбросы серьезно исказили ситуацию. Без выбросов R_B = 0.93, и критерий однозначно отвергает нулевую гипотезу, но присутствие двух выбросов понижает R_B почти в два раза до R_B = 0.42, и критерий однозначно принимает нулевую гипотезу.

Применяем АМУ. На рисунках 1, 2, 3 приведены результаты работы АМУ при N = 18 + 2 выброса в зависимости от n. АМУ выделяет 2 выброса. После удаления 2 выбросов, R_B с 0.42 становится 0.93. Аналогичный и более значимый результат получается и на выборках большего объема.

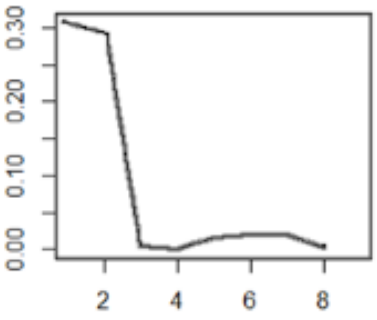

Рисунок 1 – Δ_n^1

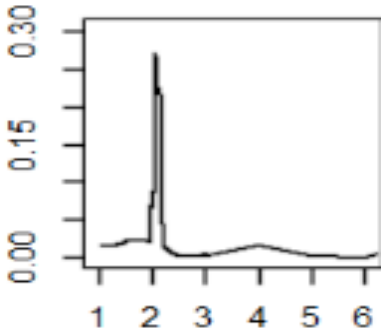

Рисунок 2 – Δ_n^2

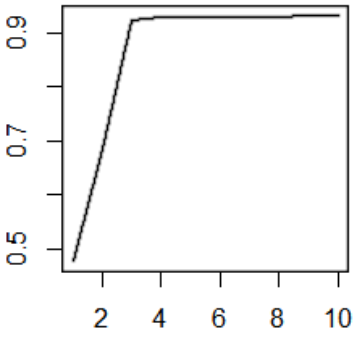

Рисунок 3 – R_B

3. Автокорреляционные и структурные функции трех компонентов скорости ветра по результатам мини-содарных измерений

АМУ использовался 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. Анализировались и 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. С помощью АМУ, осуществлялось цензурирование выборок. Рисунки 4 и

5 иллюстрируют автокорреляционные функции $\rho(\tau)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(x_{i+\tau}-\bar{x})$, а рисунки 6 и 7 – структурные функции $St(\tau)=\frac{1}{n}\sum_{i=1}^{n}(x_i-x_{i+\tau})^2$ для V_x–компонента скорости ветра и высот 45, 180, 35 и 175 м, соответственно, в утренние часы. Здесь красным цветом показана робастная непараметрическая оценка, и черным цветом – классическая неробастная оценка.

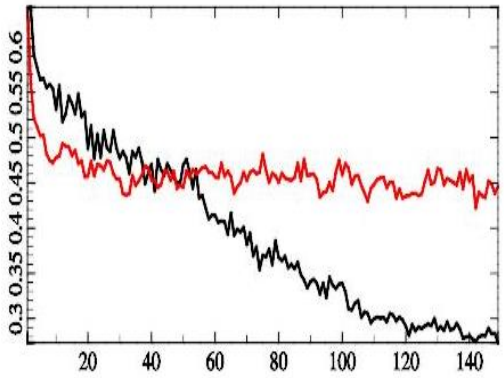

Рисунок 4 – $\rho(\tau)$ (V_x; z = 45 м, 08:00–08:10)

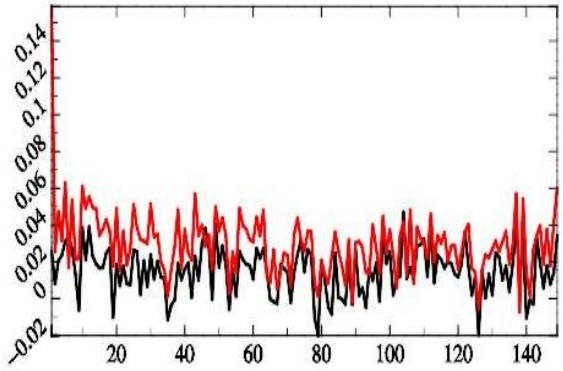

Рисунок 5 – $\rho(\tau)$ (V_x; z = 180 м, 08:00–08:10)

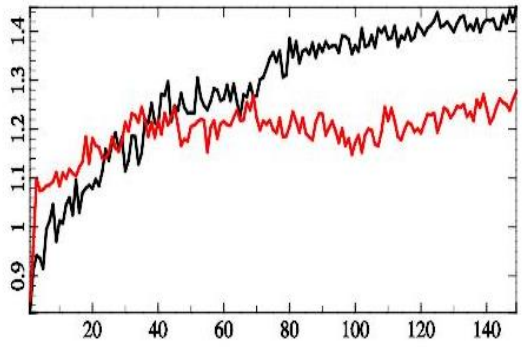

Рисунок 6 – $St(\tau)$ (V_x; z = 35 м, 08:00–08:10)

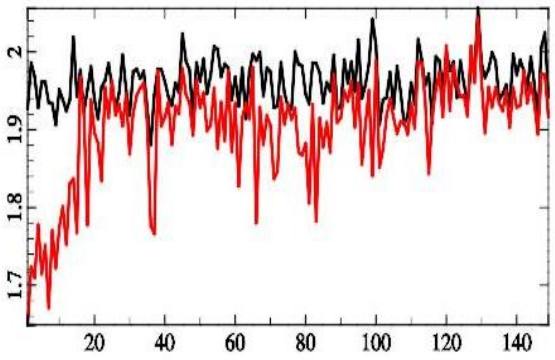

Рисунок 7 – $St(\tau)$ (V_x; z = 175 м, 08:00–08:10)

1. *Федоров В.А.* Измерения содаром «Волна 3» параметров радиальных компонент вектора скорости ветра//Оптика атмосферы и океана. 2003. Т. 16. № 2. С. 151–155.

2. *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.

3. *Шуленин В.П.* Робастные методы математической статистики. Томск: Изд-во НТЛ, 2016. 260 с.

4. *Muthukrishnan R. and Poonkuzhali G.* A Comprehensive Survey on Outlier Detection Methods // American-Eurasian Journal of Scientific Research. 2017. V. 12 (3). P. 161–171.

5. *Кобзарь А. И.* Прикладная математическая статистика. Для инженеров и научных работников. - М.: Наука, 2006. 816 с.

6. *Орлов А.И.* Неустойчивость параметрических методов отбраковки резко выделяющихся наблюдений //Заводская лаборатория. 1992. № 7. С. 40–42.

7. *Shevlyakov G.L., Vilchevski N.O.* Robustness in data analysis: criteria and methods. Utrecht: VSP, 2002. 315 p.

8. *Симахин В.А., Черепанов О.С., Шаманаева Л.Г.* Обнаружение и выделение аномальных наблюдений при акустическом зондировании скорости ветра. Настоящий сборник. 2019. Томск: Изд-во ИОА СО РАН.