

УДК 519.2
**О РАЗРАБОТКЕ БИБЛИОТЕКИ ПРОЦЕДУР НА ЯЗЫКЕ R
ДЛЯ КОМПЬЮТЕРНОГО МОНИТОРИНГА ДАННЫХ
И ПРИНЯТИЯ РЕШЕНИЙ НА ОСНОВЕ МЕТОДОВ
ПОСЛЕДОВАТЕЛЬНОГО СТАТИСТИЧЕСКОГО АНАЛИЗА**

А. Ю. Харин¹⁾, А. А. Старостина²⁾, Е. А. Крученков³⁾

*1), 2), 3) Белорусский государственный университет, пр. Независимости 4,
220030, г. Минск, Беларусь, KharinAY@bsu.by*

Кратко рассматриваются особенности разрабатываемой в Учреждении Белорусского государственного университета «Научно-исследовательский институт прикладных проблем математики и информатики» библиотеки процедур на языке R для компьютерного мониторинга стохастических данных и принятия решений на основе методов последовательного статистического анализа.

Ключевые слова: язык R, компьютерные процедуры, последовательный статистический анализ.

**ON DEVELOPMENT OF R PROCEDURES LIBRARY
FOR COMPUTER MONITORING OF DATA
AND DECISION MAKING
BASED ON SEQUENTIAL ANALYSIS METHODS**

A. Yu. Kharin¹⁾, A. A. Starostina²⁾, E. A. Kruchenkov³⁾

*1), 2), 3) Belarusian State University, 4 Independence av.,
Minsk 220030, Belarus, KharinAY@bsu.by*

Features of the R procedures library developed in the Research institute of applied mathematics and informatics at the Belarusian state university are briefly considered. The library is developed for computer monitoring of stochastic data and decision making based on the methods of sequential statistical analysis.

Keywords: R statistical software, computer procedures, sequential statistical analysis.

Введение

Задачи, возникающие на современном этапе развития информационного общества, часто приводят к обработке данных, представляющих собой потоки наблюдений, имеющих стохастическую составляющую, пренебречь которой не удаётся. Мониторинг таких потоков проводится с целью принятия возможных управляющих решений, зависящих от неко-

торых свойств наблюдаемого потока. Для обеспечения максимальной эффективности принятия решений в такой ситуации может быть использован последовательных статистический анализ Вальда [1]. Он основан на предположении о том, что число наблюдений, по которым принимается решение, не фиксируется заранее, а является случайным, зависящим от самих стохастических наблюдений специальным образом. Это позволяет обеспечить заданную точность (малые вероятности ошибочных решений) ценой лишь минимального в среднем числа наблюдений. Такой подход может использоваться при решении прикладных задач анализа данных, возникающих в области мониторинга качества выпускаемой продукции и в медицине для наблюдения за процессами лечения, что позволяет достичь существенного экономического эффекта [2].

1. Разрабатываемый функционал библиотеки

В типовом наборе процедур принятия решений на основе потоков наблюдаемых данных, предлагаемом современными компьютерными средствами анализа данных, отсутствуют процедуры, реализующие такой подход. Во многом такая ситуация обусловлена математической сложностью вычисления характеристик эффективности последовательных тестов при заданном наборе параметров даже для простейших моделей потоков данных. В Учреждении Белорусского государственного университета «НИИ прикладных проблем математики и информатики» имеется в этом направлении существенный задел в виде разработанного подхода, позволяющего решать эту задачу для широкого класса моделей наблюдений. Подход основан на дискретизации пространства значений критериальной статистики и использовании специально выстраиваемых цепей Маркова для её аппроксимации [3], [4]. Актуальность этой проблемы отмечалась в 2021 г. на международной научной конференции по робастной статистике (Вена, Австрия) [5].

В учреждении Белорусского государственного университета «НИИ прикладных проблем математики и информатики» решается система взаимосвязанных задач для разработки библиотеки процедур для мониторинга потоков данных и принятия решений на основе последовательного анализа в рамках следующих моделей данных: 1) потоки однородных наблюдений; 2) потоки с трендом; 3) потоки данных регрессионного типа; 4) потоки данных с марковской зависимостью; 5) потоки данных рекуррентного типа, включая авторегрессионные.

Для каждого типа потоков стохастических данных разрабатываются R-модули, позволяющие: а) при помощи имитационного моделирования по методу Монте—Карло лучить оценки характеристик эффективности процедур принятия решений (случаи простых и сложных параметрических гипотез рассматриваются отдельно); б) исследовать влияние искажений гипотетических моделей потоков на характеристики эффективности; в) построить робастные статистические последовательные решающие правила в рамках задаваемых семейств и исследовать их характеристики эффективности в различных условиях.

2. Примеры графических представлений результатов

Приведём для иллюстрации два примера работы процедур разрабатываемой библиотеки. Первый из примеров показывает график зависимости оценки вероятности ошибки от уровня искажений, когда последовательность представлена независимыми одинаково распределёнными наблюдениями (см. рис. 1).

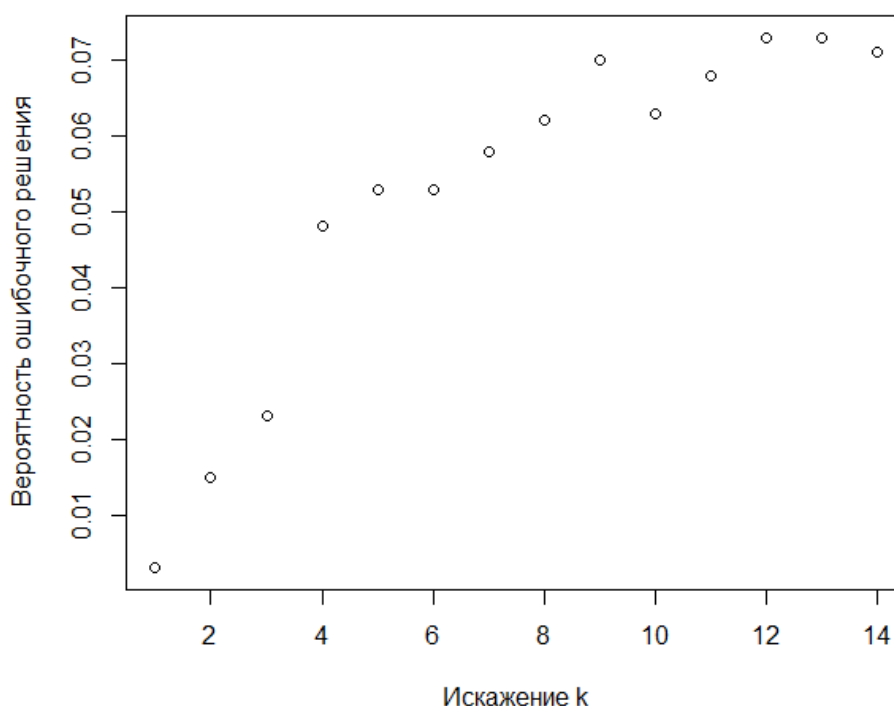


Рис. 1. Зависимость оценки вероятности ошибки от уровня искажения

Второй пример показывает траекторию принятия решений последовательным статистическим решающим правилом в задаче анализа авторства текстов, когда модель наблюдений – сложная цепь Маркова.

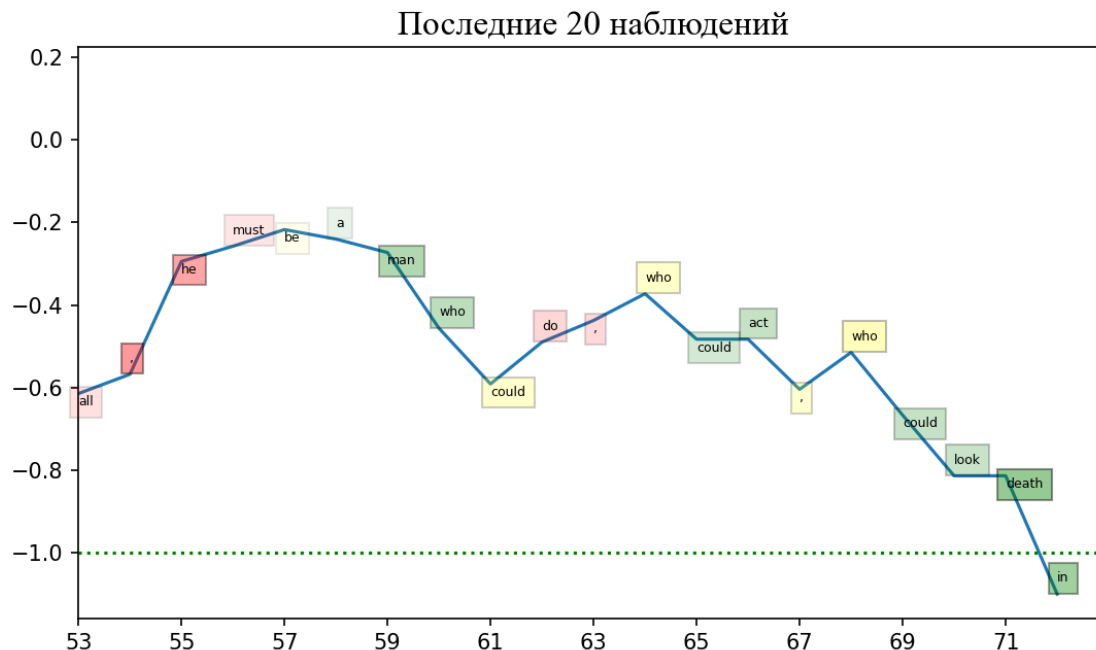


Рис. 2. Траектория критериальной статистики для сложной марковской зависимости

Исследования выполняются в рамках проекта по Государственной программе научных исследований «Цифровые и космические технологии, безопасность человека, общества и государства», а в части разработки методов и процедур для анализа рекуррентных случайных процессов поддержаны грантом Белорусского республиканского фонда фундаментальных исследований № Ф23УЗБ-080.

Библиографические ссылки

1. Wald A. Sequential analysis. New York : John Wiley and Sons, 1947.
2. Mukhopadhyay N., B. de Silva Sequential methods and their applications. New York : Marcel Dekker, 2009.
3. Kharin A. An approach to asymptotic robustness analysis of sequential tests for composite parametric hypotheses // Journal of Mathematical Sciences. 2017. Vol. 227 (2). P. 196–203.
4. Kharin A.Y., Kishylau D.V. Robust sequential test for hypotheses about discrete distributions in presence of “outliers” // Journal of Mathematical Sciences. 2015. Vol. 205 (1). P. 68–74.
5. Kharin A. Robustness in sequential decision making on parameters of stochastic data flows // International Conference on Robust Statistics’2021. Materials. TU Wien : Vienna, 2021. 2 p.