КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2013 Т. 5 № 4 С. 589–595

Ки&М

ЧИСЛЕННЫЕ МЕТОДЫ И ОСНОВЫ ИХ РЕАЛИЗАЦИИ

УДК: **004.942**

Методика и программа для накопления и статистического анализа результатов компьютерного эксперимента

А. Р. Winn[1,a], Т. Чжо[2], В. М. Трояновский[2,b], Я. Л. Аунг[2]

[1] Department of Computer Science, Defense Services Academy
the Union of Myanmar
[2]Национальный исследовательский университет «МИЭТ»
Россия, 124498, г. Москва, г. Зеленоград, проезд 4806, д. 5

E-mail: [a] apw2009@gmail.com, [b] troy40@mail.ru

Получено 6 апреля 2013 г.,
после доработки 21 сентября 2013 г.

Решается задача накопления и статистического анализа результатов компьютерного эксперимента. Программа основного эксперимента рассматривается в рамках разработанной методики как источник данных, собираемых на специально подготовленный лист Excel с заранее организованной структурой для накопления, статистической обработки и визуализации данных. Созданная методика и программа использованы при исследовании эффективности корреляционных методов выделения гармонического сигнала на фоне помех по реализации ограниченной длины.

Ключевые слова: компьютерная программа, статистический анализ, компьютерный эксперимент, доверительные интервалы, обработка данных

Methodology and program for the storage and statistical analysis of the results of computer experiment

A. P. Winn[1], Htin Kyaw[2], V. M. Troyanovskyi[2], Y. L. Aung[2]

[1]Department of Computer Science, Defense Services Academy, the Union of Myanmar
[2]National Research University "MIET", 5 pas 4806, Zelenograd, Moscow, 124498, Russia

Abstract. – The problem of accumulation and the statistical analysis of computer experiment results are solved. The main experiment program is considered as the data source. The results of main experiment are collected on specially prepared sheet Excel with pre-organized structure for the accumulation, statistical processing and visualization of the data. The created method and the program are used at efficiency research of the scientific researches which are carried out by authors.

Keywords: computer program, statistical analysis, computer experiment, confidential intervals, data processing

Citation: *Computer Research and Modeling*, 2013, vol. 5, no. 4, pp. 589–595 (Russian).

Введение

Воспроизводимость экспериментальных результатов и анализ «парадоксальных» наблюдений [Шноль, 1984] являются важными проблемами научных исследований. Большая совокупность разнородных условий и ограничений препятствует применению классических подходов для решения этих проблем не только в биологии, но и в технических науках и приложениях.

Среди них можно выделить следующие [Трояновский, 2004]:

- стохастичность воздействий и малая изученность объектов;
- принципиально ограниченная длина доступных реализаций;
- динамическое преобразование сигналов в объекте исследования.

Известны результаты преодоления указанных ограничений для задач идентификации [Serdyuk , Troyanovskyi, 2009] и для поиска скрытых периодичностей [Aung, Troyanovskyi, 2011].
Важные методические особенности решения задач с учетом названных условий состоят в следующем:

- стохастичность воздействий и малая изученность объектов требуют привлечения статистических методов для анализа результатов;
- классические статистические методы (явно или не явно) используют гипотетическое множество событий, значений или реализаций; именно на этом множестве определяются вероятности и другие статистические характеристики;
- результат каждого эксперимента можно рассматривать как отдельную случайную величину, частную реализацию искомой функции и т. п.; все они – лишь частные оценки параметров исследуемого процесса;
- в прикладных исследованиях чаще всего предполагается стационарность и эргодичность процессов; однако из-за принципиально ограниченной длины доступных реализаций эргодичность важна лишь как принципиально важная основа для применения статистического подхода, а на первое место выступает необходимость определения статистических свойств доступной выборки данных, и именно здесь возникает проблема различий при усреднении по множеству и по времени.

При сечении множества поперек ансамбля статистически независимых реализаций (рис. 1) в рассмотрение попадают независимые отсчеты. К тому же мощность гипотетического множества может быть бесконечной. Исследователь всегда работает с единственной реализацией ограниченной длины. В то же время отсчеты, выбираемые из частной реализации, могут оказаться коррелированными, а их число всегда ограничено.

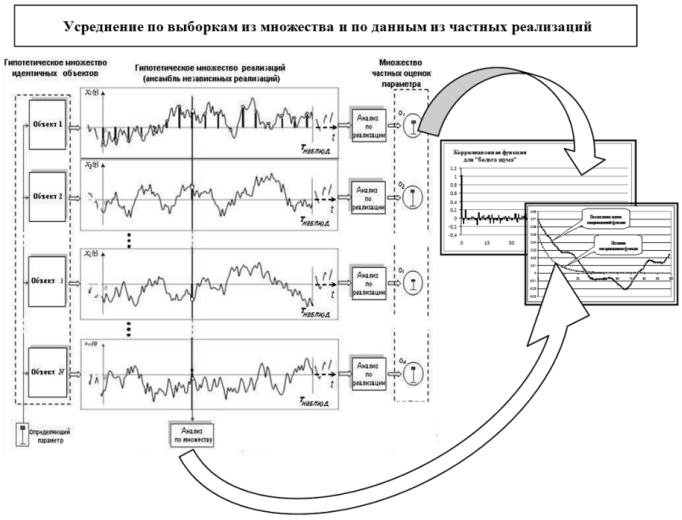

Рис. 1. Искажение вида корреляционной функции в ее оценке, вычисляемой по реализации ограниченной длины

В результате вместо искомых характеристик и функций исследователь, как правило, получает лишь их оценки, которые могут сильно отличаться от истинных характеристик и функций. Верификация, как обязательное требование к научным исследованиям, позволяет выявить степень соответствия таких оценок истинным функциям и их параметрам. Если научное исследование использует компьютерный эксперимент, то становится возможным дополнительно повысить качество верификации за счет оперативной визуализация результатов и получения доверительных интервалов для накопленных данных.

Стратегия решения задачи

Для наглядности рассмотрим задачу оценки корреляционной функции, вычисляемой по реализации ограниченной длины. Известно, что корреляционная функция является важнейшей характеристикой для случайного процесса. Она позволяет не только определить величину дисперсии у флуктуации случайного процесса, но и выявить скрытые периодичности в зашумленных процессах. Однако при ограниченной длине реализации сама корреляционная функция становится не точной, и имеет собственные флуктуации (рис. 1).

Стратегия создания специальной методики и программы для накопления и статистического анализа результатов компьютерного эксперимента представлена на рисунке 2.

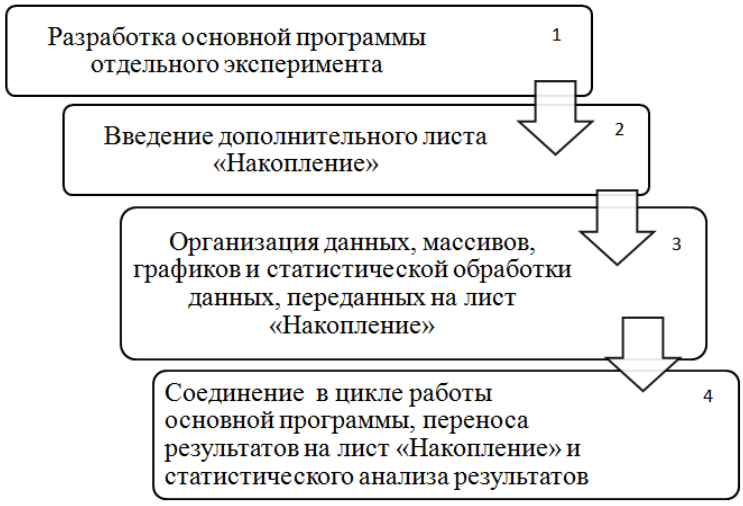

Рис. 2. Стратегия накопления и статистического анализа результатов компьютерного эксперимента

В рамках этой методики программа основного эксперимента рассматривается как источник данных, собираемых дополнительным модулем нашей программы на специально подготовленный лист Excel. Будем предполагать, что весь компьютерный эксперимент проводится в среде Excel+VBA (или результаты могут сохраняться в одном из форматов, доступных Excel). Эта среда обладает не только мощными вычислительными возможностями, наличием более 300 встроенных функций и возможностью дополнительно вводить собственные подпрограммы обработки на языке VBA. Она отличается еще и хорошей организацией данных и средств их отображения, наличием развитых средств построения диалога, разнообразных режимов помощи и презентаций.

При таком подходе на дополнительном листе может быть заранее организована структура для накопления, необходимой статистической обработки и визуализации данных. Благодаря наличию связей, установленных между табличными данными, формулами и графиками, сохраняемых в Excel-программах независимо от изменяющихся данных, такая программа с заранее запрограммированной статистической обработкой удобна для интерактивного применения в качестве помощника экспериментатору.

Алгоритм программы, реализующей предложенную методику, приведен на рисунке 3. Программа позволяет разделить этапы единичных экспериментов и верификации результатов,

соединив в едином цикле проведение эксперимента, накопление результатов (путем переноса данных в таблицу на лист «Накопление») и их немедленную статистическую обработку.

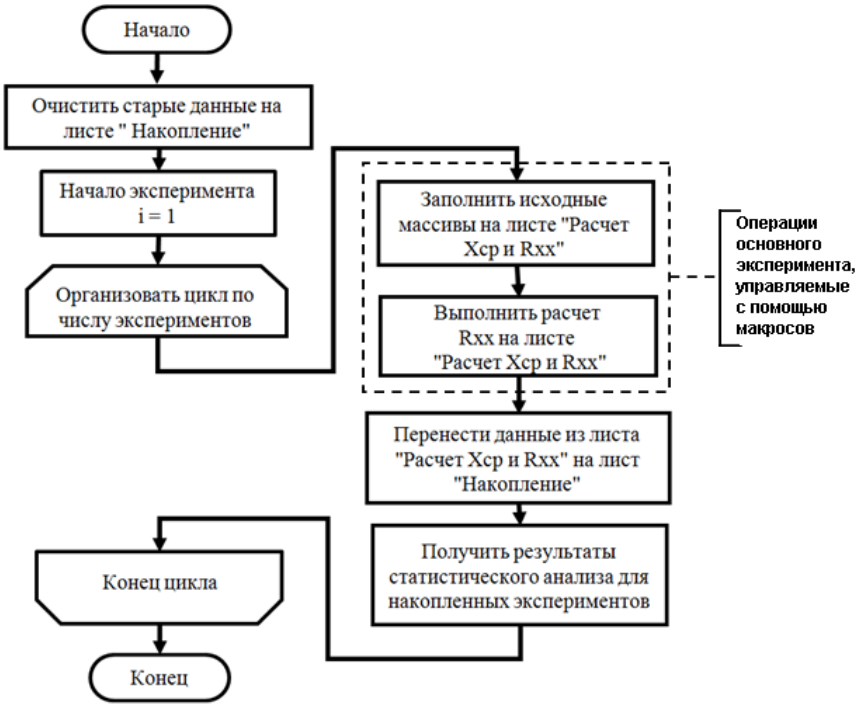

Рис. 3. Алгоритм программы для реализации методики

Применительно к рассматриваемой задаче оценки корреляционной функции все необходимые расчеты основной программы (см. блок 1 на рис. 2) проводятся на листе EXCEL, названном «Расчет Xcp и Rxx». На указанном листе выделены необходимые области ячеек, а наполнение массивов и расчеты организованы с помощью дополнительно разработанных макросов. Именно эти макросы запускаются как в ходе единичных экспериментов, так и в общем цикле верификации результатов (рис. 3).

Наиболее важны для реализации методики введение дополнительного листа «Накопление» и организация данных, массивов и статистической обработки данных, переданных на лист «Накопление» (блоки 2 и 3 на рис. 2). Для накопления данных на этом листе (рис. 4) для каждого *i*-го эксперимента организованы массивы, в элементах которых фиксируются все значимые параметры (копия данных об условиях эксперимента, а также его итоговые результаты – в рассматриваемом примере это ординаты корреляционной функции $R_{xx}(j)$). Поскольку для этих данных используется табличная форма, то по накапливаемым данным средствами EXCEL легко проводится необходимая статистическая обработка, в первую очередь, вычисление средних значений и стандартных отклонений. При этом диапазон ячеек, указываемый для статистических расчетов, может быть указан с большим запасом – EXCEL игнорирует незаполненные ячейки. Благодаря наличию связей, установленных между табличными данными, формулами и графиками, сохраняемых в Excel-программах независимо от изменяющихся данных, заранее запрограммированная статистическая обработка обеспечивает интерактивное применение для любых новых данных.

Кроме того, на листе «Накопление» может быть организован расчет оценок доверительных интервалов на основе теоретических соображений или на основе рассчитанных стандартных отклонений.

Дополнительную наглядность результатам верификации придает визуализация данных в виде графиков желаемого числа отдельных экспериментов вместе с графиками доверительных интервалов, рассчитанных по накопленным данным (см. ниже рис. 5). Так же, как и расчет

средних и стандартных отклонений, такие графики единожды создаются с использованием данных из массива накапливаемых результатов на листе «Накопление», и EXCEL автоматически перестраивает графики, отслеживая любые изменения этих данных. Последнее придает дополнительную универсальность алгоритму, представленному на рисунке 3.

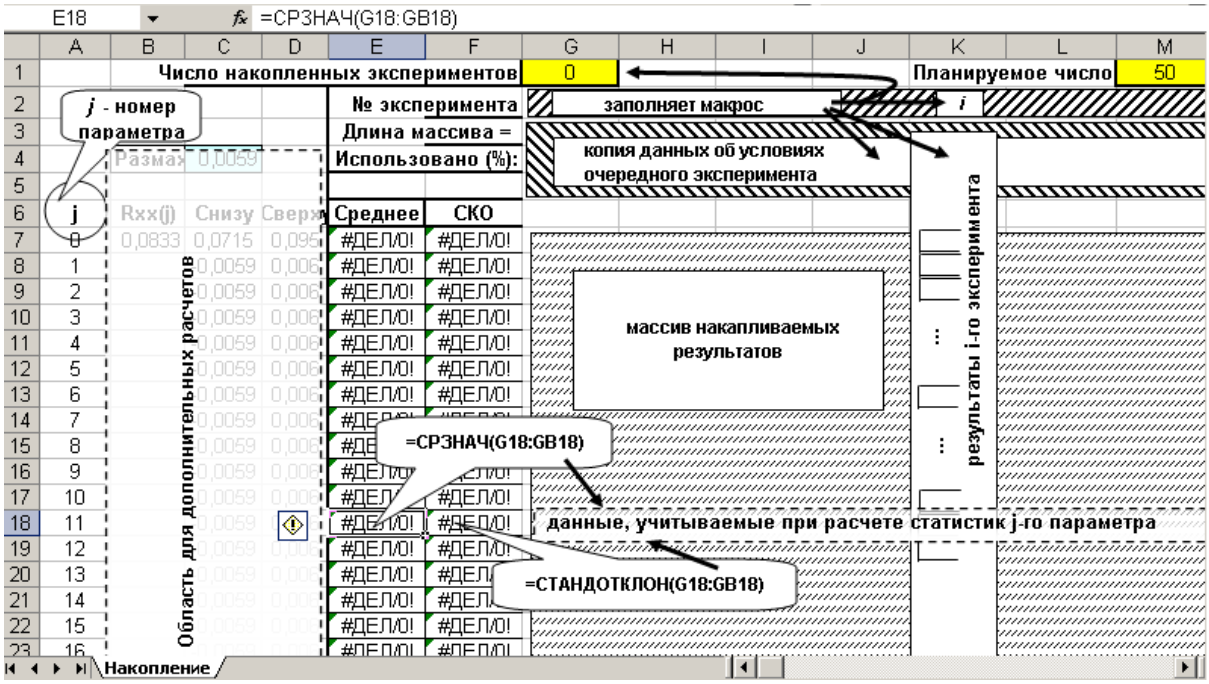

Рис. 4. Организация полей с данными и формулами на листе «Накопление»

Отдельно заметим, что после создания описанной структуры данных, формул и графиков на листе «Накопление» большая часть алгоритма на рисунке 3 реализуется в виде макросов на VBA и сводится к простым операциям организации цикла, вызова макросов для выполнения действий основного эксперимента и пересылки данных между рабочими листами рабочей книги EXCEL. Вместе с рисунком 4 это фактически позволяет использовать созданную методику и программу как свободно распространяемое ПО.

Применение методики и программы

Полученные результаты были применены авторами при проведении научных исследований, использующих корреляционные методы анализа данных.

а) Исследование статистических свойств автокорреляционной функции. Оценивание свойств автокорреляционной функции является одной из первоочередных задач, решение которой необходимо для уверенного применения корреляционных методов в любых практических применениях.

После центрирования доступной реализации путем вычитания среднего (с использованием закона распределения амплитуд сигнала — если он известен, или с вычислением среднего по доступным данным) вычисление значений ординат корреляционной функции по реализации ограниченной длины с помощью соотношения

$$R_{xx}\ j \quad \frac{1}{N}\ {}_{i\ 1}^{N}\ x\ i\ x\ i \quad j$$

дает для (центрированного) стационарного процесса несмещенную, но флуктуирующую оценку (см. рис. 1).

Определение уровня этих флуктуаций – сложная задача, решавшаяся лишь при определенных допущениях в 50–60-х годах XX века такими известными учеными, как В. С. Пугачев, Г. Дженкинс и др. В этой связи визуализация картины соответствующих флуктуаций для различных практических ситуаций представляется весьма полезной.

На рисунке 5а показаны результаты, полученные с помощью нашей программы для сигнала в виде белого шума. Они показывают, что статистические флуктуации имеют все ординаты корреляционной функции. Важно отметить, что для оценки дисперсии (для ординаты $R_{xx}(0)$) разброс величин от эксперимента к эксперименту существенно меньше самой ординаты[1], а вот для остальных ординат (при $j \neq 0$), где для белого шума $R_{xx}(j)\big|_{j\ 0}\quad 0$, наличие остаточных статистических флуктуаций – факт очень неприятный, поскольку результат всегда сравнивается с нулем.

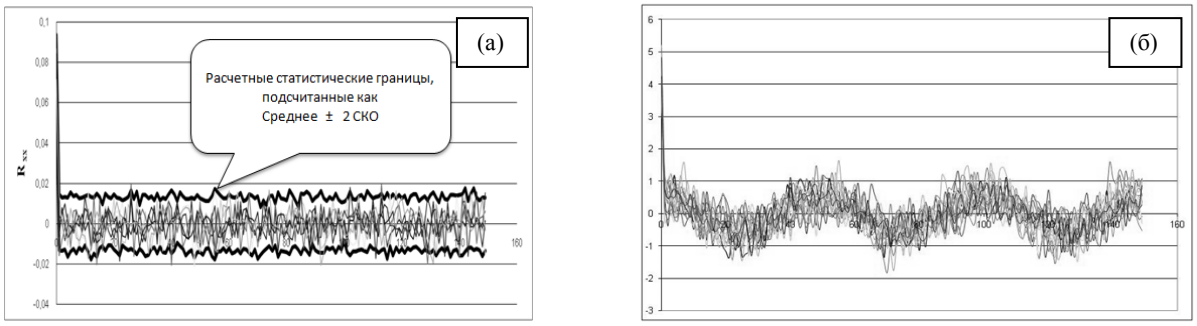

Рис. 5. Результаты работы программы при определении доверительных интервалов для оценки корреляционной функции (а) и в задаче выделения гармонического сигнала на фоне помех (б)

Количественная оценка этих явлений проведена в [Aung Phyo Winn, Serdyuk, Troyanovskyi, 2013], где авторы получили новые теоретические результаты для свойств корреляционных функций сигналов в виде белого шума и окрашенного сигнала. Методика и программа, описываемые в данной работе, позволили провести верификацию этих результатов.

б) Выделение гармонического сигнала на фоне помех. Как известно (см., например, [Вентцель, 2002]), ковариационная функция суммы независимых случайных процессов равна сумме их ковариационных функций. Соответственно, ковариационная функция суммы полезного гармонического сигнала и помехи состоит из ковариационной функции помехи и косинусоиды, частота которой равна частоте полезного сигнала. Для помехи в виде белого шума ковариационная функция – это δ-функция, и можно ожидать, что уже при небольших значениях аргумента ковариационная функция общего сигнала должна содержать чистую косинусоиду.

На практике картина выглядит иначе. При переходе от гипотетического множества к использованию реализаций ограниченной длины остаточные статистические флуктуации корреляционной функции помехи (см. выше п. а) распространяются на весь рассматриваемый диапазон аргументов. В результате оценка ковариационной функции суммарного сигнала приобретает флуктуации при любых значениях аргументов. При накоплении данных от отдельных экспериментов можно видеть как идеальная косинусоида «размывается», оставаясь внутри некоторого «коридора». На рисунке 5б показаны соответствующие результаты, где на график выведены для наглядности 50 корреляционных функций гармонического сигнала с аддитивной помехой, рассчитанные по реализациям (одинаковой) ограниченной длины.

Границы доверительных интервалов (рис. 5а) могут быть установлены программным путем. По умолчанию программа использует величину ± 2 среднеквадратических отклонений (СКО), рассчитываемых по выборке из полученных экспериментальных данных. Для данных с нормальным законом распределения это определяет интервал, в который попадают примерно 95 % всех отсчетов.

[1] И зависит от длины реализации и вида закона распределения сигнала.

Заключение

1. Разработана методика и специальный программный модуль для накопления и статистической обработки данных, получаемых при повторении заданного числа экспериментов.
2. Накопленные данные позволяют визуально оценивать границы их разброса.
3. Разработанная программа позволяет оперативно оценивать точность экспериментальных результатов.
4. Разработанная программа была успешно использована для верификации теоретических оценок статистических флуктуаций корреляционных функций, вычисляемых по реализациям ограниченной длины, а также в задаче выделения полезного сигнала на фоне помех.

Список литературы

Вентцель Е. С. Теория вероятностей / учебник для вузов. — Изд. 7-е. — М.: Высшая школа, 2002. — 380 с.

Трояновский В. М. Информационно-управляющие системы и прикладная теория случайных процессов: учебное пособие. — М.: Гелиос АРВ, 2004. — 304 с.

Шноль С. Э. Кибернетика живого. Биология и информация. — М.: Наука, 1984 — 84 с.

Aung Phyo Winn , Dr. V. M. Troyanovskyi. A New Algorithmic Approach for Detecting Hidden Periodicity of Noisy Signals in Technically Complicated Systems / Ninth International Conference on Computer Application (ICCA 2011), Yangon, 2011. — P. 103–108.

Aung Phyo Winn, Serdyuk O. A., Troyanovskyi V. M. Transformation of the structure and parameters of a shaping filter into confidence intervals for correlation function estimate // Preprints of the 2013 IFAC Conference on Manufacturing Modelling, Management, and Control, Saint Petersburg, Russia, June 19–21, 2013. — P. 1860–1865.

Serdyuk O. A., Troyanovskyi V. M. The ideal representations, rocks and realities of statistical methods' identification. 2009 IEEE International Conference on Control Applications — IEEE Catalog Number CFP09CCA-CDR, ISBN: 978-1-4244-4602-5, p. 1472–1476.