УДК 004.588

ББК 74.5

Ч-751

В. З. Чокой

Иркутск, Россия

ОБРАБОТКА И РАЗВЕДОЧНЫЙ АНАЛИЗ ЧИСЛОВЫХ МАССИВОВ ДАННЫХ

Управление процессами обслуживания авиационной техники предполагает выполнение разнообразных комплексов работ логистического, организационно-технического и технологического характера. Общими для перечисленных составляющих являются вопросы наблюдений за объектами и процессами с целью сбора количественной информации, ее обработки и первичного (разведочного) анализа перед использованием для принятия решений. В этой связи рассмотрены актуальные вопросы математического и инструментального обеспечения обработки и разведочного анализа числовых массивов в условиях эксплуатирующих организаций и при обучении в образовательных учреждениях гражданской авиации.

Ключевые слова: выборка, бинарный ряд, обработка данных, фильтрация, разведочный анализ, стационарность значений выборки, анализ на выброс, статистические критерии.

V. Z. Chokoj

Irkutsk, Russia

DATA PROCESSING AND EXPLORATORY DATA ANALYSIS

Aircraft maintenance procedures imply doing different work of logistic, organizational and technical and technological character. Their common issues are those of surveillance objects and processes **f**or the purpose of numerical data collection, processing and exploratory data analysis before using for taking decisions. The urgent

issues of mathematical and instrumental support for processing and exploratory data analysis within the operating company and under teaching in educational establishments of civil aviation are discussed in this regard.

Key words: sample, binary sequence, data processing, filtering, exploratory analysis, stationarity of sample values, outlier analysis, statistical criteria.

Необходимые объемы и требуемое качество (достоверность) исходных данных во многом являются определяющими при управлении сложными организационно-техническими системами гражданской авиации, где цена ошибок управления чрезвычайно велика. Это обусловливает важность математического, методического и программного обеспечения вопросов первичной обработки и анализа числовых данных. При этом важно, чтобы рекомендуемые средства обработки и анализа соответствовали реальным условиям и возможностям эксплуатирующих организаций и образовательных учреждений.

К классу задач первичной обработки результатов наблюдений и к предварительному анализу результатов наблюдений можно отнести, прежде всего, разведочный анализ, фильтрацию, анализ стационарности данных, анализ динамики нестационарных процессов.

Разведочный анализ предполагает выявление в данных значимо отличающихся значений, называемых выбросами, и близких к ним значений, называемых маргиналами. Такие значения, если они случайны, могут в последующем существенно исказить результаты использования накопленных данных, например, при формировании прогнозных моделей. В этой связи выявленные аномалии, после дополнительного анализа, должны быть своевременно устранены.

Разведочный анализ обычно сводится к проверке гипотезы о статистической однородности выборки после добавления к ней нового значения x_n, не укладывающегося в текущий размах значений. С этой целью используются различные алгоритмы, инвариантные к аномальным максимуму и минимуму, и основанные на критериях Диксона, ***r***-статистики, Граббса, трех сигм, Ирвина, Романовского, Кокрена и других.

Алгоритм с критерием Диксона. После ранжирования исходной выборки в порядке не убывания рассчитывают фактическое значение критерия:

$d_n = \frac{x_n - x_{n-1}}{x_n - x_1}$ – при подозрении x_n на аномальный максимум;

$d_1 = \frac{x_2 - x_1}{x_n - x_1}$ – при подозрении x_1 на аномальный минимум.

Далее, задавшись уровнем значимости α, определяют табулированное (например, в [Айвазян, 1998]) пороговое значение критерия $d_{\alpha, n}$. На завершающем этапе сравнивают фактическое и пороговое значения критерия. Если $d_{\alpha, n} \leq d_n$ (или $d_{\alpha, n} \leq d_1$), то исследуемое значение x_n (или x_1) признается значимо аномальным, в противном случае – нормальным (то есть значимо не отличающимся от предшествующих значений выборки).

Алгоритм с *r*-статистикой. Вначале выполняют ранжирование выборки в порядке не убывания и расчет фактического значения *r*-статистики:

$r_n = \frac{x_n - \bar{x}}{D_x \cdot \sqrt{\frac{n-1}{n}}}$ – при подозрении x_n на аномальный максимум;

$r_1 = \frac{\bar{x} - x_1}{D_x \cdot \sqrt{\frac{n-1}{n}}}$ – при подозрении x_1 на аномальный минимум,

где: $\bar{x}$ – среднее по выборке (с учетом анализируемого значения);

D_x – дисперсия выборки (с учетом анализируемого значения).

Далее, задавшись уровнем значимости α, и рассчитав число степеней свободы $f = n–2$, определяют табулированное (например, в [Айвазян, 1998]) пороговое значение критерия $r_{\alpha, f}$. На завершающем этапе сравнивают фактическое и пороговое значения критерия. Если $r_{\alpha, f} \leq r_n$ (или $r_{\alpha, f} \leq r_1$), то исследуемое значение x_n (или x_1) признается значимо аномальным, в противном случае – нормальным.

Алгоритм с критерием Граббса. Вначале выполняют ранжирование выборки в порядке не убывания и расчет фактического значения критерия:

$g_n = \frac{x_n - \bar{x}}{D_x}$ – при подозрении x_n на аномальный максимум; $g_1 = \frac{\bar{x} - x_1}{D_x}$ – при подозрении x_1 на аномальный минимум.

Далее, задавшись уровнем значимости α, определяют табулированное (например, в [Айвазян, 1998]) пороговое значение критерия $g_{\alpha,n}$. На завершающем этапе сравнивают фактическое и пороговое значения критерия. Если $g_{\alpha,n} \leq g_n$ (или $g_{\alpha,n} \leq g_1$), то исследуемое значение x_n (или x_1) признается значимо аномальным, в противном случае – нормальным.

Алгоритм с критерием «трех сигм». Используется для нормально и квазинормально распределенных выборок. Вначале рассчитываются среднее $\overline{x}$ и СКО σ_x. Далее проверяется выполнение условия $|\overline{x} - x_n| > 3 \cdot \sigma_x$. Если условие выполняется, то x_n признается аномальным, в противном случае – нормальным. Часто условие «трех сигм» считают чрезмерно жестким, приводящим к излишней отбраковке значений исходной выборки. В этой связи используют более лояльные условия:

- при $6 < n \leq 100$ используют условие $|\overline{x} - x_n| > 4 \cdot \sigma_x$;
- при $100 < n \leq 1000$ используют условие $|\overline{x} - x_n| > 4{,}5 \cdot \sigma_x$;
- при $1000 < n \leq 10000$ используют условие $|\overline{x} - x_n| > 5 \cdot \sigma_x$.

Алгоритм с критерием Ирвина. Вначале выполняют ранжирование выборки в порядке не убывания, оценку СКО σ_x и расчет фактического значения критерия:

$\eta_n = \frac{x_n - x_{n-1}}{\sigma_x}$ – при подозрении x_n на аномальный максимум;

$\eta_1 = \frac{x_2 - x_1}{\sigma_x}$ – при подозрении x_1 на аномальный минимум.

Далее, задавшись уровнем значимости α, определяют пороговое значение критерия:

$\eta_{\alpha,n} = 2 \cdot \sqrt{n}$+0,6 – при $\alpha = 0{,}1$;

$\eta_{\alpha,n} = 2{,}5 \cdot \sqrt{n} + 0{,}75$ – при $\alpha = 0{,}05$;

$\eta_{\alpha,n} = 3 \cdot \sqrt{n}$+1,15 – при $\alpha = 0{,}01$.

На завершающем этапе сравнивают фактическое и пороговое значения критерия. Если $\eta_{\alpha,n} \leq \eta_n$ (или $\eta_{\alpha,n} \leq \eta_1$), то исследуемое значение x_n (или x_1) признается значимо аномальным, в противном случае – нормальным.

Алгоритм с критерием Романовского. Часто используют для коротких выборок (n< 20). Вначале выполняют ранжирование выборки в порядке не убывания, оценку СКО σ_x и расчет фактического значения критерия:

$\beta_n = \frac{x_n - \bar{x}}{\sigma_x}$ – при подозрении x_n на аномальный максимум;

$\beta_1 = \frac{\bar{x} - x_1}{\sigma_x}$ – при подозрении x_1 на аномальный минимум.

Далее, задавшись уровнем значимости α, определяют пороговое значение критерия:

$\beta_{\alpha,n} = 0{,}571 \cdot \ln(n-1) + 0{,}951$ – при $\alpha = 0{,}1$;

$\beta_{\alpha,n} = 0{,}651 \cdot \ln(n-1) + 0{,}883$ – при $\alpha = 0{,}05$;

$\beta_{\alpha,n} = 0{,}837 \cdot \ln(n-1) + 0{,}642$ – при $\alpha = 0{,}01$.

На завершающем этапе сравнивают фактическое и пороговое значения критерия. Если $\beta_{\alpha,n} \le \beta_n$ (или $\beta_{\alpha,n} \le \beta_1$), то исследуемое значение x_n (или x_1) признается значимо аномальным, в противном случае – нормальным.

Фильтрация (элиминирование) данных выполняется в ситуациях, когда текущая выборка мала, и/или наблюдаемые значения имеют недостаточную достоверность. Формально получение элиминированного значения x^*_i $(i = \overline{1,n})$ сводится к корректировке наблюдаемого значения x_i с учётом смежных значений $\{\ldots, x_{i-1}, x_i, x_{i+1}, \ldots\}$. Один из простых вариантов элиминирования может выполняться по алгоритму «нелинейного сглаживания по 7 точкам», предполагающему вычисление последовательности:

$$x_1^* = \frac{39 \cdot x_1 + 8 \cdot x_2 - 4 \cdot x_3 - 4 \cdot x_4 + \cdot x_5 + 4 \cdot x_6 - 2 \cdot x_7}{42};$$

$$x_2^* = \frac{8 \cdot x_1 + 19 \cdot x_2 + 16 \cdot x_3 + 6 \cdot x_4 - 4 \cdot x_5 - 7 \cdot x_6 + 4 \cdot x_7}{42};$$

$$x_3^* = \frac{-4 \cdot x_1 + 16 \cdot x_2 + 19 \cdot x_3 + 12 \cdot x_4 + 2 \cdot x_5 - 4 \cdot x_6 + x_7}{42};$$

$$x_i^* = \frac{7 \cdot x_i + 6 \cdot (x_{i+1} + x_{i-1}) + 3 \cdot (x_{i+2} + x_{i-2}) - 2 \cdot (x_{i+3} + x_{i-3})}{21},$$

$$4 \le i \le n - 3;$$

$$x_{n-2}^* = \frac{x_{n-6} - 4 \cdot x_{n-5} + 2 \cdot x_{n-4} + 12 \cdot x_{n-3} + 19 \cdot x_{n-2} + 16 \cdot x_{n-1} - 4 \cdot x_n}{42};$$

$$x_{n-1}^{*} = \frac{4 \cdot x_{n-6} - 7 \cdot x_{n-5} - 4 \cdot x_{n-4} + 6 \cdot x_{n-3} + 16 \cdot x_{n-2} + 19 \cdot x_{n-1} + 8 \cdot x_n}{42};$$

$$x_n^{*} = \frac{-2 \cdot x_{n-6} + 4 \cdot x_{n-5} + x_{n-4} - 4 \cdot x_{n-3} - 4 \cdot x_{n-2} + 8 \cdot x_{n-1} + 39 \cdot x_n}{42}.$$

Анализ стационарности данных обычно выполняется с целью своевременного обнаружения в наблюдаемых процессах восходящего или нисходящего тренда. Для оценки стационарности (нестационарности) значений в выборках используют различные алгоритмы на базе тех или иных критериев, например, серий, модифицированного критерия серий, поворотных точек, Аббе, Стьюдента, Фишера.

Алгоритм с критерием серий. Вначале на основе исходной выборки $\{x_i, i = \overline{1,n}\}$ формируется бинарный ряд $\{y_i\}$ по правилу:

$$y_i = \begin{cases} 1, \text{если } x_i \geq x_m; \\ 0, \text{если } x_i < x_m. \end{cases}$$ где x_m – медиана исходной выборки.

Далее в бинарном ряде подсчитывают число однородных серий ν, задаются уровнем значимости α и определяют табулированные (например, в [Айвазян, 1998]) пороговые значения критерия $\nu_{n,1-\alpha/2}$ и $\nu_{n,\alpha/2}$. Если ν удовлетворяет условию $\nu_{n,1-\alpha/2} \leq \nu \leq \nu_{n,\alpha/2}$, то гипотеза о стационарности исходной выборки принимается, в противном случае – отвергается.

Алгоритм с модифицированным критерием серий. Здесь, помимо числа серий ν, дополнительно фиксируется длина самой протяженной серии τ. Гипотеза о стационарности исходной выборки принимается, если одновременно выполняются два условия:

$$\nu \geq 0{,}5 \cdot \left(n + 1 - U_\alpha \cdot \sqrt{n-1}\right);$$

$$\tau \leq 3{,}3 \cdot \log_{10}(n+1),$$

где U_α – квантиль нормального распределения при уровне значимости α.

Алгоритм с критерием поворотных точек исследует бинарный ряд, образованный по правилу:

$$y_i = \begin{cases} 1, \text{если } x_{i+1} - x_i \geq 0; \\ 0, \text{если } x_{i+1} - x_i < 0. \end{cases}$$

Статистики ν и τ вычисляют аналогично предыдущему алгоритму, но для стационарной исходной выборки эти статистики должны удовлетворять двум условиям:

$$\nu > 0{,}33 \cdot (2 \cdot n - 1) - U_\alpha \cdot \sqrt{\frac{18 \cdot n - 29}{90}}; \qquad \tau > \tau_0,$$

где $\tau_0 = \begin{cases} 5, \text{если } n \le 26; \\ 6, \text{если } 26 < n \le 153; \\ 7, \text{если } n > 153 \end{cases}$.

Алгоритм с критерием Аббе. Вначале рассчитывают фактическое значение критерия

$$\gamma_\text{ф} = \frac{q}{D_x},$$

где: $q = \frac{0{,}5 \cdot \sum_{i=1}^{n-1} (x_i - x_{i+1})^2}{n-1}$;

D_x – дисперсия значений исходной выборки.

Далее, задавшись уровнем значимости α, определяют пороговое значение критерия $\gamma_{n,\alpha}$. Для $n \le 59$ пороговые значения табулированы (например, в [Айвазян, 1998]), а при $n \ge 60$ пороговое значение критерия определяют по формуле

$$\gamma_{n,\alpha} = 1 + \frac{U_\alpha}{\sqrt{n + 0{,}5 \cdot (1 + U_\alpha^2)}}.$$

Если $\gamma_\text{ф} \le \gamma_{n,\alpha}$, то гипотеза о стационарности исходной выборки принимается, в противном случае – отвергается.

Алгоритм с *t*-критерием Стьюдента. Вначале исходную выборку разбивают на два участка, в каждом из которых по n значений. Для каждого участка вычисляют дисперсии (D_1, D_2), после чего рассчитывают фактическое значение критерия

$$t_\text{ф} = \frac{|D_1, - D_2|}{\sqrt{\frac{D_1, + D_2}{n}}}.$$

Далее, задавшись доверительной вероятностью p и скорректировав ее по формуле $p' = \frac{1+p}{2}$, рассчитывают число степеней свободы $\nu = (n+1) \cdot \left(1 + \frac{2 \cdot m}{m^2 + 1}\right)$,

где $m = \frac{max\{D_1; D_2\}}{min\{D_1; D_2\}}$, после чего определяют табулированное (например, в [1]) пороговое значение критерия $t_{v,p}$. На завершающем этапе сравнивают фактическое и пороговое значения критерия, если условие $t_{\phi} \leq t_{v,p}$ выполняется, то исходная (объединенная) выборка признается стационарной, в противном случае – нестационарной.

Информацию о характере, месте возникновения и величине нестационарности исходной выборки обеспечивают дополнительные математические подходы, например, аппарат анализа процессов по первой и второй производным.

Функциональность и интерфейсные решения по инструментам обработки и анализа данных. В соответствии с рассмотренными выкладками на факультете Эксплуатации летательных аппаратов Иркутского филиала МГТУ ГА в последние годы сформирован пакет инструментов обработки и разведочного анализа числовых массивов данных. Эти инструменты включены в рассчетно-информационный пакет Модельер 2.1 и для пользователей доступны через группу «Решения при риске» головного меню (*рис. 1*).

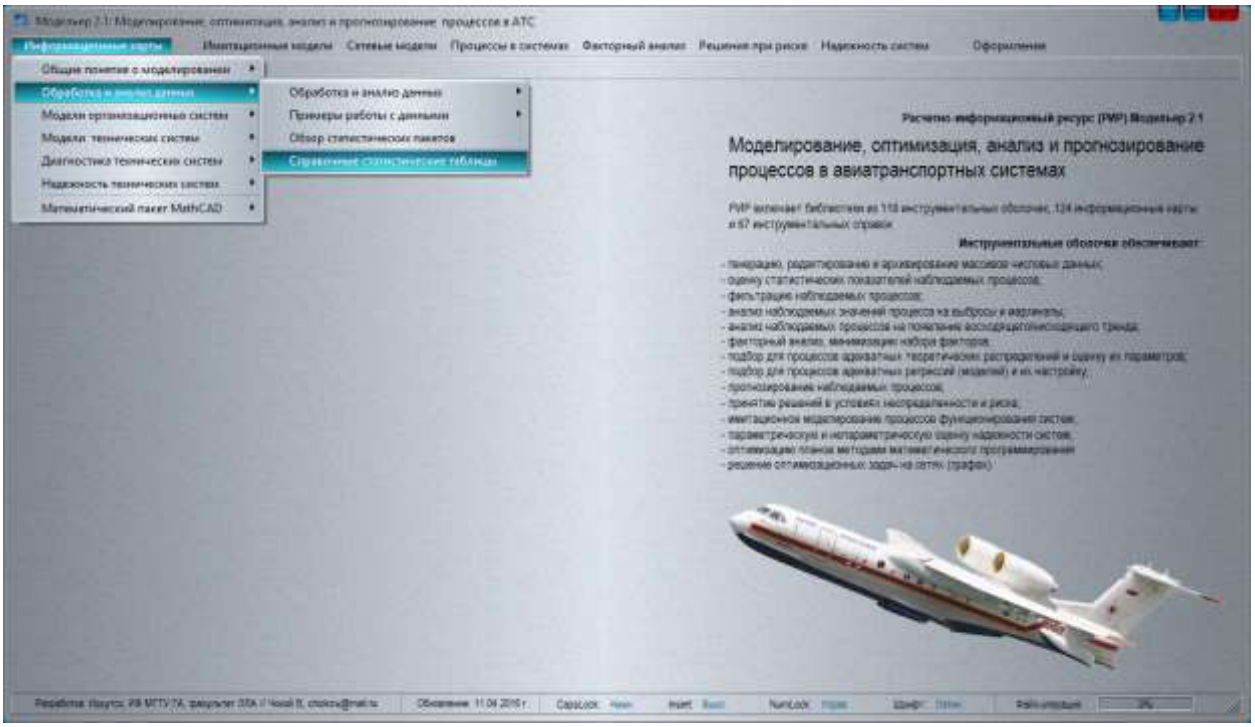

Рис. 1. Головная панель пакета Модельер 2.1 (раскрыта группа «Информационные карты» головного меню)

В частности, в пакет включены следующие инструменты обработки и разведочного анализа:

– фильтрация (элиминирование) данных (*рис. 2*);

– анализ процесса на стационарность (*рис. 3*);

– анализ процесса по производным (*рис.* 4);

– анализ процесса на выбросы (*рис. 5*).

Основные интерфейсные решения по перечисленным инструментам, и в целом по пакету, соответствуют требованиям действующих стандартов и норм CALS-технологий для компьютерных ресурсов [Ганьшин, 1993]. Пакет представляет собой полнофункциональное windows-приложение, функционирующее на типовых IBM-подобных ЭВМ с операционной системой Windows-xx. Для инсталляции пакета на жестком диске достаточно 1,8 Гб памяти. При необходимости каждая из инструментальных оболочек пакета может изыматься из состава пакета и использоваться как автономное windows-приложение.

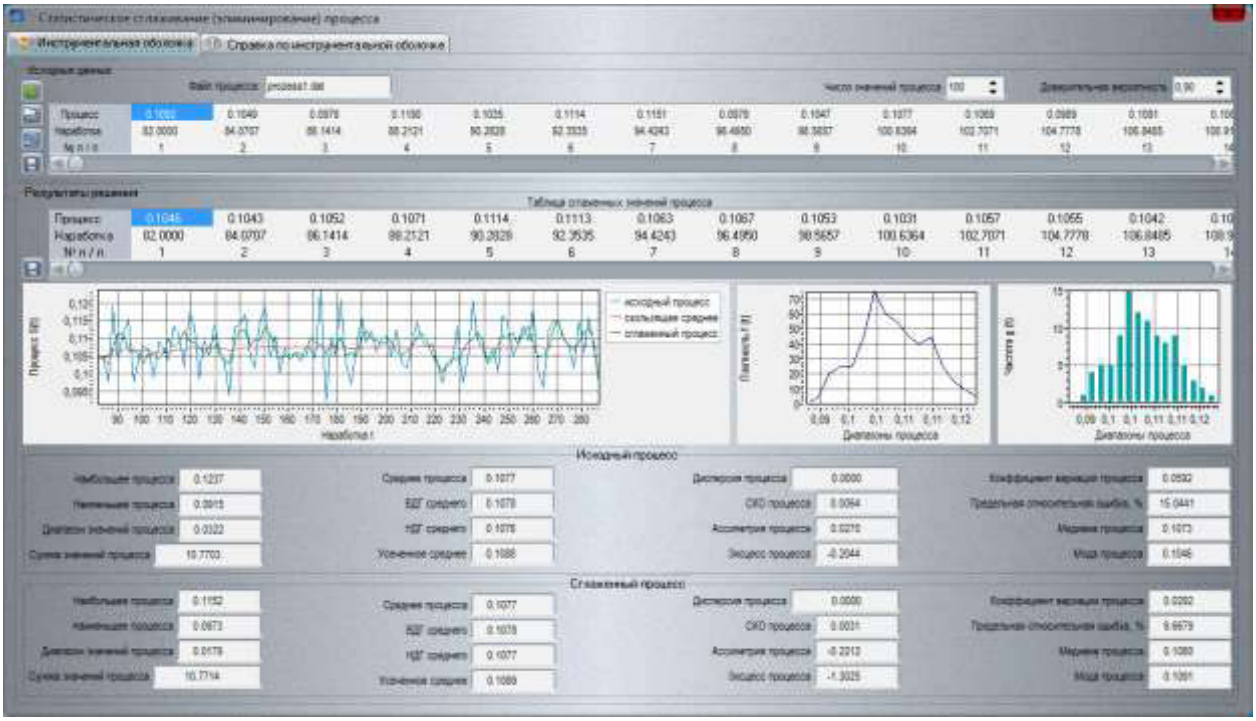

Рис. 2. Панель инструмента «Фильтрация (элиминирование) данных»

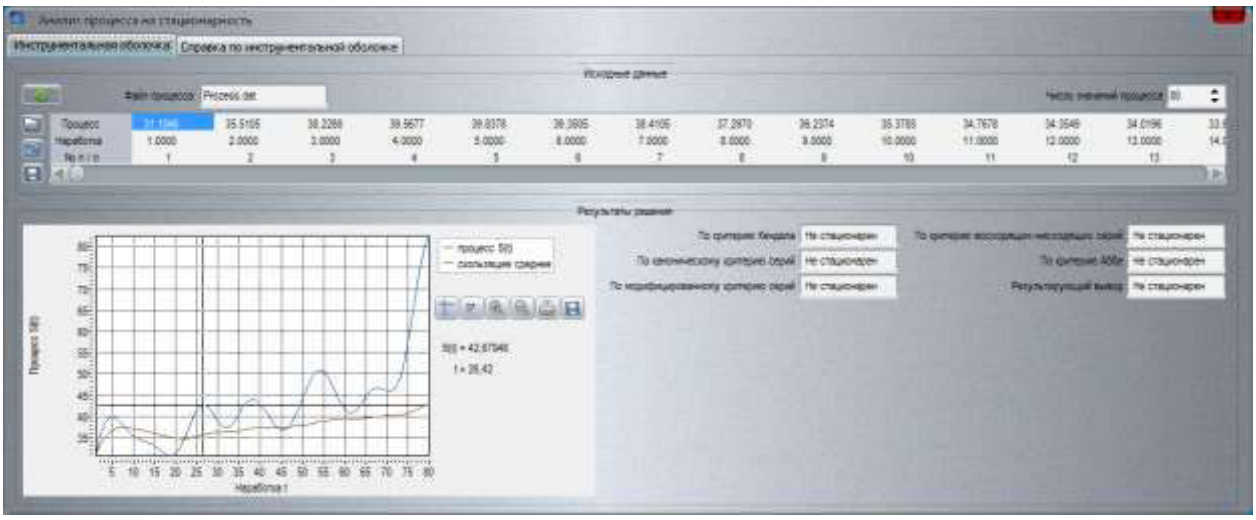

Рис. 3. Панель инструмента «Анализ процесса на стационарность»

Особенностями пакета являются:

- наличие справочной системы, как по общим вопросам охваченной пакетом предметной области, так и по частным вопросам, относящимся к конкретным инструментам;

- возможность выбора пользователем комфортного дизайна экранных панелей;

- наличие всплывающих подсказок по назначению кнопок управления, а также по формату и по допустимому диапазону вводимых числовых исходных данных;

- использование в справке и в наименованиях полей для данных терминов, доступных пользователям без углубленной математической подготовки.

Рис. 4. Панель инструмента «Анализ процесса по производным»

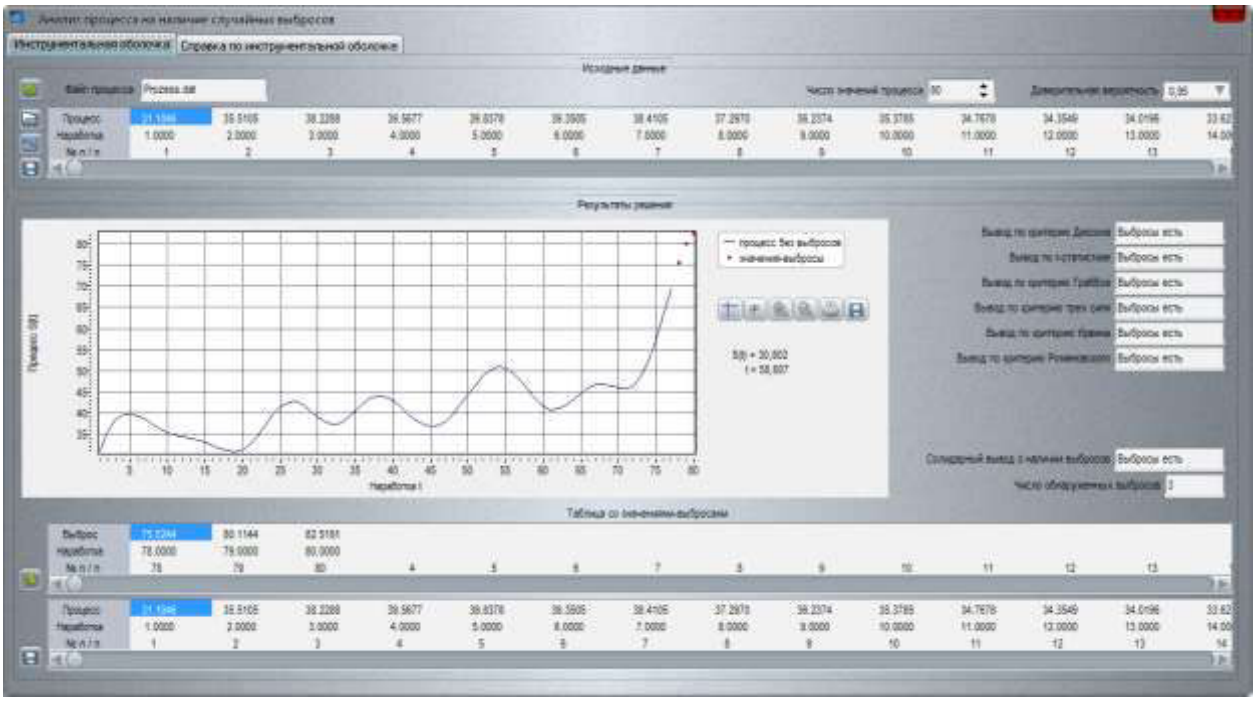

Рис. 5. Панель инструмента «Анализ процесса на выбросы»

Представленные инструменты по обработке и предварительному анализу массивов числовых данных используются при изучении ряда учебных дисциплин, например, «Методы и алгоритмы обработки статистических данных», «Прикладные методы вычислений». При этом, как правило, они используются не самостоятельно, а в связке с другими (основными) инструментами (например,

моделирования, регрессионного анализа, прогнозирования, факторного анализа, принятия решений). Такое комплексирование позволяет повысить корректность итоговых результатов, получаемых с помощью основных инструментов пакета.

Библиографический список

1. *Айвазян С. А*. Прикладная статистика и основы эконометрики / С. А. Айвазян, В. С. Мхитарян. М.: ЮНИТИ, 1998. 1022 с.

2. *Ганьшин В. Н.* Применение методов математической статистики в авиационной практике / В. Н. Ганьшин и др. М.: Транспорт, 1993. 211 с.

3. Информационная поддержка жизненного цикла изделий. Информационные материалы. НИЦ CALS-технологий «Прикладная логистика» // [Электронный ресурс]. – 2008. URL: http://www.cals.ru (дата обращения: 10.04.2017).

4. *Крянев А. В.* Математические методы обработки неопределенных данных / А. В. Крянев, Г. В. Лукин. М.: Физматлит, 2003. 311 с.

5. *Тьюки Дж.* Анализ результатов наблюдений. М.: Мир, 1981. 696 с.

References

1. *Ajvazyan S. A.* (1998). Applied Statistics and the Fundamentals of Econometrics / S. A. Ajvazyan, V. S. Mhitaryan. M.: YUNITI, 1998. 1022 p. (in Russian).

2. *Gan'shin V.N.* (1993). The Use of the Methods of Mathematical Statistics in Aviation / V.N. Gan'shin, etc. M.: Transport, 1993. 211 p. (in Russian).

3. The Information Support of the Product Lifecycle. Informational Materials. CALS R&D Center «Applied Logistics» // [Electronic resource]. – 2008. URL: http://www.cals.ru [accessed 10 April 2017] (in Russian)

4. *Kryanev A. V. (2003).* Mathematical Methods of Ambiguous Data Processing / *A. V Kryanev, G. V. Lukin*. M.: Fizmatlit, 2003. 311 p. (in Russian).

5. *Tukey J*. (1981). Exploratory Data Analysis. M.: Mir, 1981. 696 p. (in Russian).