УДК 519.233.5:519.242.5

С.Г. Радченко, д.т.н., доц.

МЕТОДЫ НАИЛУЧШЕГО РЕШЕНИЯ РЕГРЕССИОННЫХ ЗАДАЧ

Национальный технический университет Украины «КПИ»
E-mail: lapach@ukr.net

Рассмотрены новые методы и алгоритмы решения многофакторных регрессионных задач. Показано, что их использование позволяет строить планы экспериментов, которые не приведены в каталогах, формализованно определять структуру моделей, неизвестную исследователю заранее, решать некорректно поставленные задачи с получением статистических моделей, имеющих наилучшие из возможных критерии качества.

Ключевые слова: некорректно поставленные задачи, планирование экспериментов, регрессионный анализ, устойчивое решение регрессионных задач.

Розглянуто нові методи й алгоритми розв'язання багатофакторних регресійних задач. Показано, що їх використання дозволяє побудувати плани експериментів, які не наведено в каталогах, формалізовано визначити структуру моделей, невідому досліднику заздалегідь, розв'язувати некоректно поставлені задачі з отриманням статистичних моделей, що мають найкращі з можливих критерії якості.

Ключові слова: некоректно поставлені задачі, планування експериментів, регресійний аналіз, стійке розв'язання регресійних задач.

Постановка проблемы

Создание новой техники и технологий, повышение точности и надежности средств измерений требуют формализованной информации для получения наилучших результатов.

Сложность и системность решаемых прикладных задач не всегда позволяет использовать теоретико-аналитический подход. Тогда необходимо применять экспериментально-статистический подход.

Формализованная информация предоставляется исследователем в виде математических моделей, необходимых для принятия решений, изучения, управления и оптимизации объекта исследования.

Получение многофакторных статистических моделей, линейных по параметрам и нелинейных по факторам, часто сводится к решению некорректно поставленных обратных задач, требующих разработки специальных методов их решений [1].

Анализ исследований и публикаций

Для решения некорректно поставленных задач было предложено использовать метод регуляризации.

Значительные работы проведены академиками А.Н. Тихоновым и М.М. Лаврентьевым, их учениками и последователями [2; 3, с. 161–171; 4, с. 47–72].

А. Хёрл использовал регуляризацию в решении регрессионных задач и обобщил ее в работе [5]. Метод получил название гребневой регрессии (ridge regression).

Регуляризация приводит к систематическим ошибкам, т. е. к смещению коэффициентов b_i уравнения регрессии, однако уменьшает среднеквадратические ошибки их оценок.

При практическом использовании регуляризации конкретный выбор значения параметра регуляризации r затруднен.

В линейном относительно параметров регрессионном анализе регуляризация используется редко.

Для формализованного получения математических моделей можно использовать метод группового учета аргументов (МГУА), разработанный академиком Национальной академии наук Украины А.Г. Ивахненко, его учениками и последователями [6; 7].

Авторы МГУА обращают внимание на некоторые проблемы при получении моделей:

1) слишком большое число поколений или рядов селекции модели ведет к вырождению, информационная матрица становится плохо обусловленной [7, с. 33];

2) при полном переборе вариантов модели можно решать сравнительно простые задачи [7, с. 33];

3) планирование эксперимента позволяет повышать точность и помехоустойчивость моделирования [7, с. 50].

Однако конкретные рекомендации по выбору планов экспериментов исследователями не приводятся.

Нерешенные проблемы

Анализ публикаций показал, что в области регрессионного анализа нерешенными проблемами являются следующие:

– не рассмотрено решение проблем и задач регрессионного анализа с позиций системного подхода;

– нет рекомендаций по формализованному получению устойчивых структур многофакторных статистических моделей;

– не приведены случаи устойчивого получения статистических моделей для нестандартных областей факторного пространства, отличающихся от куба, сферы, симплекса;

– нет концепции ортогональности эффектов уравнения регрессии как одного из основных направлений устойчивого оценивания регрессионных моделей;

– не опубликовано последовательное построение многофакторных регулярных планов экспериментов;

– отсутствует литература о новых методах планирования эксперимента.

Цель работы – рассмотреть новые результаты в области множественного регрессионного анализа как технологию наилучшего решения регрессионных задач.

Основные подходы к решению регрессионных задач

Наилучшее решение регрессионных задач возможно при системном подходе в границах триединой проблемы [8]:

– устойчивое (робастное) планирование эксперимента;

– устойчивая структура статистической модели;

– устойчивое оценивание коэффициентов модели.

Под устойчивым (робастным) планом эксперимента понимается план полного или дробного факторного эксперимента, позволяющий выбрать неизвестные исследователю структуры «истинных» статистических моделей $\hat{y}_w$ полиномиального вида, линейных по параметрам, и получить адекватные модели (w – текущий номер определяемой модели, $1 \le w \le m$; m – общее число определяемых моделей по устойчивому плану эксперимента).

План эксперимента не изменяется для получаемых различных структур моделей.

К устойчивым планам экспериментов относятся планы полного факторного эксперимента:

– многофакторные регулярные планы;

– планы на основе $ЛП_\tau$ равномерно распределенных последовательностей.

Разработан новый подход к формализованному выбору устойчивых структур многофакторных статистических моделей, линейных относительно параметров и в общем случае нелинейных по факторам.

Устойчивая структура многофакторной статистической модели – это структура, которая характеризуется неизменностью множества главных эффектов и взаимодействий многофакторной статистической модели полиномиального вида при изменении значений результатов экспериментов (откликов), порождаемых случайными ошибками (погрешностями) результатов наблюдений, измерений, вычислений и неопределенностью искомой структуры модели.

Структурные элементы моделей выбираются из множества структурных элементов модели полного факторного эксперимента с ортогональными или слабо коррелированными (коэффициент парной корреляции $|r_{ij}| < 0{,}3$) эффектами с использованием устойчивого (робастного) плана эксперимента.

Необходимо также обеспечить устойчивость коэффициентов модели.

Под устойчивостью коэффициентов статистической модели будем понимать минимально возможную изменчивость коэффициентов многофакторной статистической модели полиномиального вида к случайным ошибкам (погрешностям) результатов наблюдений, измерений и вычислений.

Для оценки устойчивости коэффициентов используется число обусловленности $\mathrm{cond}(\mathbf{X}^{\mathrm{T}}\mathbf{X})$:

– наилучшая устойчивость, если $\mathrm{cond}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) = 1$;

– хорошая устойчивость, если $1 < \mathrm{cond}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) \le 10$,

– удовлетворительная устойчивость, если $10 < \mathrm{cond}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) \le 100$,

– неудовлетворительная устойчивость, если $\mathrm{cond}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) > 100$.

Коэффициенты будут максимально устойчивыми, если их эффекты ортогональны друг к другу или близки к ортогональным и нормированы.

Структуру многофакторной статистической модели, неизвестной заранее исследователю, предложено выражать в виде множества эффектов:

$$\prod_{i=1}^{k} (1 + x_i^{(1)} + x_i^{(2)} + \ldots + x_i^{(s_i - 1)}) \rightarrow N_{\Pi},$$

где k – общее число факторов, $1 \le i \le k$;

1 – значение фиктивного фактора $x_0 \equiv 1$;

$x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(s_i - 1)}$ – ортогональные контрасты факторов X_i;

(1), (2), …, $(s_i - 1)$ – порядок контрастов фактора X_i;

s_i – число различных уровней для факторов X_i;

N_{Π} – число структурных элементов полного факторного эксперимента, равное числу опытов эксперимента.

Структура включает множество главных эффектов и множество взаимодействий главных эффектов, необходимых и достаточных для получения адекватных статистических моделей. Необходимость обосновывается теоремами Вейерштрасса, Стоуна, Джексона, а достаточность подтверждается многочисленным использованием структуры для адекватной аппроксимации различных полных и дробных факторных экспериментов.

Разработанный алгоритм RASTA3 позволяет выбирать структурные составляющие многофакторной статистической модели [9, с. 179–182].

Для получения моделей разработано программное средство «Планирование, регрессия и анализ моделей» (ПС ПРИАМ) [9, с. 45–47].

Для получения «наилучших» структур статистических моделей необходимо выполнение следующих условий [10]:

– статистическая независимость коэффициентов моделей;

– статистическая значимость коэффициентов;

– соответствие плана эксперимента устойчивому (робастному) плану;

– эффекты, введенные в модель, должны быть нормированы.

Методология теории планирования эксперимента разработана для областей факторного пространства [9, с. 197–199]:

– прямоугольный параллелепипед (куб);

– сфера;

– симплекс.

В нестандартных областях факторного пространства факторы и их эффекты (главные и взаимодействия) коррелированы между собой.

В качестве общего метода получения возможно наилучших статистических моделей для произвольных областей факторного пространства предложено использовать метод топологического отображения прообраза факторного пространства в образ (см. рисунок).

Разработаны два метода отображения прообраза в образ [11]:

– построение функции отображения прообраза в образ – алгоритмы RASTA4, RASTA4К [9, с. 211–258];

– установление собственной кодированной системы координат в области прообраза и в области образа – алгоритмы RASTA5.1, RASTA10 [12; 13].

Разработанное отображение факторного пространства сохраняет ортогональность факторов $X_{i\text{пр}}$, $X_{j\text{пр}}$ в прообразе факторного пространства и в образе факторного пространства при отображении точек прообраза в точки образа, т. е. координат уровней факторов прообраза в координаты уровней факторов образа, при условии введения в образе собственной кодированной системы координат.

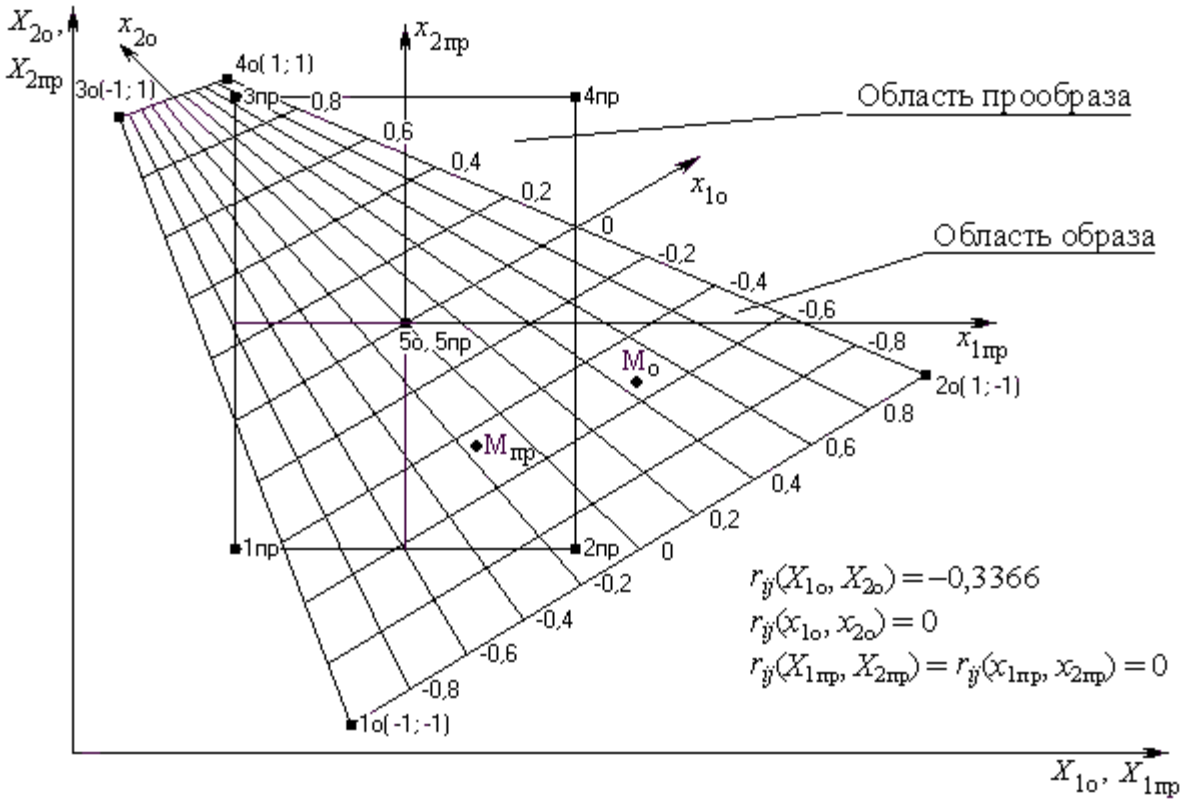
Свойство ортогональности факторов в прообразе и образе является инвариантным свойством преобразования, или топологическим инвариантом, а равенство нулю коэффициентов парной корреляции факторов в прообразе и образе инвариантным числом:

$r_{ij}\,(x_{i\text{пр}}, x_{j\text{пр}}) = r_{ij}\,(x_{i\text{о}}, x_{j\text{о}}) = 0;$

$1 \le i < j \le k.$

Оптимальные свойства статистических оценок коэффициентов уравнения регрессии, полученных в прообразе с использованием метода максимального правдоподобия, сохраняются для условий, полученных топологическим отображением прообраза в образ. Единственность оценок также выполняется.

Разработанные методы, алгоритмы и подходы устойчивого оценивания статистических моделей позволяют:



Системы натуральных и собственных кодированных координат областей образа и прообраза факторного пространства при линейном ограничении образа

– выбрать устойчивый план эксперимента;

– устойчиво оценивать структуру и коэффициенты многофакторных статистических моделей ($1 \le \mathrm{cond}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) < 10$) в условиях высокой мультиколлинеарности факторов ($0{,}6 < < |r_{ij}(X_{io}, X_{jo})| < 1$).

Возможно получение устойчивого решения некорректно поставленных задач построения многофакторных статистических моделей в различных нестандартных областях факторного пространства.

Концепция ортогональности включает следующие основные положения.

1. В полном факторном эксперименте любые эффекты ортогональны друг к другу (теорема Бродского) [14, с. 26–29]. Если эффекты факторов и взаимодействий выразить в виде системы ортогональных нормированных контрастов, т. е.

$$\sum_{u=1}^{N} x_{iu}^{(p)} = 0;$$

$$\sum_{u=1}^{N} x_{iu}^{(p)} \times x_{ju}^{(q)} = 0;$$

$$\sum_{u=1}^{N} [x_{iu}^{(p)}]^2 = N;$$

$$\sum_{u=1}^{N} [x_{iu}^{(p)} \times x_{ju}^{(q)}]^2 = N,$$

то матрица дисперсий-ковариаций примет вид:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\sigma^2(\varepsilon) = (1/N)\mathbf{E}\sigma^2(\varepsilon),$$

где $x_{iu}^{(p)}$ – значение p-го ортогонального контраста i-го фактора для u-й строки матрицы планирования, $1 \le u \le N$, $1 \le p \le s_i - 1$;

$x_{ju}^{(q)}$ – значение q-го ортогонального контраста j-го фактора для u-й строки матрицы планирования, $1 \le q \le s_j - 1$, $1 \le i < j \le k$;

$\mathbf{X}$ – матрица эффектов полного факторного эксперимента;

$\sigma^2(\varepsilon)$ – теоретическое значение дисперсии воспроизводимости результатов опытов;

N – число опытов в плане эксперимента;

$\mathbf{E}$ – единичная матрица.

Ортогональный многофакторный план эксперимента является максимально устойчивым (робастным) при получении неизвестной ранее исследователю структуры модели и оценке ее коэффициентов.

2. Для дробного факторного эксперимента необходимо использовать многофакторные регулярные планы экспериментов. В этих планах все главные эффекты ортогональны друг к другу. В планах на основе ЛПτ равномерно распределенных последовательностей эффекты близки к ортогональным [9, с. 166–169].

3. Главные эффекты в статистических моделях необходимо представлять в виде ортогональных контрастов, которые следует нормировать.

4. В тех случаях, когда условия решения прикладной задачи не позволяют представить эффекты ортогонально друг к другу, необходимо использовать метод топологического отображения прообраза факторного пространства в образ и другие подходы [9, с. 199, 204, 211–258].

Разработаны последовательные регулярные планы экспериментов с хорошими статистическими свойствами [9, с. 148–150, 429–431]:

$3^4//9 \rightarrow 3^4//27$;

$3^4//18 \rightarrow 3^4//27 \rightarrow 3^4//54$;

$5^6//25 \rightarrow 5^6//50$;

$4^8//32 \rightarrow 4^8//64$.

Разработаны алгоритмы, позволяющие получать планы с хорошими статистическими свойствами:

– алгоритм генерирования квазиортогональных планов экспериментов с ограничением числа проводимых опытов и минимально возможной коррелированностью столбцов эффектов RASTA1;

– алгоритм генерирования квази-D-оптимальных планов экспериментов RASTA2 [9, с. 150–152];

– алгоритм генерирования квазирегулярных и квазиравномерных планов экспериментов RASTA8.

Разработанные методы наилучшего решения регрессионных задач были использованы при решении более ста прикладных задач и показали хорошие результаты.

В работе [15] рассмотрены математическое моделирование и компромиссная оптимизация технологического процесса электроэрозионной прошивки отверстий в стали 1Х12СЮ.

Критериями качества (откликами) были выбраны:

y_1 – производительность обработки $П$:

$П$ = max;

y_2 – износ электроэрозионного инструмента J:

J = min.

В результате анализа априорной информации и потребностей производства исследовали влияние на критерии качества следующих факторов:

X_1 – давление прокачиваемой жидкости P_E, уровни фактора 0,5; 1,0; 1,5;

X_2 – рабочий ток в межэлектродном зазоре I_E, уровни 18; 22; 26;

X_3 – частота импульсов f_E, уровни 2; 3; 4;

X_4 – напряжение на вибраторе U_E, уровни 50; 70; 90;

X_5 – напряжение на двигателе подачи электрода-инструмента W_E, уровни 60; 80; 100.

В качестве плана эксперимента целесообразно использовать регулярный равномерный план $3^5//27$.

Структуры математических моделей выбирались из структурного множества эффектов модели полного факторного эксперимента:

$$(1 + x_1 + z_1)(1 + x_2 + z_2)\times\ldots\times(1 + x_5 + z_5) \rightarrow N_П,$$

где $x_1, x_2, \ldots, x_5$ и $z_1, z_2, \ldots, z_5$ – соответственно линейные и квадратичные ортогональные контрасты факторов X_1, X_2, …, X_5.

По плану эксперимента $3^5//27$ и значениям уровней факторов была построена рабочая матрица. По условиям каждой строки рабочей матрицы рандомизированно проведено по два повторных опыта. Получение математических моделей и их статистический анализ были проведены с использованием ПС ПРИАМ:

$$\hat{y}_1 = 88,80 + 27x_2 + 10,56x_3 + 4,85z_2z_3 - 3,15z_2 + 2,28x_1 + 2,89z_2x_3 - 1,70z_3;$$

$$\hat{y}_2 = 54,89 + 5,38x_1 + 4,31x_3 + 3x_3x_4z_5 + 2,75z_1x_2z_3 + 2,56z_1x_2z_5 + 3,29z_1z_2z_4,$$

где

$x_1 = 2(X_1 - 1)$;

$z_1 = 1,5(x_1^2 - 0,666667)$;

$x_2 = 0,25(X_2 - 22)$;

$z_2 = 1,5(x_2^2 - 0,666667)$;

$x_3 = X_3 - 3$;

$z_3 = 1,5(x_3^2 - 0,666667)$;

$x_4 = 0,05(X_4 - 70)$;

$z_4 = 1,5(x_4^2 - 0,666667)$;

$x_5 = 0,05(X_5 - 80)$;

$z_5 = 1,5(x_5^2 - 0,666667)$.

Формирование структур моделей $\hat{y}_1$ и $\hat{y}_2$ по алгоритму RASTA3 производилось выбором из всех главных эффектов и взаимодействий по два и три эффекта с соблюдением условия $|r_{ij}| < 0,4$.

Выбранные эффекты в модели $\hat{y}_1$ все ортогональны друг к другу ($r_{ij} = 0$), в модели $\hat{y}_2$ – только $|r_{ij}(x_1, z_1x_2z_5)| = 0,2887$, для остальных эффектов $r_{ij} = 0$.

Построенные модели адекватны, высокоинформативны, максимально устойчивы. Подтверждается устойчивость их структуры и оценок коэффициентов. Эти модели были использованы для анализа влияния различных факторов на изучаемые критерии качества (отклики) y_1 и y_2.

По моделям $\hat{y}_1$ и $\hat{y}_2$ была проведена многокритериальная компромиссная оптимизация по Парето, получены значения факторов $X_1 \ldots X_5$, дающих оптимальные значения критериев качества.

Выводы

1. Впервые разработана и исследована система устойчивого решения многофакторных регрессионных задач в условиях исходной априорной неопределенности и мультиколлинеарности факторов.

2. Выбор плана эксперимента на основе концепции ортогональности нормированных эффектов позволяет установить истинную структуру модели, неизвестную заранее исследователю, выбором статистически значимых и ортогональных эффектов из множества эффектов модели полного факторного эксперимента.

3. Разработаны топологические методы устойчивого оценивания статистических моделей для произвольных областей факторного пространства с использованием отображения прообраза факторного пространства в образ:

– получением математических функций отображения прообраза в образ;

– установлением собственных кодированных систем координат в области прообраза и в области образа, топологически эквивалентных (гомеоморфных) между собой;

– планированием эксперимента с использованием фиктивных факторов.

4. Разработан инвариантно-групповой подход в теории планирования эксперимента, позволивший устойчиво оценивать коэффициенты модели в области прообраза, где можно наилучшим образом планировать эксперимент, тогда как в области образа в исходной системе координат факторы могут быть статистически сильно взаимосвязаны.

5. Разработанные алгоритмы генерирования планов экспериментов RASTA1, RASTA2, RASTA8 позволяют получать планы, не приведенные в известных каталогах, с хорошими статистическими свойствами и решать разнообразные ранее нерешаемые задачи.

С методами решения регрессионных задач и полученными результатами можно ознакомиться в работе [16].

Литература

1. *Тихонов А. Н.* [Выступление на годичном общем собрании Академии наук СССР] / А. Н. Тихонов // Вестник Академии наук СССР. – 1989. – № 2. – С. 94–95.

2. *Тихонов А.Н.* Методы решения некорректных задач: учеб. пособие для вузов. – 3-е изд., испр. / А. Н. Тихонов , В. Я. Арсенин.– М.: Наука, 1986. – 299 с.

3. *Тихонов А. Н.* Об обратных задачах / А. Н. Тихонов // Некорректные задачи математической физики и анализа. – Новосибирск: Наука, 1984. – 264 с.

4. *Жуковский Е .Л.* Статистическая регуляризация решений обратных некорректно поставленных задач обработки и интерпретации результатов эксперимента / Е. Л. Жуковский // Методы математического моделирования, автоматизация обработки наблюдений и их применения: сб. / под ред. А. Н. Тихонова, А. А. Самарского. – М., 1986. – С. 47–72.

5. *Hoerl, A. E.; Kennard, R. W.* 1970. Ridge regression: biased estimation for non-orthogonal problems. – Technometrics. Vol. 12: P. 55–67.

6. *Ивахненко А. Г.* Моделирование сложных систем: информационный подход / А. Г. Ивахненко. – К.: Вища шк., Голов. изд-во, 1987. – 63 с.

7. *Ивахненко А. Г.* Самоорганизация прогнозирующих моделей / А. Г. Ивахненко, Й. А. Мюллер. – К. : Техніка, 1985 ; Берлин : ФЕБ Ферлаг Техник, 1984. – 223 с.

8. *Радченко С. Г.* Формализация постановки многофакторного экспериментального исследования / С. Г. Радченко // Математичні машини і системи. – 2011. – № 1. – С. 96–102.

9. *Радченко С. Г.* Устойчивые методы оценивания статистических моделей: моногр. / С. Г. Радченко . – К.: ПП «Санспарель», 2005. – 504 с.

10. *Радченко С .Г.* Стійке оцінювання статистичних моделей технічних систем : Автореф. дис. на здобуття наук. ступеня д-ра техн. наук / С. Г. Радченко . – К.: НТУУ «КПІ», 2009. – 35 с.

11. *Радченко С. Г.* Устойчивые методы оценивания статистических моделей / С. Г. Радченко // Матеріали IX Міжнар. наук. конф. ім. акад. М. Кравчука, К., 16–19 трав. 2002 р. – К., 2002. – С. 451–452.

12. *Радченко С. Г.* Стійке оцінювання статистичних моделей у довільних опуклих областях факторного простору. Ч. 1. Теорія / С. Г. Радченко // Наукові вісті НТУУ «КПІ». – 2005. – № 3(41). – С. 38–45.

13. *Радченко С. Г.* Стійке оцінювання статистичних моделей у довільних опуклих областях факторного простору. Ч. II. Обчислювальний експеримент / С. Г. Радченко // Наукові вісті НТУУ«КПІ». – 2005. – № 4(42). – С. 48–55.

14. *Бродский В. З.* Введение в факторное планирование эксперимента / В. З. Бродский. – М.: Наука, 1976. – 224 с.

15. *Радченко С. Г.* Багатофакторне математичне моделювання та компромісна оптимізація технологічного процесу електроерозійного прошиття отворів / С. Г. Радченко // Математичні машини і системи. – 2003. – № 3, 4. – С. 186–200.

16. *Лаборатория* экспериментально-статистических методов исследований. – Режим доступа: http://www.n-t.org/sp/lesmi/

Статья поступила в редакцию 16.09.2011.