Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
ИНСТИТУТ КОСМИЧЕСКИХ И ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ
Кафедра Информатики

УТВЕРЖДАЮ
Заведующий кафедрой
_____ А. И. Рубан
«11» 10 2016 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

231000.62 «Программная инженерия»

Программная система идентификации комбинаторным методом группового учета аргументов

Руководитель _____ доцент, к.т.н. А.А. Пьяных
подпись, дата должность, ученая степень инициалы, фамилия

Выпускник _____ О. А. Полежаева
подпись, дата инициалы, фамилия

Нормоконтроль _____ доцент, к.т.н. О. А. Антамошкин
подпись, дата должность, ученая степень инициалы, фамилия

Красноярск 2016

Федеральное государственное автономное

образовательное учреждение

высшего образования

«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

ИНСТИТУТ КОСМИЧЕСКИХ И ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Кафедра Информатики

УТВЕРЖДАЮ

Заведующий кафедрой

_____ А. И. Рубан

«30» мая 2016 г.

ЗАДАНИЕ

НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ

в форме бакалаврской работы

Студенту Полежаевой Ольге Андреевне
фамилия, имя, отчество

Группа КИ 12-18Б Направление (специальность) 231000.62
номер код

Программная инженерия
наименование

Тема выпускной квалификационной работы Программная система идентификации комбинаторным методом группового учета аргументов

Утверждена приказом по университету № 6145/с от 10.05.2016

Руководитель ВКР А.А. Пьяных, доцент, к. т. н
инициалы, фамилия, должность, ученое звание и место работы

Исходные данные для ВКР: материалы, справочная, научная, методическая литература, ресурсы Интернет

Перечень разделов ВКР: Анализ теоретического материала, Обзор существующих аналогов, Технологии разработки программной системы, Описание программной системы идентификации комбинаторным методом группового учета аргументов

Перечень графического материала Подготовить раздаточный материал, и презентацию с использованием мультимедийного оборудования

Руководитель ВКР _____ А. А. Пьяных
подпись инициалы и фамилия

Задание принял к исполнению _____ С.А. Полежаева
подпись, инициалы и фамилия студента

« ___ » _____ 20__ г.

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ
Федеральное государственное автономное образовательное учреждение
Высшего образования
«Сибирский федеральный университет»

ОТЗЫВ

руководителя о бакалаврской работе студента ИКИТ группы КИ 12-18б Полежаева Ольга Андреевна на тему: «Программная система идентификации комбинаторным методом группового учета аргументов».

Целью бакалаврской работы является разработка программной системы для идентификации процессов различной природы происхождения с помощью комбинаторного метода группового учет аргументов.

Содержание бакалаврской работы соответствует теме работы и утвержденному заданию. Полежаева О. А. в ходе выполнения бакалаврской работы проявила себя как специалист, знающий методы исследования проблемы, представленные в бакалаврской работе и уместность их применения. Пояснительная записка имеет логичную структуру, выводы и правильно составленные введение и заключение, также имеется наличие очевидной связи между параграфами и главами.

Пропорциональность объемов глав и параграфов в работе соответствует достаточно полному анализу исследуемых проблем. Таким образом, бакалаврская работа имеет достаточно сбалансированную структуру, определенную соответствием важности решаемой задачи и ее объему.

В бакалаврской работе имеется достаточный список использованных источников, который соответствует теме исследования. В тексте имеются ссылки на источники, анализ которых позволяет сделать вывод об отсутствии в данной работе плагиата. Автором была проведена довольно критическая оценка вторично информации, и выполнены собственные выводы и предложения.

Замечания и предложения: Работа над программой не была доведена до конца.

Считаю, что бакалаврская работа Полежаевой О. А. заслуживает оценки «отлично», а ее автор присвоения квалификации «Бакалавр» по направлению «Программная инженерия».

Доцент «Информатика» ИКИТ СФУ
канд. тех. наук. _____ А. А. Пьяных

РЕФЕРАТ

Выпускная квалификационная работа (ВКР) по теме «Автоматизированная ВКР входит введение, четыре главы и заключение

Целью работы является разработка программной системы идентификации группового метода учета аргументов.

ВКР. содержит 37 страниц текстового документа, 6 рисунков, 1 таблицу, 14 библиографических источников.

Во введении раскрывается актуальность работы, ставятся цель и задачи.

В первой главе происходит анализ литературы, обзор метода группового учета аргументов и предлагаемых алгоритмов, описывается выбранный для реализации алгоритм.

Во второй главе идет обзор существующих аналогов разработанной программной системы.

В третьей главе описываются и выбираются методы и средства разработки.

В четвертой главе дается описание системы, алгоритма ее работы, основных функциональных возможностей и интерфейса.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Обработка больших объемов данных – задача, которую необходимо решать практически во всех сферах деятельности. На основании данных, полученных в результате эксперимента или в ходе производственного процесса, можно получить закономерности распределения данных, предсказать результаты, которые будут получены в дальнейшем, выявить необходимые изменения для оптимизации производственного процесса. В общем случае эти задачи решаются построением математической модели на основании имеющихся данных.

Такая задача возникает, например, в теплоэнергетике при работе тепловых котлов. Выявление взаимосвязей между паропроизводительностью и температурой в топке, паропроизводительностью и давлением в трубопроводе, паропроизводительностью и коэффициентом избытка воздуха позволило бы оптимизировать работу котла.

Одним из существующих методов математического моделирования является метод группового учета аргументов, он сравнительно новый и пока не часто реализуется программно, но, в то же время, он предлагает быстрые и эффективные алгоритмы обработки данных, что делает его перспективным для изучения.

В связи с этим, целью данной работы являлась разработка программного обеспечения, реализующего комбинаторный (однорядный) алгоритм МГУА на двумерной выборке и проверка работы приложения на тестовых примерах с целью дальнейшего применения в теплоэнергетике.

Для достижения данной цели решались следующие задачи:

1. Анализ теоретического материала о методе группового учета аргументов.
2. Обзор существующих аналогов.
3. Разработка приложения.

1 Анализ теоретического материала

1.1 Математическое моделирование

Математическое моделирование – динамично развивающаяся область математической науки, изучающая построение и описание математических моделей, а также сам процесс построения математических моделей.

Математическая модель (в общем смысле) – это математическое описание объекта (явления или процесса), максимально к нему приближенное.

Согласно Ляпунову, математическая модель – это опосредованное практическое или теоретическое исследование объекта, при котором непосредственно изучается не сам объект, а некоторая вспомогательная искусственная или естественная система (модель), находящаяся в некотором объективном соответствии с познаваемым объектом, способная замещать его в определенных отношениях и дающая при её исследовании, в конечном счете, информацию о самом моделируемом объекте [1].

Математические модели, в зависимости от используемых математических средств, можно разделить на различные группы, в частности на:

- линейные и нелинейные [2];
- статические и динамические [3];
- детерминированные и стохастические [3];
- дедуктивные, индуктивные и комбинированные [4][5, с. 38].

Последнее разделение хотелось бы рассмотреть подробнее.

Сначала определим понятия дедукции и индукции.

Дедукция – метод логических рассуждений, предполагающий получение частных суждений из общего [6]. В случае математического моделирования использование дедуктивного подхода предполагает, что исходными данными являются закономерности функционирования системы, а выходными – некие частные результаты [5, с. 38].

Индукция – метод логических рассуждений, предполагающий получение общего суждения из частных [6]. Случае математического моделирования использование индуктивного подхода предполагает, что исходными данными являются некоторые входные данные системы и результаты ее работы. Алгоритм моделирования в таком случае заключается в построении и полном (полная индукция) или неполном (неполная индукция) переборе всех вариантов. [5, с. 3839].

Приведем пример полной индукции.

Утверждение: любое четное число n, $4<n<24$ представимо в виде суммы двух простых чисел.

Проверим все варианты:

n = 4 4 = 2 + 2
n = 6 6 = 3 + 3
n = 8 8 = 3 + 5
n = 10 10 = 5 + 5
n = 12 12 = 5 + 7
n = 14 14 = 7 + 7
n = 16 16 = 5 + 11
n = 18 18 = 7 + 11
n = 20 20 = 13 + 7
n = 22 22 = 19 + 3
n = 24 24 = 19 + 5

Мы перебрали все возможные варианты, и убедились, что наше утверждение верно.

Комбинированный метод предполагает, что часть информации о системе нам известна, и мы используем индуктивное моделирование только для тех случаев, когда у нас нет никакой априорной информации [5, с. 40].

Одним из методов индуктивного моделирования является метод группового учета аргументов.

1.2 Задача идентификации

Задача идентификации формулируется следующим образом. Пусть в результате каких либо экспериментов над объектом замерены его входные $X=(x_1, x_2, \ldots x_n)$ и выходные переменные $Y=(y_1, y_2, \ldots y_m)$ как функции времени. Требуется определить вид (структуру) и параметры некоторого оператора $\hat{A}$, ставящего в соответствие переменные X и Y.

Задачи идентификации могут быть поставлены в узком (параметрическая идентификация) и широком (структурная идентификация) смысле соответственно. В первом случае неизвестна структура и параметры оператора$\hat{A}$, во втором – лишь параметры этого оператора.

Таким образом, очевидна тесная взаимосвязь задачи идентификации с проведением эксперимента и обработкой экспериментальных зависимостей.

Задача параметрической идентификации сводится к отысканию таких оценок параметров математической модели $\hat{A}$, которые обеспечивают в каком либо смысле близость расчетных и экспериментальных значений выходных переменных при одинаковых входных. В общем случае необходимы измерения *«m»* компонент вектора Y, которые могут производиться при *«k»* повторениях эксперимента при *«l»* дискретных отметках времени (если идентифицируемый объект функционирует во времени). В качестве критериев количественной меры близости модели и оригинала чаще всего используются максимальные δ_y, средние m_y и среднеквадратичные δ_y величины погрешностей рассогласования расчетных и экспериментальных значений y_{pi} и $y_{эi}$, соответственно, т.е

$$d_y = max\,|y_{pi} - y_{эi}|$$

$$m_y = 1/N\sum (y_{pi} - y_{эi}) \qquad (1)$$

$$\sigma_y = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{pi} - y_{эi})^2}$$

где: i = 1, 2…, N = m + l + k - номер опыта по измерению компоненты $y_{эi}$

Таким образом, задача параметрической идентификации сводится к минимизации одной из функций вида (1). Для минимизации могут быть использованы известные численные методы решения экстремальных задач.

Обилие существующих методов идентификации отражает разнообразие используемых математических моделей и методов их исследования. Очевидно, что идентифицировать модель детерминированного, линейного, стационарного процесса (модель считается стационарной, если её параметры либо постоянные, либо меняются медленно по сравнению со временем, необходимым для их идентификации) известной размерности, с одним входом – существенно проще, чем аналогичного стохастического процесса неизвестного порядка и степени стационарности.

Идентификация моделей с помощью регрессионного метода.

Регрессионный анализ представляет собой классический статистический метод. Благодаря своим широким возможностям регрессионные методы давно и успешно используются в инженерной практике. Регрессионный анализ основывается на двух главных принципах.

1. Методы применяются для линейных по идентифицируемым параметрам моделям. Структура математической модели процесса представляется функцией вида:

$$y = \sum_{i=1}^{N} a_i f_i(\vec{x}) \qquad (2)$$

где a_i – i-тый оцениваемый параметр; $f_i(x)$ – i-тая известная функция, x - вектор входных воздействий, y – выходная переменная.

На практике, чаще всего в качестве $f_{i\,(x)}$ выбираются степенные функции, а соответственно выражение (2) является полиномиальной, либо дробнорациональными зависимостью. При этом точность описания достигается увеличением числа членов полинома, обеспечивающих их сходимость к

реальному процессу. Заметим, что получающаяся модель практически никогда не соответствует физической сущности моделируемого реального процесса, его истинному виду, однако инженерная простота вычислений, удобство практического использования модели, возможность получения результата без «особых размышлений» служит основной причиной широкого распространения на практике регрессионных методов.

Естественно, и в этом случае с помощью удачно выбранного вида полинома можно существенно сократить размер модели, а значит и трудоемкость вычислительного процесса, как при идентификации, так и при использовании модели.

2. Минимизируемой функцией ошибки (разности между прогнозируемой моделью и данными эксперимента) при регрессионном анализе является сумма квадратов ошибок. Благодаря этому удается применить метод наименьших квадратов, математический аппарат которого предельно прост, а вычислительные методы сводятся к методам линейной алгебры.

Регрессионные модели могут быть как линейными, так и нелинейными с любым числом входов и выходов.

1.2 Метод наименьших квадратов

Параметры могут входить в модель линейно и нелинейно.

Пример нелинейного вхождения параметров в модель:

$$y = a_0 + a_1 \cos(b_1 x_1) + a_2 \sin(b_2 x_2),$$

где a_i – линейно входящий параметр, b_i – нелинейно входящий параметр. В таких случаях для оценки параметров моделей используется нелинейное программирование.

В случае линейного вхождения параметров в модель для их оценки используется метод наименьших квадратов.

Метод наименьших квадратов (МНК) — математический метод, применяемый для решения различных задач, основанный на минимизации суммы квадратов отклонений некоторых функций от искомых переменных.

Сущность метода наименьших квадратов.

Пусть x – набор n неизвестных переменных (параметров), $f_i(x)$,

$i = 1,\ldots,m,\ m > n$ – совокупность функций от этого набора переменных. Задача заключается в подборе таких значений x, чтобы значения этих функций были максимально близки к некоторым значениям y_i. По существу речь идет о «решении» переопределенной системы уравнений $f_i(x) = y_i,\ i = 1,\ldots,m$ в указанном смысле максимальной близости левой и правой частей системы. Сущность МНК заключается в выборе в качестве «меры близости» суммы квадратов отклонений левых и правых частей $| f_i(x) - y_i |$. Таким образом, сущность МНК может быть выражена следующим образом:

$$\sum_i e_i^2 = \sum_i (y_i - f_i(x))^2 \rightarrow \min_x$$

В случае, если система уравнений имеет решение, то минимум суммы квадратов будет равен нулю и могут быть найдены точные решения системы уравнений аналитически или, например, различными численными методами оптимизации. Если система переопределена, то есть, говоря нестрого, количество независимых уравнений больше количества искомых переменных, то система не имеет точного решения и метод наименьших квадратов позволяет найти некоторый «оптимальный» вектор x в смысле максимальной близости векторов y и $f(x)$ или максимальной близости вектора отклонений e [8].

МНК в случае линейной функции. Пусть исходная функция, параметры которой надо найти – $y = ax + b$.

Подбираем y=ax+b таким образом, чтобы сумма квадратов отклонений была наименьшей. Чтобы найти минимум функции, надо вычислить частные производные по каждому из параметров a и b и приравнять их к нулю.

$$\sum (y_i - \tilde{y}_{x_i})^2 \to \min$$

Обозначим сумму квадратов отклонений ($\Sigma\varepsilon_i^2$) через S, тогда:

$$S = (\overline{y_1} - y_1)^2 + (\overline{y_2} - y_2)^2 + \ldots + (\overline{y_n} - y_n)^2$$

$$S = (ax_1 + b - y_1)^2 + (ax_2 + b - y_2)^2 + \ldots + (ax_n + b - y_n)^2$$

$$S = \sum_{i=1}^{n} (ax_i + b - y_i)^2$$

S зависит от a и b, т.е. функция двух переменных принимает наименьшее значение в стандартной точке, которая находится из условия:

$$\begin{cases} S'_a = 0 \\ S'_b = 0 \end{cases}$$

$$S'_a = 2\sum_{i=1}^{n} (ax_i + b - y_i) \cdot x_i = 2\left(a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i \cdot x_i\right)$$

$$S'_b = 2\sum_{i=1}^{n} (ax_i + b - y_i) = 2\left(a\sum_{i=1}^{n} x_i + bn - \sum_{i=1}^{n} y_i\right)$$

Приравняем каждую частную производную к нулю:

$$\begin{cases} a\sum\limits_{i=1}^{n} x_i^2 + b\sum\limits_{i=1}^{n} x_i - \sum\limits_{i=1}^{n} y_i x_i = 0 \\ a\sum\limits_{i=1}^{n} x_i + bn - \sum\limits_{i=1}^{n} y_i = 0 \end{cases}$$

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \cdot x_i \\ a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i \end{cases}$$

Метод наименьших квадратов для линейной регрессии.

В общем случае уравнение линейной регрессии с константой выглядит так:

$$y = a_0 + \sum_{i=1}^{N} a_i x_i = x^T a,$$

где a_i – коэффициенты модели, x_i – данные наблюдений, $x_0 = 1$.

Пусть y — вектор-столбец наблюдений объясняемой переменной, а X – это (n × k)-матрица наблюдений факторов (строки матрицы — векторы значений факторов в данном наблюдении, по столбцам — вектор значений данного фактора во всех наблюдениях). Матричное представление линейной модели имеет вид:

$$y = Xb$$

Тогда вектор оценок объясняемой переменной и вектор остатков регрессии будут равны:

$$\hat{y} = Xb$$

$$e = y - \hat{y} = y - Xb$$

Соответственно, сумма квадратов остатков регрессии будет равна:

$$\sum_{i=1}^{n} e_i^2 = e^T e = (y - Xb)^T (y - Xb)$$

Дифференцируя эту функцию по вектору параметров b и приравняв производные к нулю, получим систему уравнений (в матричной форме):

$$(X^T X)b = X^T y$$

В расшифрованной матричной форме эта система уравнений выглядит следующим образом:

$$\begin{pmatrix} \sum 1 & \sum x_{i1} & \cdots & \sum x_{in} \\ \sum x_{i1} & \sum x_{i1}^2 & \cdots & \sum x_{i1}x_{in} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_{in} & \sum x_{in}x_{i1} & \cdots & \sum x_{in}^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_{i1}y_i \\ \vdots \\ \sum x_{in}y_i \end{pmatrix}$$

Так как в модель включена константа, в левом верхнем углу матрицы системы уравнений находится количество наблюдений n, а в остальных элементах первой строки и первого столбца — просто суммы значений переменных и первый элемент правой части системы.

Решение этой системы уравнений и дает общую формулу МНК-оценок для линейной модели [8][9].

1.3 Метод группового учета аргументов

Метод группового учета аргументов (МГУА) – метод, описывающий семейство индуктивных алгоритмов математического моделирования многопараметрических систем.

Метод был разработан в 1968 году академиком Алексеем Григорьевичем Иваненко, сотрудником Института АН Украины.

МГУА предполагает построение постепенно усложняющихся моделей заданного класса и выбор из них оптимальной для заданного набора экспериментальных данных.

Для реализации метода группового учета аргументов выборку делят на 2 части: тренировочную (Training) и проверочную (Checking). На тренировочной выборке мы рассчитываем оценки коэффициентов модели, а на проверочной проверяем их для выбора модели, наиболее хорошо описывающей всю выборку. Иногда выделяют также удостоверяющую (Verifying) выборку, которая используется для проверки качества выбранной модели. Выборку не обязательно делить поровну.

МГУА в общем случае следует следующему алгоритму:

1. Построение моделей возрастающей сложности заданного класса.
2. Расчет параметров моделей на тренировочной выборке.
3. Проверка соответствия модели некоему внешнему критерию на проверочной выборке.
4. Если значение критерия удовлетворяет некоторому пороговому условию, то выбираем эту модель как подходящую для данного набора данных и останавливаемся, иначе:
   a. в случае многорядного алгоритма возвращаемся на шаг 1.
   b. в случае однорядного алгоритма выбираем модель, для которой значение критерия максимально приближено к пороговому условию, выбираем эту модель как подходящую для данного набора данных и останавливаемся.

1.3.1 Построение моделей

На этом шаге нужно определить класс моделей, которые будут использоваться для описания выборки.

Обычно выбор класса моделей зависит от конкретной задачи, это может быть комбинация тригонометрических, степенных или других функций одной или нескольких переменных.

Алгоритм построения моделей зависит от конкретного алгоритма.

1.3.2 Внешние критерии

Внешний критерий выбора модели – это критерий полученный на дополнительной информации, которой не было в данных, использовавшихся при вычислении параметров модели. Например, если разделить выборку, то параметры модели можно рассчитывать на одной (обучающей) части, авнешний критерий на другой (тестовой) части выборки.

В алгоритмах метода группового учет а аргументов используются и внешний, и внутренний критерий. Внутренний критерий используется для настройки параметров модели, внешний критерий используется для выбора модели оптимальной структуры. Возможен выбор моделей по нескольким внешним критериям.

1. Критерий регулярности

Критерий регулярности $\Delta^2(C)$включает среднеквадратичную ошибку на обучающей подвыборке C полученную при параметрах модели, настроенных на тестовой подвыборке ℓ.

$$\Delta^2(C) = |y_C - A_C\widehat{\omega_l}|^2 = (y_C - A_C\widehat{\omega_l})^T(y_C - A_C\widehat{\omega_l}) \quad (3)$$

где

$$\widehat{w}_\ell = (A_\ell^T A_\ell)^{-1}(A_\ell^T y_\ell) \quad (4)$$

и

$$\widehat{y}_C(\ell) = A_C\widehat{w}_\ell. \quad (5)$$

Другие модификации критерия регулярности:

$$\Delta^2(C) = \frac{|y_C - A_C \hat{w}_\ell|^2}{|y_C|^2} \quad (7)$$

$$\Delta^2(C) = \frac{|y_C - A_C \hat{w}_\ell|^2}{|y_C - \bar{y}_C|^2}, \quad (8)$$

где $\bar{y}$ — среднее значение вектора y.

Критерий $\Delta^2(C)$ также обозначается $\Delta^2(C \setminus \ell)$, то есть ошибка на подвыборке C, при параметрах, полученных на подвыборке ℓ.

2. Критерий минимального смещения.

Критерий минимального смещения по-другому называется критерий непротиворечивости модели: модель которая имеет на обучающей выборке одну невязку, а на контрольной — другую, называется противоречивой. Смысл критерия минимального смещения в том, что оптимальной считается модель, дающая наименьшую разницу значений выходной переменной, полученных в одной точке из моделей, обученных на разных частях выборки. Иными словами, критерий минимального смещения требует, чтобы оценки параметров оптимальной моднли на разных частях выборки различались минимально.

Критерий непротиворечивости как критерий минимума смещения имеет вид

$$n_{\text{см}}^2 = \frac{\sum_1^N (y_T - y_C)^2}{\sum_{i=1}^N y_{i0}^2},$$

Где $n_{см}$ - коэффициент минимального смещения, y_T - значения выходной переменной, полученные на обучающей выборке, y_C - значения выходной переменной, полученные на проверяющей выборке, y_{0i} – i-е значение выходной переменной из исходных данных.

Другие модификации этого критерия:

$$\eta_{\text{bs}}^2 = \frac{|A_W \widehat{w}_\ell - A_W \widehat{w}_C|^2}{|y_C - \bar{y}_C|^2}$$

и

$$\eta_a^2 = |\widehat{w}_\ell - \widehat{w}_C|^2,$$

где $\widehat{w}_\ell$ $\widehat{w}_C$ и — векторы коэффициентов, полученные с использованием подвыборок ℓ и C. При использовании последнего варианта следует помнить, что число элементов вектора параметров w в различных моделях может быть различно.

Критерий absolute noise-immune. Утверждается, что с помощью этого критерия, из сильно зашумленных данных возможно найти скрытые физические закономерности.

$$V^2 = (A_W \widehat{w}_\ell - A_W \widehat{w}_W)^T (A_W \widehat{w}_W - A_W \widehat{w}_C) = \\ = (\widehat{w}_\ell - \widehat{w}_W)^T A_W^T A_W (\widehat{w}_W - \widehat{w}_C). \quad (12)$$

где $\widehat{w}_W$ — вектор коэффициентов, полученный на всей выборке W.

Комбинированный критерий. Этот критерий позволяет использовать при выборе моделей линейную комбинацию нескольких критериев. Комбинированный критерий

$$k^2 = \sum_{i=1}^{K} \alpha_i k_i^2, \text{ при условии нормировки } \sum_{i=1}^{K} \alpha_i = 1. \quad (13)$$

Здесь k_i — принятые на рассмотрение критерии, а α_i — веса этих критериев, назначенные в начале вычислительного эксперимента.

Используются также нормализованные значения критериев. При этом предыдущая формула имеет вид

$$k^2 = \sum_{i=1}^{K} \alpha_i \frac{k_i^2}{k_{i\max}^2}. \quad (14)$$

Максимальное значение критерия $k_{i\max}^2$ берется по вычисленным значениям критериев для всех порожденных моделей. В данном случае оптимальная модель может быть найдена только после завершения настройки параметров всех моделей.

Пример комбинированного критерия — смещение плюс регулярность.

$$c_2^2 = \bar{\eta} bs^2 + \overline{\Delta}^2(C). \quad (15)$$

Второй пример — смещение плюс ошибка на тестовой выборке.

$$c_3^2 = \bar{\eta} bs^2 + \overline{\Delta}^2(B \backslash W). \quad (16)$$

Такой критерий обеспечивает выбор наиболее несмещенных, устойчивых и точных моделей. Здесь $\Delta(C \backslash W)$ — среднеквадратичная ошибка, вычисленная на выборке C, с весами, настроенными на всей выборке W.

1.3.3 Алгоритмы моделирования

Целю МГУА является получение модели в результате перебора моделей из индуктивно-порождаемого множества. Параметры каждой модели настраиваются так, чтобы доставить минимум выбранному внешнему критерию. Различают два основных типа алгоритмов МГУА — однорядный и многорядный.

Все алгоритмы МГУА воспроизводят схему массовой селекции: последовательно порождаются модели возрастающей сложности. Каждая модель настраивается — методом наименьших квадратов находятся значения параметров. Из моделей-претендентов выбираются лучшие в соответствии с выбранным критерием. Многорядные алгоритмы могут вычислять остатки регрессионных моделей после каждого ряда селекции или не вычислять; при этом используются исходные данные.

Комбинаторный (однорядный) алгоритм использует только один ряд выбора. При этом порождаются все возможные линейные комбинации ограниченной сложности. Так как под сложностью понимается число линейно входящих параметров w, то сложность не превосходит заданное значение F_0. Пусть, как и ранее

$$y = w_0 + w_1 a_1 + w_2 a_2 + w_3 a_3, \ldots, w_{F_0} a_{F_0}.$$

Алгоритм выполняет следующие шаги. Для всех комбинаций входных аргументов строятся модели-претенденты неубывающей сложности.

Параметры каждой модели настраиваются методом наименьших квадратов по обучающей выборке. Наилучшая модель выбирается исходя из минимума значения внешнего критерия. Как вариант — назначается порог и выбираются несколько моделей, значения критерия для которых не превышает этот порог.

Именно этот алгоритм был выбран для реализации в данной выпускной работе.

1.3.4 Полное описание

В комбинаторных алгоритмах МГУА полное описание – функция, задающая класс генерируемых алгоритмом моделей.

На основе полного описания генерируются частные описания – модели, которые строятся алгоритмом.

Частные описания получаются при приравнивании к нулю коэффициентов полного описания.

Пусть, например частное описание задано формулой:

$$y = a_0 + a_1 x_1^2 + a_2 x_2^2 + \cdots + a_n x_n^2,$$

тогда среди частных описаний будут, генерированных комбинаторным алгоритмом будут, например, такие:

$$y = a_0 + a_1 x_1^2,$$
$$y = a_0 + a_1 x_1^2 + a_3 x_3^2,$$
$$y = a_0 + a_2 x_{21}^2 + a_4 x_4^2 + a_7 x_7^2,$$

Вывод

Для реализации задачи идентификации комбинаторным методом учета аргументов в качестве внешнего критерия был выбран критерий несмещенности, как наиболее подходящий для данной задачи. Для расчета оценок параметров моделей был выбран метод наименьших квадратов.

В качестве полного описания была выбрана следующая функция:

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1 x_2 + a_4 x_1^2 + a_5 x_2^2$$

2 Обзор существующих аналогов

На текущий момент, существует много приложений, реализующих МГУА, но большинство из них написаны для конкретной задачи и/или предприятия и не находятся в открытом доступе [10].

Сравнение приложений, реализующих МГУА, информация о которых находится в открытом доступе, приведено в таблице 1.

Таблица 1. Анализ существующих систем.

Название	Цена, р	Платформа	Интерфейс	ТП	Тестовый период	Прочее
KnowledgeMiner (ух) for Excel	От 19631 до 25520	Mac OS	8/10	+	+	Нужно приобретать весь пакет
Zapron MineMaster	Неизвестна	Windows 7/8, Windows Server 2012	5/10	+	+	Сотрудничают напрямую с клиентами, основной язык-китайский

Самый популярный аналог «KnowledgeMiner (ух) for Excel» имеет развитый функционал, относительно неплохую техническую поддержку, но работает только на платформе Mac OS и имеет достаточно высокую цену. Версия для обработки очень больших массивов данных (так называемая Gold-версия) – наиболее дорогая [11].

Второй по популярности аналог «Zapron MineMaster» также имеет ряд недостатков, в первую очередь, он ориентирован в основном на китайский рынок, английская версия в разработке, получить информацию о продукте очень сложно. Кроме того, компания Zapron работает напрямую с предприятиями и получить версию продукта для разработчика довольно сложно [12].

Преимущества системы «PSI», в том, что она, во-первых, работает на платформе Windows 7/8, суммарная доля которых на рынке операционных систем составляет более 50 % [13], во-вторых, ее интерфейс на русском языке,

что повышает удобство использования для российских специалистов, и в-третьих, оно ориентировано на отдельных специалистов.

3 Технологии разработки программной системы

3.1 Методологии разработки программного обеспечения

В данном подразделе описано несколько методологий разработки программного обеспечения, рассмотрены их преимущества и недостатки. В конце подраздела происходит выбор оптимальной методологии для разработки автоматизированной информационной системы учета стажировок.

3.1.1 Каскадная модель

Каскадная модель – модель разработки программного обеспечения, суть которой состоит в том, что процесс разработки разбивается на этапы, которые выполняются строго один за другим.

В каскадной модели каждая из процессных областей представляет собой отдельную фазу проекта.

Основная причина выбора именно этой модели разработки заключается в том, что при строго документированных и несколько раз проверенных до завершения фазы проектирования требованиях к программному обеспечениювероятность ошибки на более поздних стадиях крайне мала, что существенно снижает затраты на доработку проекта. Каскадная модель предполагает, что стадии выполняются только после завершения предыдущих, то есть, разработка кода не может начаться раньше, чем закончится разработка архитектуры, например.

Использование данного подхода в проектах со строго документированными требованиями позволяет обеспечить качественное выполнение работы и уложиться в заданные финансовые и временные рамки.

Однако у данной модели есть и недостатки.

1. В случае нечетких требований к проекту может получиться так, что в проектировании была допущена ошибка, и из-за особенностей модели, она выявится только на одном из заключительных этапов, например при тестировании, что может привести к необходимости полной переработки проекта.

2. В случае, если заказчик вносит изменения в требования по ходу разработки проекта, использование данной модели является весьма неэффективным, так как при каждом внесенном изменении предполагается возвращение к предыдущим фазам разработки.

3. Сложно управлять рисками некоторых типов (таких, как риски, связанные с использованием новых технологий или риски некорректного определения требований). Подобные риски могут проявить себя только на этапе реализации (если не тестирования), когда число возможных путей исправления ситуации намного меньше, чем в начале проекта.

4. Весьма ограничены возможности оценки и корректировки важных атрибутов проекта – скорости разработки, качества продукта, обоснованности принятых архитектурных решений. Адекватно оценить эти атрибуты становится возможным только на поздних этапах проекта.

В случае, если требования к проекту четко определены и он является типовым, использование модели водопада будет оптимальным выбором.

3.1.2 Итеративная модель

Процесс итеративной (или инкрементальной) разработки стал эволюционным развитием модели «водопада» [14].

В случае использования итеративной модели процесс разработки состоит из повторяющихся мини-процессов (итераций), в каждом из которых выполняется полный цикл разработки со всеми фазами. Каждая новая итерация предполагает добавление в проект новой функциональности или улучшения

существующей. Полный набор требований, зафиксированный границами проекта, оказывается реализованным после завершения финальной итерации.

Итеративная разработка обладает рядом преимуществ по сравнению с каскадной моделью.

1. Наиболее важные функции (модули) могут быть реализованы после первых нескольких итераций, что позволяет заказчику начать использование системы уже в середине разработки.

2. Заказчик получает возможность оценить конечный результат уже на начальных стадиях разработки и внести необходимые исправления.

3. Основные проектные риски могут (и должны) быть разрешены на первых итерациях. Например, архитектурное решение, приводящее к неприемлемой производительности может быть обнаружено и исправлено уже в первой итерации.

Данные преимущества возникают только при условии правильного планирования итераций. Итеративная модель используется во многих процессах разработки, включая RUP и гибкие методологии.

Для разработки данного проекта была выбрана итеративная модель.

3.2 Средства разработки

Проект разрабатывался на языке программирования C# с использованием среды разработки Microsoft Visual Studio 2012, для реализации отображения трехмерных графиков был использован компонент среды разработки ComponentOne.

4 Описание программной системы идентификации комбинаторным

методом группового учета аргументов

4.1 Алгоритм работы системы

На рисунке 1 представлен общий алгоритм работы приложения.

Рисунок 1 – Алгоритм работы приложения

4.2 Полное наименование

Полное наименование: программная система идентификации

комбинаторным методом группового учета аргументов «PSI»

4.3 Назначение разрабатываемой системы

Приложение «PSI» предназначено для анализа двумерных массивов данных с целью идентификации. Предназначено для аналитиков.

4.4 Перечень модулей, их назначение и основные характеристики

В состав приложения «PSI» входят следующие модули:

- модуль «Вычисление»;
- модуль «Работа с выборкой».
- модуль «Построение графиков».

В модуль «Вычисление» входят алгоритмы вычисления коэффициентов моделей МНК, алгоритм образования моделей и алгоритм расчета внешнего критерия и выбора из получившегося набора моделей оптимальной.

В модуль «Работа с выборкой» входит возможность загрузки данных выборки из сторонних файлов и просмотра выборки для обеспечения возможности работы модуля «Вычисление».

В модуль «Построение графиков» входит построение точечных графиков модели и выборки по данным, полученным в результате работы модуля «Вычисление».

4.5 Описание интерфейса системы

Интерфейс пользователя выполнен в стандартном стиле Windows.

На основной форме есть кнопки «Рассчитать», «График», «Добавить выборку», «Выход», элементы управления «вверх-вниз» «Выборка 1» и «Выборка 2», отображающие в процентах значения тренировочной и обучающей выборок, поля «Оптимальная модель» и «Все модели».

После нажатия кнопки «Добавить выборку» открывается окно «Выборка», на котором есть кнопки «Открыть файл», «ОК» и поле «Выборка». После нажатия кнопки «Открыть файл» появляется стандартный диалог открытия файла. После выбора файла, содержащего выборку, в поле «Выборка», появляются значения данных. После нажатия кнопки «ОК» выборка сохраняется и окно закрывается.

После нажатия кнопки «Рассчитать» появляются модели, сгенерированные для данной выборки и зеленым цветом подсвечивается оптимальная.

4.6 Диаграмма вариантов использования

На рисунке 2 приведена диаграмма вариантов использования для процесса вычисления моделей.

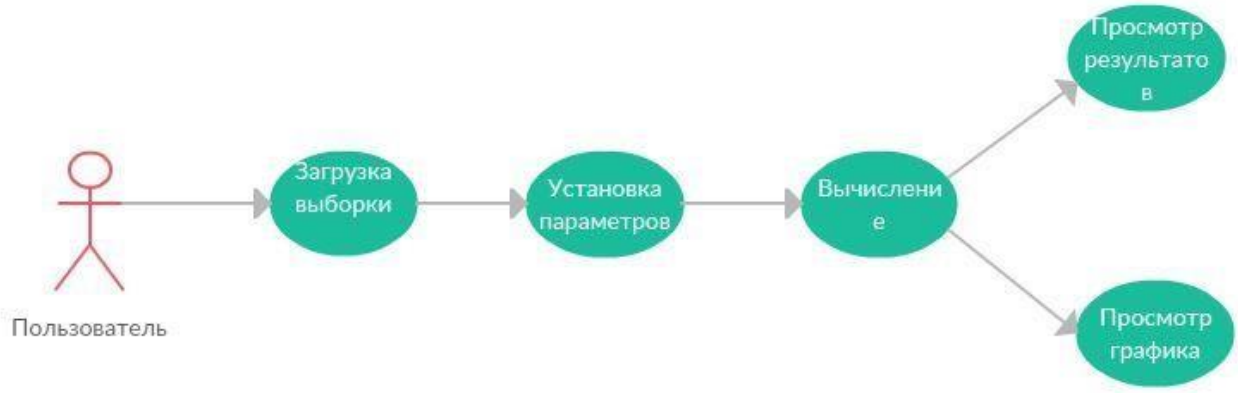


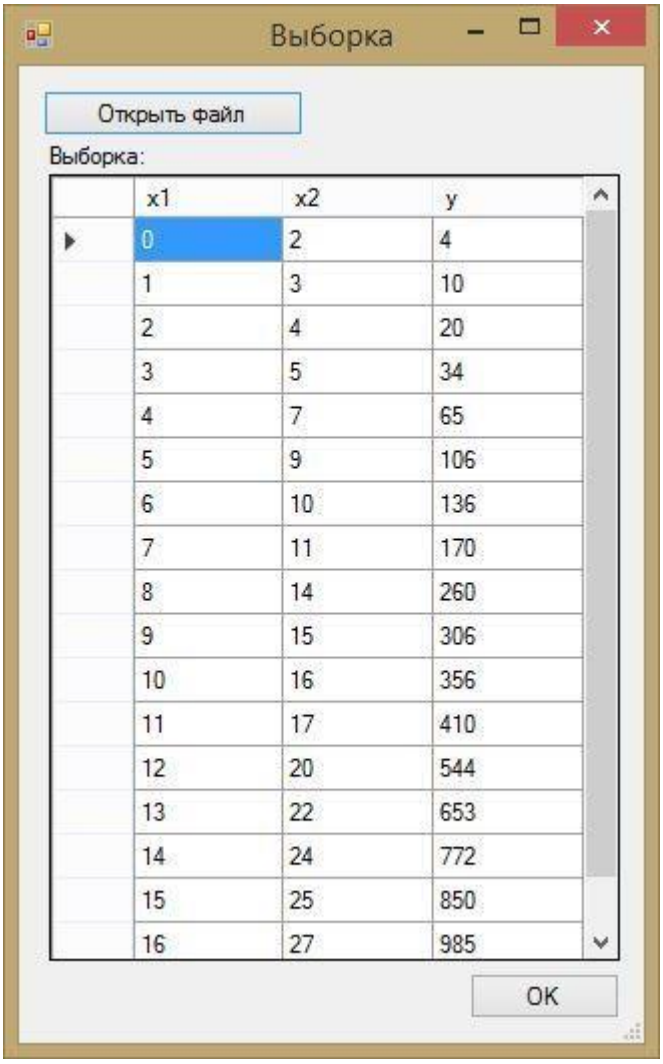
Рисунок 2 – Получение оптимальной модели

4.7 Описание разработанных модулей

В состав приложения «PSI» входят следующие модули:

- модуль «Выборка»;
- модуль «Вычисление»;
- модуль «График».

Модуль «Выборка» предоставляет возможность открытия выборки и загрузки ее в программу. На рисунке 3 изображено окно модуля «Выборка» с уже загруженными данными. Данная выборка является тестовой, она идентифицируется моделью $y = x_1^2 + x_2^2$.



	x1	x2	y
▶	0	2	4
	1	3	10
	2	4	20
	3	5	34
	4	7	65
	5	9	106
	6	10	136
	7	11	170
	8	14	260
	9	15	306
	10	16	356
	11	17	410
	12	20	544
	13	22	653
	14	24	772
	15	25	850
	16	27	985

Рисунок 3 – Модуль «Выборка»

Модуль «Вычисление» предоставляет возможность установки параметров разбиения выборки, генерирования всех возможных моделей из данного полного описания и выбора оптимальной модели. Окно модуля «Вычисление» является главным окном программной системы. Главное окно модуля «Вычисление» с результатами работы представлено на рисунке 4. Как мы видим, программная система верно идентифицировала модель.

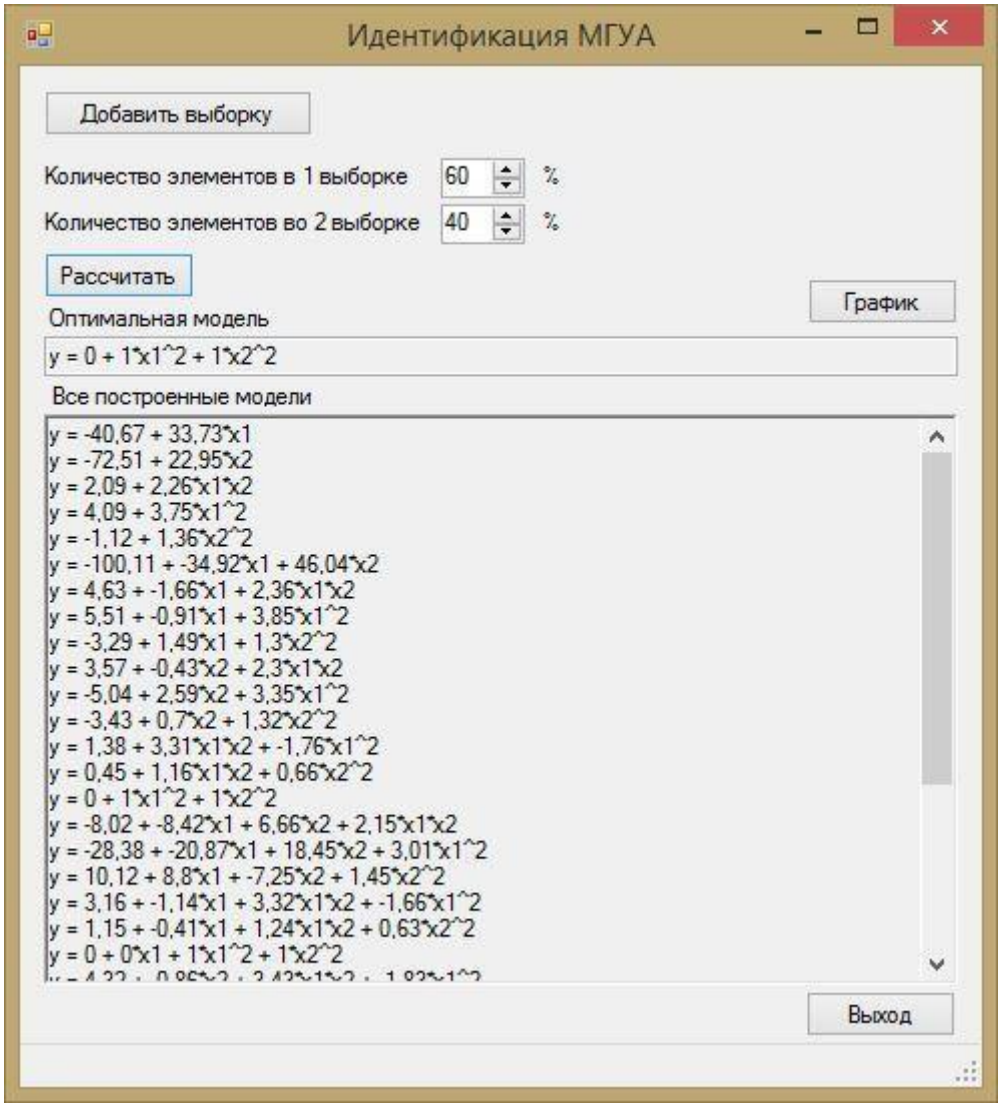


Рисунок 4 – Модуль «Вычисление»

Модуль «График» предоставляет возможность получить точечные графики модели и выборки для определения точности модели. На рисунках 5-6 представлены графики, построенные для выборки и модели соответственно. Данные графики наглядно показывают, что модель верно описывает выборку.

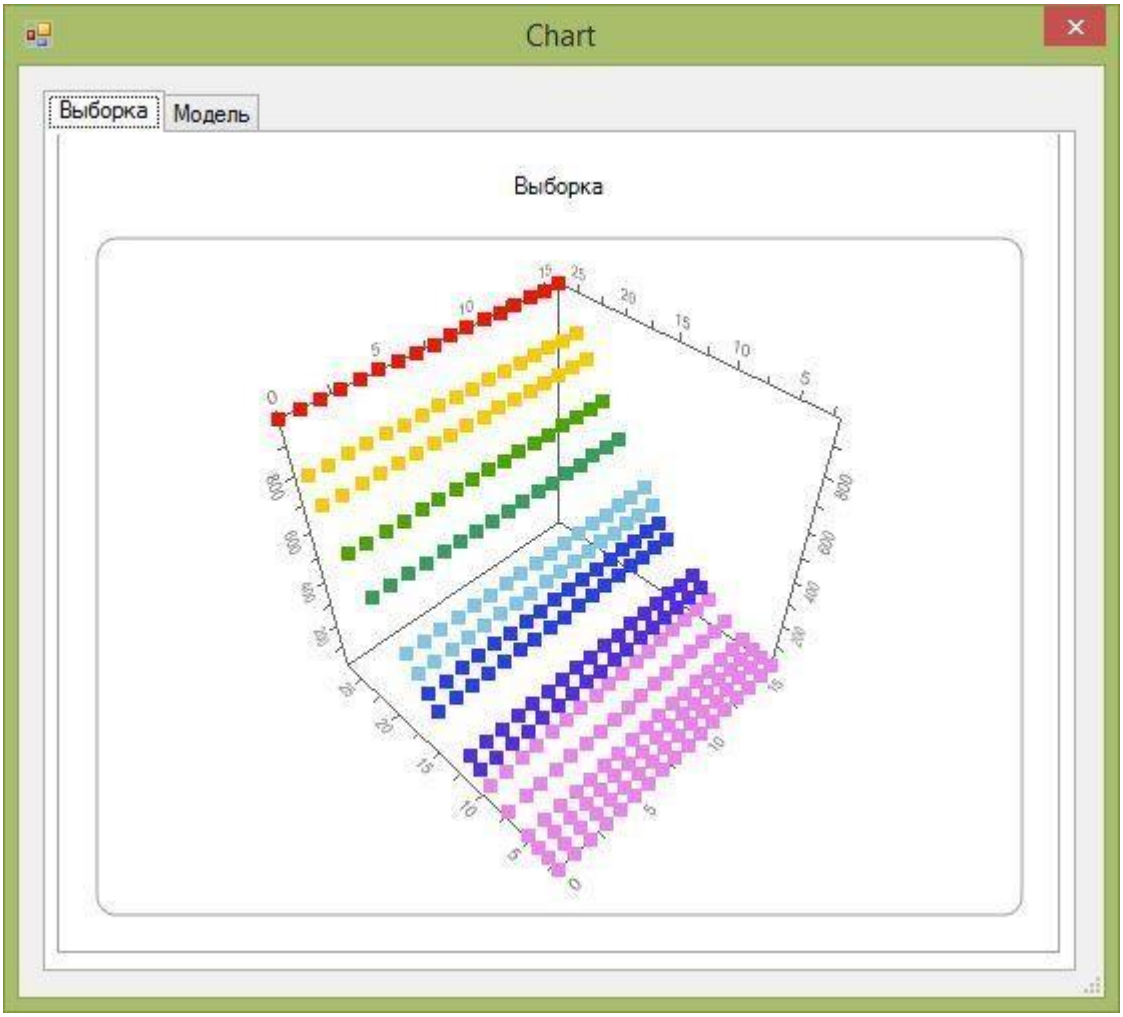


Рисунок 5 – График выборки.

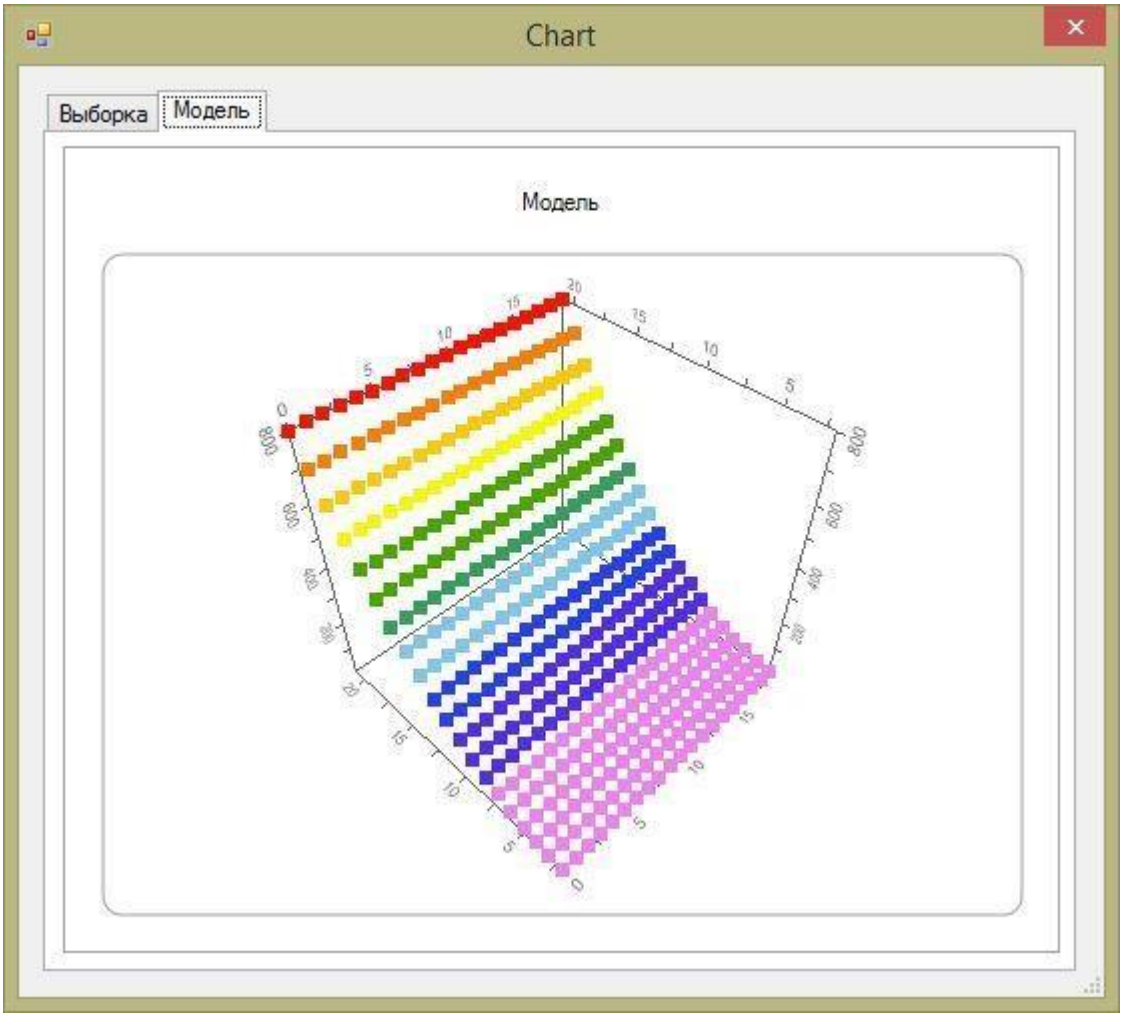


Рисунок 6 – График модели.

ЗАКЛЮЧЕНИЕ

В результате ВКР была реализована программная система идентификации комбинаторным методом учета аргументов. Все поставленные задачи были выполнены, программная система успешно отработала на тестовых примерах .

Сильной стороной данного проекта является наглядность представления результатов, что достигнуто использованием специализированного компонента среды разработки Microsoft Visual Studio ComponentOne.

В дальнейшем планируется использовать данную систему в теплоэнергетике для выявления новых зависимостей между технологическими параметрами теплоэнергетических объектов.

ПЕРЕЧЕНЬ СОКРАЩЕНИЙ

МГУА – Метод Группового Учета Аргументов.

МНК – Метод Наименьших Квадратов.

ПО – Программное Обеспечение.

ПК – Персональный Компьютер.

ВКР – Выпускная Квалификационная Работа.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Новик, И. Б. О философских вопросах кибернетического моделирования. / И. Б. Новик – М.: Знание, 1964.

2. Данилов, Ю. А. Лекции по нелинейной динамике. Элементарное введение. Серия «Синергетика: от прошлого к будущему». Изд.2. / Ю.А. Данилов – М.: URSS, 2006. — 208 с.

3. Советов, Б. Я. Моделирование систем: Учеб. для вузов — 3-е изд., перераб. и доп. / Б. Я. Советов, С. А. Яковлев – М.: Высш. шк., 2001. — 343 с.

4. Andreski, S. Social Sciences as Sorcery. / S. Andreski – New York: St. Martin’s Press, 1972.

5. Ивахненко, А. Г. Индуктивный метод самоорганизации моделей сложных систем. / А. Г. Ивахненко – Киев: Наукова думка, 1982.

6. Прохоров, А. М. Большая советская энциклопедия / Ред. А. М. Прохоров, Н. К. Байбаков, А. А. Благонравов – М.: Советская Энциклопедия, 1969— 1978.

7. Радлов, Э. Л. Индукция в логике / Э. Л. Радлов // Энциклопедический словарь Брокгауза и Ефрона : в 86 т. (82 т. и 4 доп.). — СПб.: 1890—1907

8. Линник, Ю. В. Метод наименьших квадратов и основы математикостатистической теории обработки наблюдений. – 2-е изд. / Ю. В. Линник – М.: 1962.

9. Рубан, А. И. Теория вероятностей и математическая статистика /А. И.

Рубан. – Красноярск: ИПЦ КГТУ, 2002.

10. Software // GMDH National Institute for Strategic Studies International Center for Informational Technologies and Systems of the National Academy of Science of Ukraina [Электронный ресурс]. – Режим доступа: http://www.gmdh.net/GMDH_sof.htm

11. KnowledgeMiner – Store // KnowledgeMiner Software That Extracts Knowledge From Data [Электронный ресурс]. – Режим доступа: http://www.knowledgeminer.com/index.htm

12. 产品 // Zaptron [Электронный ресурс]. – Режим доступа:

http://zaptron.com/?page_id=61

13. Desktop Operating System Market Share // NetMarketShare Market Share Statistics for Internet Technologies [Электронный ресурс] – Режим доступа: https://www.netmarketshare.com/operating-system-market-share.aspx?qprid=10&qpcustomd=0

14. Рахимбердиев, А. Современные процессы разработки программного обеспечения / А. Рахимбердиев // RSDN [Электронный ресурс]: – Режим доступа: http://www.rsdn.ru/article/Methodologies/SoftwareDevelopmentProcesses.xml