На правах рукописи

Сапкина Наталья Владимировна

ВОССТАНОВЛЕНИЕ ЗАКОНОМЕРНОСТЕЙ НА ОСНОВЕ НЕЧЕТКИХ РЕГРЕССИОННЫХ МОДЕЛЕЙ

Специальность 05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Воронеж – 2014

Работа выполнена в ФГБОУ ВПО «Воронежский государственный университет»

Научный руководитель — доктор технических наук, профессор **Леденёва Татьяна Михайловна**

Официальные оппоненты:

Буховец Алексей Георгиевич, доктор технических наук, доцент, ФГБОУ ВПО «Воронежский государственный аграрный университет», профессор кафедры прикладной математики и математических методов в экономике

Сербулов Юрий Стефанович, доктор технических наук, профессор, ФГБОУ ВПО «Воронежская государственная лесотехническая академия», профессор кафедры вычислительной техники и информационных систем

Ведущая организация: ФГАОУ ВПО «Белгородский государственный национальный исследовательский университет»

Защита состоится «25» июня 2014 г. в 15.00 на заседании диссертационного совета Д.212.038.24 при ФГБОУ ВПО «Воронежский государственный университет» по адресу: 394006, г. Воронеж, Университетская пл., 1, ауд. 226.

С диссертацией можно ознакомиться в библиотеке и на сайте ФГБОУ ВПО «Воронежский государственный университет», http://www.science.vsu.ru

Автореферат разослан «___» мая 2014 г.

Ученый секретарь
диссертационного совета — Воронина И.Е.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Использование компьютерных технологий привело к пониманию важности задач, связанных с обработкой накопленной информации для извлечения знаний. Все более востребованным становится создание систем интеллектуального анализа данных, позволяющих выявить полезную скрытую информацию на основе классификации, кластеризации, статистического анализа, поиска ассоциативных правил и других подходов. Если данные представлены в виде динамических рядов каких-то показателей или их можно сгенерировать из базы данных, то для восстановления закономерностей используется техника регрессионного моделирования, при этом подразумевается, что данные являются числовыми. Однако, если информация относится к другому типу, например, является частично или полностью приближенной, то классические методы регрессионного анализа не применимы, и этот факт обусловливает необходимость их модификации. Одним из способов формализации приближенной информации является использование понятия нечеткого множества и его частного случая – нечеткого числа. Задача разработки регрессионных моделей, ориентированных на нечеткие числа, решалась зарубежными (H.Tanaka, P. Diamond, D. Dubois, M.S. Yang, M. Sakawa, M. Albrecht) и отечественными (Р.А. Алиев, А.Э. Церковный, Г.А. Мамедова, Н.Г. Ярушкина и др.) учеными. В общем случае методы нечеткого регрессионного моделирования могут быть разделены на две группы: первая базируется на методе наименьших квадратов и его модификациях, а вторая − на линейном программировании. Анализ показал, что рассмотрены далеко не все возможные постановки задач, учитывающих нечеткость исходных данных и/или параметров модели, кроме того, во многих исследованиях отсутствует комплексность подхода к реализации всех этапов регрессионного моделирования. Построение нечетких регрессионных моделей опирается на математический аппарат, включающий определение арифметических операций над нечеткими числами и их сравнение. Только для некоторых типов нечетких чисел результат арифметической операции представляет собой нечеткое число того же типа. В других случаях требуется дополнительная аппроксимация. Необходимость совершенствования существующих методов нечеткого регрессионного моделирования за счет учета различных типов данных и параметров, представленных нечеткими числами *L*-*R*-типа, и их реализации в рамках информационной системы интеллектуального анализа данных обусловливает актуальность диссертационного исследования.

Диссертационная работа выполнена в соответствии с одним из основных научных направлений Воронежского государственного университета «Математическое моделирование, программное и информационное обеспечение, методы вычислительной и прикладной математики и их применение к фундаментальным исследованиям в естественных науках».

Объект исследования – информационная система интеллектуального анализа данных, в которой реализуются нечеткие линейные регрессионные модели с коэффициентами в виде нечетких чисел *L*-*R*-типа.

Предмет исследования – нечеткий линейный регрессионный анализ на множестве нечетких чисел *L*-*R*-типа.

Цель диссертационной работы заключается в развитии подходов к решению задачи восстановления закономерностей в данных на основе нечеткого регрессионного моделирования.

Для достижения поставленной цели решаются следующие **задачи:**

1. Анализ существующих подходов к восстановлению закономерностей в данных на основе регрессионного моделирования и выявление путей их совершенствования на случай приближенной исходной информации.

2. Выявление алгебраических свойств операций над нечеткими числами *L*-*R*-типа и разработка теоретической основы нечеткого регрессионного моделирования.

3. Определение оценок параметров нечетких регрессионных моделей и модификация общей процедуры регрессионного моделирования для выявления закономерностей в приближенной информации.

4. Разработка программного комплекса с применением современных компьютерных технологий для анализа и интеллектуальной обработки данных на основе предложенных алгоритмов нейросетевого и нечеткого регрессионного моделирования.

Методы исследования. В диссертационной работе использовались методы нечеткого и нейросетевого моделирования, линейной алгебры, математической статистики, теории нечетких множеств и нечеткой арифметики, объектно-ориентированного и модульного программирования.

Основные результаты, выносимые на защиту, и их научная новизна:

1) совокупность теоретических результатов, касающихся свойств арифметических операций над нечеткими числами *L*-*R*-типа и существования алгебраических структур, что позволяет осуществлять вычисления при построении нечетких регрессионных моделей;

2) модификация процедуры регрессионного моделирования для восстановления закономерностей в данных, отличающаяся оценками параметров нечетких линейных (парной и множественной) регрессионных моделей, в которых коэффициенты представлены обычными и/или нечеткими числами *L*-*R*-типа;

3) альтернативные подходы к выявлению множества существенных независимых переменных в рамках нечеткого регрессионного моделирования, основанные на нечетком коэффициенте корреляции, стандартизированном уравнении нечеткой множественной линейной регрессии и применении автоассоциативных нейронных сетей, «работающих» с приближенной информацией, что обеспечивает комплексность анализа данных на различных этапах процесса выявления закономерностей;

4) структура программного комплекса, включающая блок нечеткой арифметики, который может использоваться как самостоятельное приложение, и средства для проведения нечеткого линейного регрессионного моделирования, а также основанная на ней информационная система интеллектуального анализа данных, функционал которой ориентирован на выявление закономерностей в данных, содержащих частично или полностью приближенную информацию.

Область исследования. Диссертационная работа соответствует следующему пункту Паспорта специальности 05.13.17 «Теоретические основы информатики»: п. 5. «Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях...».

Практическая значимость работы. Разработанная информационная система, в которой реализован предложенный комплекс алгоритмов нейросетевого анализа и нечеткого регрессионного моделирования, предназначена для обработки приближенной информации, выявления в ней функциональных зависимостей и проведения исследований в ситуациях, когда традиционные методы неприменимы. Результаты диссертационной работы используются для оценки качества выпущенной продукции с целью обоснования управленческих решений по совершенствованию технологических процессов специалистами ЗАО ЛЦ «АВС Фарбен», а также в учебном процессе ФГБОУ ВПО «Воронежский государственный университет» при чтении спецкурсов и выполнении курсовых и выпускных квалификационных работ.

Апробация работы. Основные результаты, полученные в диссертационной работе, докладывались и обсуждались на следующих международных и всероссийских конференциях: Международная научно-практическая конференция «Глобальная научная интеграция» (Тамбов, 2011); Международная конференция «Актуальные проблемы прикладной математики, информатики и механики» (Воронеж, 2011-2012); Всероссийская молодежная научная школа «Инженерия знаний. Представление знаний: состояние и перспективы» (Воронеж, 2012); Международная конференция «ExploIT Dynamics PhD Workshop» (Германия, г. Ольденбург, 2012); Международная конференция «Современные методы прикладной математики, теории управления и компьютерных технологий» (Воронеж, 2013); Международный научный семинар «Emerging Trends in Informations Systems (IS)» (Нижний Новгород, 2013).

Публикации. Основные результаты диссертации опубликованы в 12 научных работах, в том числе 5 – в изданиях, рекомендованных ВАК РФ. В работах, опубликованных в соавторстве, лично соискателю принадлежат: [1] – метод оценки параметров нечеткой линейной множественной регрессионной модели, анализ данных; [10] – детальная разработка и наполнение шагов нечеткого парного линейного регрессионного анализа.

Объём и структура работы. Диссертация состоит из введения, четырёх глав, заключения, списка использованных источников из 110 наименований,

двух приложений. Основная часть работы изложена на 151 странице и включает 42 рисунка и 17 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснованы актуальность темы, научная новизна и значимость работы, приведены цели и задачи исследования.

В **первой** главе рассмотрены информационные системы интеллектуального анализа данных, технология их построения и архитектура; приведена классификация задач интеллектуального анализа данных; представлены существующие подходы к восстановлению закономерностей на основе регрессионного моделирования; рассмотрен подход к отбору наиболее информативных признаков для проведения множественного регрессионного анализа данных.

Во **второй** главе рассмотрены теоретические основы обработки приближенной информации, представленной нечеткими числами *L*-*R*-типа. В основу исследования положены арифметические операции, определенные A. Piegat. В диссертации исследованы свойства введенных операций, построены алгебраические структуры на множестве нечетких чисел *L*-*R*-типа.

В общем виде нечеткие числа задаются с помощью *L*-*R*-функций, при этом *L*-функция – это непрерывная, неубывающая функция $L:\Re\to[0,1]$, которая удовлетворяет следующим дополнительным условиям: $\lim\limits_{x\to\infty}L(x)=0$ и существует значение $x'\in\Re$, такое, что $L(x)=1$. Функция $R(x)$ обладает аналогичными свойствами. Функции $L(x)$ и $R(x)$ описывают изменение функции принадлежности нечеткого числа на промежутках неопределенности.

Число A есть унимодальное нечеткое число *L*-*R*-типа, если существуют константы $\alpha,\beta>0$, такие, что функция принадлежности нечеткого числа A имеет вид:

$$m_a(x)=\begin{cases}L\left(\dfrac{\mu-x}{\alpha}\right), \text{если } x\le\mu,\\ R\left(\dfrac{x-\mu}{\beta}\right), \text{если } x\ge\mu.\end{cases}$$

где α и β – соответственно левый и правый коэффициенты нечеткости, μ – модальное значение нечеткого числа. Условно нечеткое число A обозначается тройкой параметров (μ_a,α_a,β_a).

Нечеткое число A считается положительным, если его модальное значение положительно, у отрицательного нечеткого числа мода отрицательна.

В диссертации исследованы свойства арифметических операций над нечеткими числами *L*-*R*-типа: коммутативность, ассоциативность, дистрибутивность, наличие нейтрального и обратного элементов. Установлено, что нейтральным элементом для операции сложения нечетких чисел является обычное число 0, но обратного элемента не существует. Для

операции умножения нейтральным элементом является обычное число 1, а обратный элемент для заданного числа A определяется формулой $\overline{A} = \left(\frac{-1}{\mu_a}, \frac{\beta_a}{\mu_a(\mu_a + \beta_a)}, \frac{\alpha_a}{\mu_a(\mu_a - \alpha_a)} \right)$.

Пусть $Fuz^{LR}(\Re)$ – семейство нечетких чисел *L*-*R*-типа, определенных на множестве действительных чисел $\Re$, $*$ – арифметическая операция над нечеткими числами, тогда пара $\langle Fuz^{LR}(\Re), * \rangle$ образует нечеткий группоид.

В диссертации исследованы свойства следующих нечетких группоидов: 1) $\langle Fuz^{LR}(\Re), * \rangle$; 2) $\langle Fuz^{LR}(\Re), + \rangle$; 3) $\langle Fuz^{LR}(\Re), \times \rangle$. Полученные результаты представлены в таблице 1.

Таблица 1– Алгебраические структуры с одной операцией

Свойства\Группоид	1	1	2	3	3	3
Коммутативность		+				+
Ассоциативность	+	+	+	+	+	+
Единичный элемент 0			+			
Единичный элемент 1				+	+	+
Обратный элемент					+	+
Алгебраическая структура	**полу-группа**	**абелева полугруппа**	**моноид**		**группа**	**абелева группа**

Таким образом, множество нечетких чисел *L*-*R*-типа образует абелеву группу по умножению и является моноидом по сложению.

В диссертации получены формулы для вычисления некоторых выражений над нечеткими числами, которые в дальнейшем используются в рамках нечеткого регрессионного моделирования.

В **третьей** главе рассмотрена задача оценки параметров нечеткой линейной регрессионной модели (парной и множественной), которая учитывает различные типы данных в выборке: а) четкие независимые переменные, нечеткие коэффициенты и зависимая величина; б) нечеткие зависимая и независимые переменные модели с четкими коэффициентами. Проведена оценка адекватности и точности моделей и предложены методы отбора существенных независимых переменных.

Пусть исходные данные представлены выборкой $\{(x_i, Y_i)\}_{i=\overline{1,n}}$, где $x_i \in \Re$, а $Y_i = (\mu_{yi}, \alpha_{yi}, \beta_{yi})$ – нечеткие числа *L*-*R*-типа. Предполагается, что *линейная парная регрессионная модель с нечеткими коэффициентами* имеет вид

$$Y_i = B_0 + B_1 x_i + E_i, i = \overline{1,n}, \quad (1)$$

где $x_i \in \Re$, $Y_i = (\mu_{yi}, \alpha_{yi}, \beta_{yi})$ – нечеткие числа *L*-*R*-типа, $B_0 = (\mu_{b0}, \alpha_{b0}, \beta_{b0})$ и $B_1 = (\mu_{b1}, \alpha_{b1}, \beta_{b1})$ – теоретические коэффициенты регрессии, $E_i = (\mu_{ei}, \alpha_{ei}, \beta_{ei})$ – отклонения в виде нечетких чисел *L*-*R*-типа наблюдаемых значений от значений, полученных на основе модели, $i = \overline{1,n}$ – номер наблюдения.

Для получения коэффициентов регрессии $\tilde{B}_0 = (\tilde{\mu}_{b0}, \tilde{\alpha}_{b0}, \tilde{\beta}_{b0})$ и $\tilde{B}_1 = (\tilde{\mu}_{b1}, \tilde{\alpha}_{b1}, \tilde{\beta}_{b1})$ составляется оценочная модель

$$\tilde{Y}_i = \tilde{B}_0 + \tilde{B}_1 x_i, i = \overline{1,n}, \quad (2)$$

где $x_i \in \Re$ – наблюдаемые значения независимой переменной, $\tilde{Y}_i = (\tilde{\mu}_{yi}, \tilde{\alpha}_{yi}, \tilde{\beta}_{yi})$ – оценки значений зависимой переменной.

В соответствии с методом наименьших квадратов оценки коэффициентов регресии находятся из решения задачи минимизации функции расстояния между нечеткими переменными Y_i и $\tilde{Y}_i$

$$F(\tilde{B}_0, \tilde{B}_1) = \sum_{i=1}^{n} (\tilde{Y}_i - Y_i)^2 = \sum_{i=1}^{n} \begin{pmatrix} (\tilde{\mu}_{b0} + \tilde{\mu}_{b1} x_i - \mu_{yi})^2 + (\tilde{\alpha}_{b0} + \tilde{\alpha}_{b1} x_i - \alpha_{yi})^2 + \\ + (\tilde{\beta}_{b0} + \tilde{\beta}_{b1} x_i - \beta_{yi})^2 \end{pmatrix} \to \min$$

Необходимым условием существования минимума этой функции является равенство нулю ее частных производных по переменным $\tilde{\mu}_{b0}, \tilde{\alpha}_{b0}, \tilde{\beta}_{b0}$ и $\tilde{\mu}_{b1}, \tilde{\alpha}_{b1}, \tilde{\beta}_{b1}$. С учетом алгебраических свойств операций над нечеткими числами получены следующие формулы для оценок параметров регрессии:

$$\begin{cases} \tilde{\mu}_{b1} = \dfrac{\overline{x\mu_y} - \bar{\mu}_y \bar{x}}{\overline{x^2} - \bar{x}^2} = \dfrac{\operatorname{cov}(x, \mu_y)}{\sigma_x^2}, \tilde{\mu}_{b0} = \bar{\mu}_y - \tilde{\mu}_{b1} \bar{x}, \\ \tilde{\alpha}_{b1} = \dfrac{\overline{x\alpha_y} - \bar{\alpha}_y \bar{x}}{\overline{x^2} - \bar{x}^2} = \dfrac{\operatorname{cov}(x, \alpha_y)}{\sigma_x^2}, \tilde{\alpha}_{b0} = \bar{\alpha}_y - \tilde{\alpha}_{b1} \bar{x}, \\ \tilde{\beta}_{b1} = \dfrac{\overline{x\beta_y} - \bar{\beta}_y \bar{x}}{\overline{x^2} - \bar{x}^2} = \dfrac{\operatorname{cov}(x, \beta_y)}{\sigma_x^2}, \tilde{\beta}_{b0} = \bar{\beta}_y - \tilde{\beta}_{b1} \bar{x}. \end{cases} \quad (3)$$

Адекватность построенной модели проводится на основе анализа построенного в диссертации *нечеткого коэффициента корреляции*

$$r = \left(\tilde{\mu}_{b1} \frac{\sigma_x}{\sigma_\mu}, \tilde{\alpha}_{b1} \frac{\sigma_x}{\sigma_\alpha}, \tilde{\beta}_{b1} \frac{\sigma_x}{\sigma_\beta} \right) = \left(\frac{\operatorname{cov}(x, \mu_y)}{\sigma_x \sigma_\mu}, \frac{\operatorname{cov}(x, \alpha_y)}{\sigma_x \sigma_\alpha}, \frac{\operatorname{cov}(x, \beta_y)}{\sigma_x \sigma_\beta} \right).$$

Для практических расчетов приведены модификации этой формулы.

Оценка точности модели осуществляется путем вычисления средней относительной ошибки $E_{отн} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{Y_i - \tilde{Y}_i}{Y_i} \right| \cdot 100\%$.

В диссертации также предложена обобщенная *нечеткая линейная множественная регрессионная модель с нечеткими коэффициентами*

$$Y_i = A_0 + A_1 x_{i1} + A_2 x_{i2} + \ldots + A_n x_{in} + E_i, i = \overline{1,m}, \quad (4)$$

где $x_{ik} \in R, k = \overline{1,n}$, $Y_i = (\mu_{yi}, \alpha_{yi}, \beta_{yi})$, $i = \overline{1,m}$ – нечеткие числа *L*-*R*-типа; $A_j = (\mu_{aj}, \alpha_{aj}, \beta_{aj})$, $j = \overline{0,n}$ – нечеткие регрессионные параметры; $E_i = (\mu_{ei}, \alpha_{ei}, \beta_{ei}), i = \overline{1,m}$ – случайные ошибки, нечеткие числа *L*-*R*-типа; $i = \overline{1,m}$ – номер наблюдения, $k = \overline{1,n}$ – номер независимой переменной.

С помощью метода наименьших квадратов найдены оценки $\tilde{A}_j = (\tilde{\mu}_{aj}, \tilde{\alpha}_{aj}, \tilde{\beta}_{aj}),\ j = \overline{0,n}$, параметров A_j в матричной форме

$$\tilde{\mu}_a = (X^T X)^{-1} X^T \mu_y, \tilde{\alpha}_a = (X^T X)^{-1} X^T \alpha_y, \tilde{\beta}_a = (X^T X)^{-1} X^T \beta_y. \quad (5)$$

Под качеством нечеткой линейной множественной регрессионной модели подразумевается адекватность модели, которая оценивается на основе анализа остаточной последовательности $\{E_i = Y_i - \tilde{Y}_i, i = \overline{1,n}\}$, и точность модели, определяемая путем вычисления средней относительной ошибки $E_{отн} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \tilde{Y}_i}{Y_i}\right| \cdot 100\%$. Для обеспечения применимости классических методов оценки адекватности осуществляется переход от нечетких остатков к их дефазифицированным значениям.

Для того, чтобы коэффициенты регрессии в модели (4) выражались в сравнимых единицах измерения, построено *стандартизованное уравнение нечеткой линейной множественной регрессии*

$$T_y = B_1 t_{x1} + B_2 t_{x2} + \ldots + B_p t_{xp},\ p = \overline{1,n}, \quad (6)$$

где нечеткие коэффициенты $B_p = \left(\frac{\mu_{ap} \cdot \sigma_{xp}}{\sigma_\mu}, \frac{\alpha_{ap} \cdot \sigma_{xp}}{\sigma_\alpha}, \frac{\beta_{ap} \cdot \sigma_{xp}}{\sigma_\beta}\right)\ \left(p = \overline{1,n}\right)$ являются коэффициентами регрессии в стандартизированном масштабе, $T_y = \left(\frac{\mu_y - \bar{\mu}_y}{\sigma_\mu}, \frac{\alpha_y - \bar{\alpha}_y}{\sigma_\alpha}, \frac{\beta_y - \bar{\beta}_y}{\sigma_\beta}\right)$ и $t_{xp} = \frac{\left(x_p - \bar{x}_p\right)}{\sigma_{xp}},\ p = \overline{1,n}$ – соответствующие значения зависимой и независимых переменных регрессионной модели в стандартизованном виде. Коэффициенты B_p позволяют выявить наиболее существенные независимые переменные, оказывающие влияние на зависимую величину. Это приводит к сокращению количества независимых переменных, участвующих в модели.

В диссертационной работе предложен метод оценки неизвестных параметров *нечеткой множественной линейной регрессионной модели с четкими коэффициентами*

$$Y_i = a_0 + a_1 X_{i1} + \ldots + a_n X_{in} + E_i,\ i = \overline{1,m}, \quad (7)$$

где $Y_i = (\mu_{yi}, \alpha_{yi}, \beta_{yi}), i = \overline{1,m}$ – значения зависимой переменной, нечеткие числа *L-R*-типа; $E_i = (\mu_{ei}, \alpha_{ei}, \beta_{ei}), i = \overline{1,m}$ – случайные ошибки, нечеткие числа *L-R*-типа, $i = \overline{1,m}$ – номер наблюдения; $X_{ik} = (\mu_{xik}, \alpha_{xik}, \beta_{xik})$ – значения независимой переменной, $k = \overline{1,n}$ – номер независимой переменной.

В результате найдена следующая формула оценок параметров регрессии

$$\tilde{a} = (\mu_x^T \mu_x + \alpha_x^T \alpha_x + \beta_x^T \beta_x)^{-1}\left(\mu_x^T \mu_y + \alpha_x^T \alpha_y + \beta_x^T \beta_y\right). \quad (8)$$

Для решения задачи отбора независимых переменных модели (7) предложен подход, основанный на использовании *автоассоциативной нейронной сети*, принцип работы которой был адаптирован для анализа

информации в нечеткой среде. Сеть содержит три слоя нейронов: входной и выходной слой, а также средний слой – «узкое горло», который в результате обучения выдает сжатое представление данных (вектор Z). Число выходов n совпадает с числом входов, а внутренний слой содержит меньшее количество нейронов $m < n$.

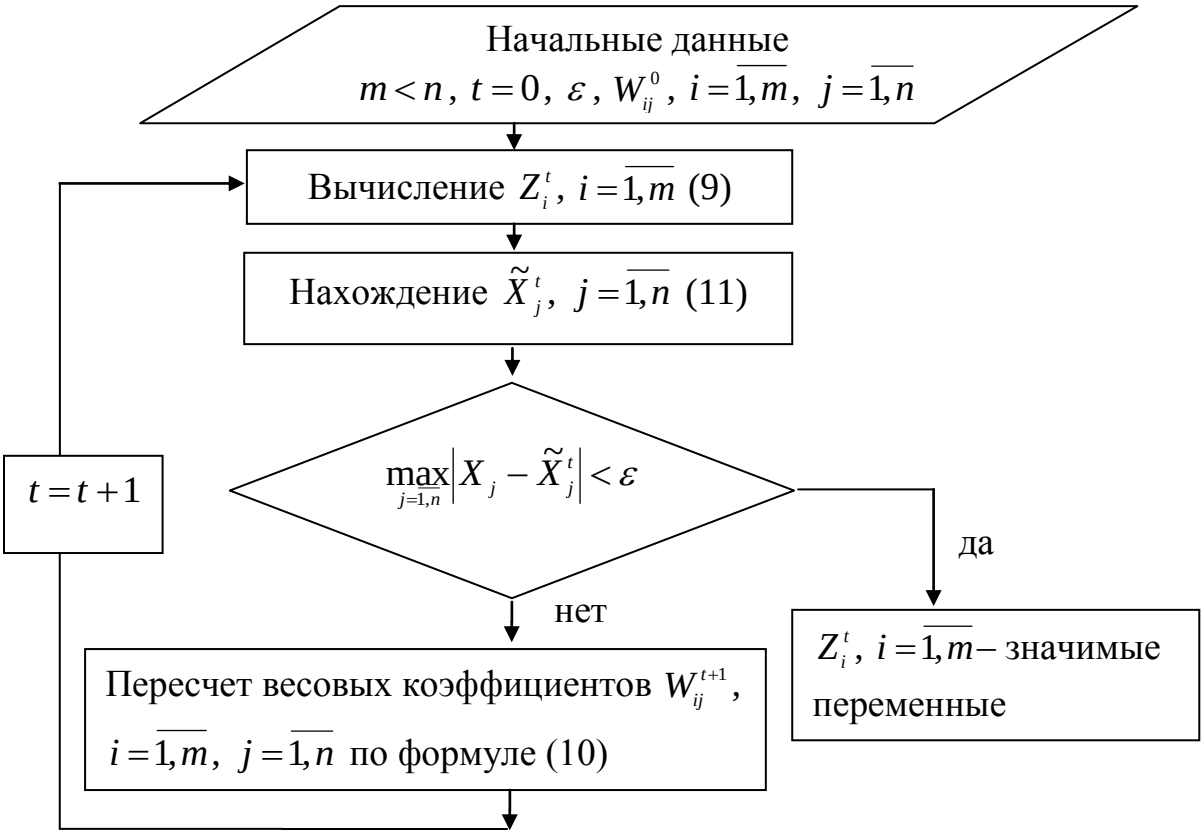

Рисунок 1 – Схема обучения автоассоциативной нейронной сети

Обучение нейронной сети заключается в следующем (рисунок 1): пусть на наборе n-мерных данных обучается m линейных нейронов, выходное значение каждого из этих нейронов в случае нечетких данных вычисляется по формуле

$$Z_i=(\mu_{zi},\alpha_{zi},\beta_{zi})=\sum_{j=1}^{n}W_{ij}X_j=\sum_{j=1}^{n}(\mu_{wij},\alpha_{wij},\beta_{wij})\cdot(\mu_{xj},\alpha_{xj},\beta_{xj}),\ i=\overline{1,m},\qquad(9)$$

тогда, согласно правилу обучения Ойя, весовые коэффициенты изменяются следующим образом:

$$W_{ij}^{t+1}=W_{ij}^{t}+\Delta W_{ij}^{t}=(\mu_{wij}^{t}+\Delta\mu_{wij}^{t},\alpha_{wij}^{t}+\Delta\alpha_{wij}^{t},\beta_{wij}^{t}+\Delta\beta_{wij}^{t}),\ i=\overline{1,m},\ j=\overline{1,n},\qquad(10)$$

где

$$\Delta W_{ij}^{t}=\eta Z_i^t\left(X_j^t-\tilde{X}_j^t\right)=\eta(\mu_{zi}^t,\alpha_{zi}^t,\beta_{zi}^t)\left((\mu_{xj}^t,\alpha_{xj}^t,\beta_{xj}^t)-\sum_{k=1}^{m}(\mu_{zk}^t,\alpha_{zk}^t,\beta_{zk}^t)\cdot(\mu_{wkj}^t,\alpha_{wkj}^t,\beta_{wkj}^t)\right).$$

Сеть самообучается на воспроизведение входов – то есть ответ нейросети считается правильным, когда значения сигналов на каждом выходе совпадает со значением на соответствующем ему входе ($X_i=\tilde{X}_i$).

Нейроны выходного слоя являются линейными с тождественной функцией активации

$$\tilde{X}_j = \left(\tilde{\mu}_{xj}, \tilde{\alpha}_{xj}, \tilde{\beta}_{xj}\right) = \sum_{k=1}^{m} Z_k W_{kj} = \sum_{k=1}^{m} \left(\mu_{zk}, \alpha_{zk}, \beta_{zk}\right) \cdot \left(\mu_{wkj}, \alpha_{wkj}, \beta_{wkj}\right),\ j = \overline{1,n}. \quad (11)$$

Таким образом, сеть с узким горлом из скрытых линейных нейронов обучается воспроизводить на выходе значения своих входов. Скрытый слой такой сети при этом осуществляет оптимальное кодирование входных данных и содержит максимально возможное при данных ограничениях количество информации.

В **четвёртой** главе предложен подход к восстановлению закономерностей с использованием информационной системы интеллектуального анализа данных (ИАД), разработана структура информационного хранилища, рассмотрены функции системы администрирования, приведена структура программного комплекса, разработанного в среде программирования CodeGear Borland C++ Builder и предназначенного для проведения нечеткого регрессионного моделирования.

Структура программного комплекса представлена на рисунке 2. К его основным *функциональным возможностям* относятся: реализация калькулятора нечетких чисел для осуществления различных арифметических операций над нечеткими числами *L*-*R*-типа и построения графиков их функций принадлежности; выполнение отбора существенных независимых переменных на основе автоассоциативных нейронных сетей; проведение нечеткого линейного парного и множественного регрессионного анализа с нахождением коэффициентов модели и средней ошибки вычислений; построение стандартизированного уравнения нечеткой множественной линейной регрессии.

Система интеллектуального анализа данных основана на технологии информационного хранилища. Ее структура (рисунок 3) предусматривает наличие двух приложений – аналитического (основного) и системы администрирования (вспомогательного). Последнее предназначено для выполнения SQL-запросов к базам данных информационной системы при участии аналитика.

Информация о формах и переходах содержится в специальной базе данных ИС, которая может быть локальной или удаленной. На рисунке 4 приведена физическая модель информационного хранилища в виде таблиц сущностей, которые взаимосвязаны между собой.

Сущность «Показатель» включает описание экономических, технических и иных показателей, которые необходимы для проведения аналитической работы. Они имеют иерархическую структуру. Сущность «Единица измерения» создана для хранения данных в единой форме и содержит информацию об используемых измерениях, коэффициент пересчета между которыми представлен атрибутом «Множитель». Сущность «Данные» содержит значения показателей по каждому измерению в виде трех атрибутов для внесения модального значения, левого и правого коэффициентов

неопределенности нечеткого числа. Кроме того, она ссылается на идентификатор показателя, его тип, период действия и дату актуальности, что позволяет при необходимости иметь информацию о значениях определенных показателей в более детальном виде.

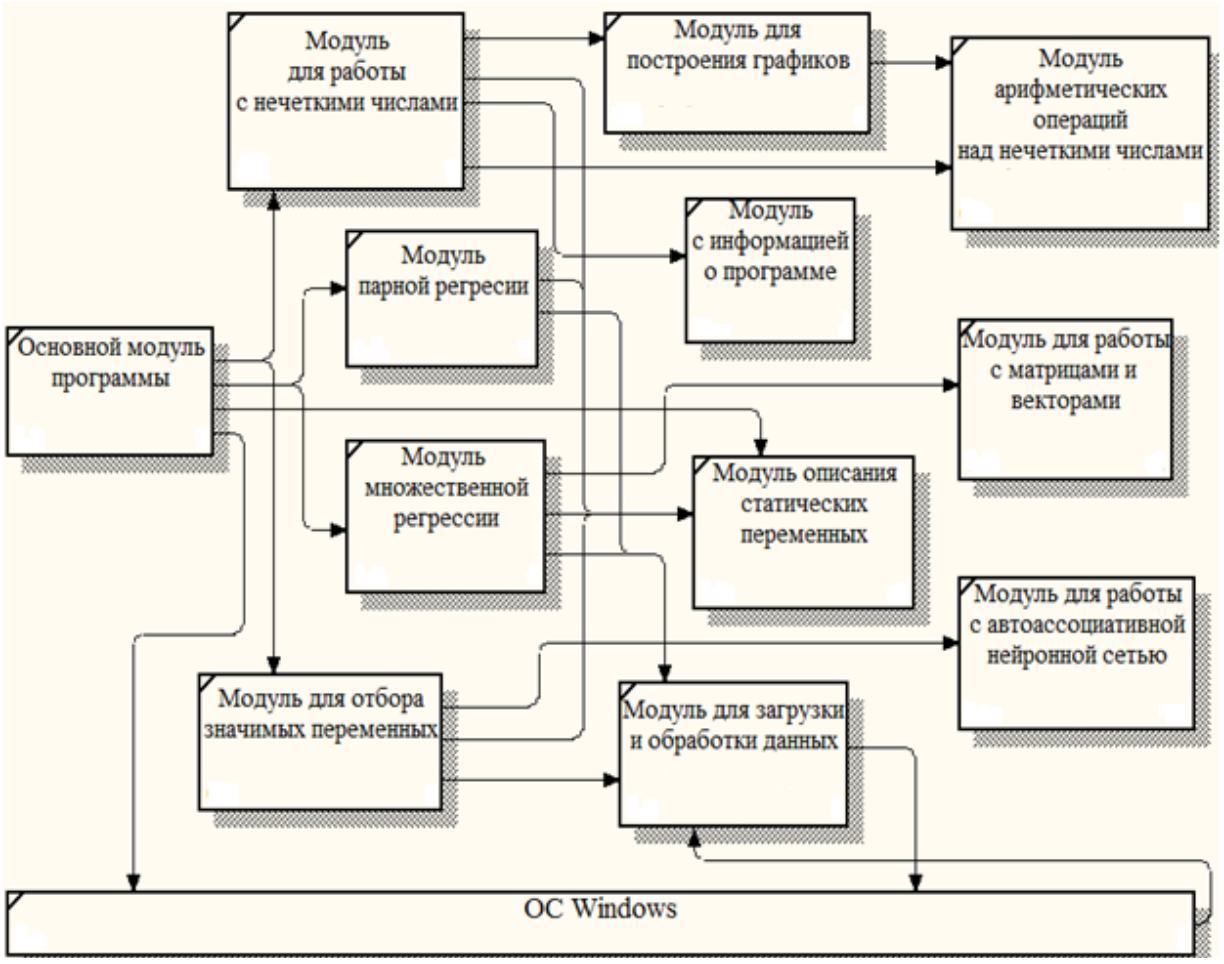

Рисунок 2 – Структура ПО нечеткого регрессионного моделирования

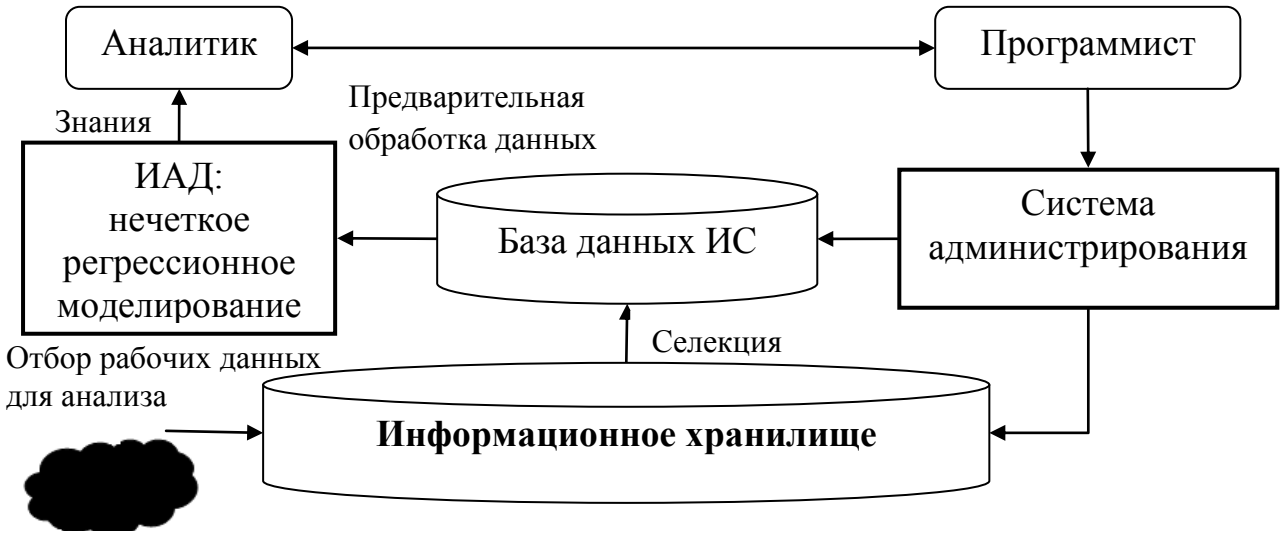

Рисунок 3 – Структура информационной системы ИАД

Разработанная информационная система использовалась для анализа данных по выпускаемой лакокрасочной продукции и проведения оценки качества товаров с целью принятия управленческих решений по совершенствованию технологических процессов.

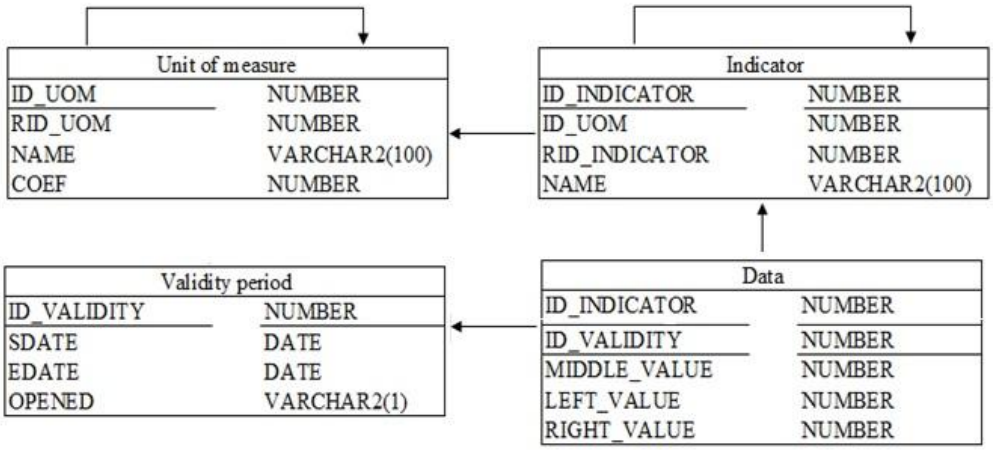

Рисунок 4 – Физическая модель информационного хранилища

Процесс интеллектуального анализа данных состоял из следующих этапов:

1. Подготовка данных, которая заключалась в предварительной обработке информации для проведения нечеткого регрессионного моделирования. В таблице 2 представлены исходные данные, в которых в качестве показателей (независимых переменных) выступают проценты наполнителей, растворителей и связующих веществ эмалированной краски, а результирующая величина (зависимая переменная) представляет собой примерное качество выпускаемой продукции и принадлежит множеству нечетких чисел *L*-*R*-типа.

Таблица 2 – Исходные данные

x_1	11	10	11	12	13,5	14	15	16	17	17,5	19	20
x_2	10	10,5	12,5	12	13	13,5	14	16	14,5	15	17	16,5
x_3	12	12,5	13	14,5	16	16,5	17	18	19,5	20,5	21	22
μ_y	11,2	12,5	12,9	14,1	14,8	16,1	17,5	18,9	18,9	20	21,1	22,2
α_y	0,9	0,1	1	0,5	1,1	1,2	0,1	0,5	0,3	1,3	1,1	0,2
β_y	0,2	0,4	0,5	1,1	0,7	0,1	0,08	1,2	0,7	0,48	1,9	0,4

В результате обучения автоассоциативной нейронной сети с одним нейроном среднего слоя было получено сжатое представление независимых переменных: вектор $x = (11, 11.5, 13, 14, 15, 15.5, 16, 17, 18, 19, 21, 22)$.

2. Проведение нечеткого линейного регрессионного моделирования, в ходе которого было получено уравнение парной регрессии $\tilde{Y} = (0.6, 0.6, 0.01) + (1, 0.004, 0.04)x$ (таблица 3).

3. Проверка построенной модели, включающая оценку ее качества путем нахождения средней погрешности вычислений $\varepsilon = (0.4, 0.4, 0.3)$.

Таблица 3 – Полученные значения результирующей переменной

$\tilde{\mu}_y$	11,5	12,1	13,6	14,6	15,6	16,1	16,6	17,6	18,6	19,6	21,6	22,6
$\tilde{\alpha}_y$	0,65	0,65	0,65	0,66	0,66	0,66	0,67	0,67	0,68	0,68	0,68	0,69
$\tilde{\beta}_y$	0,45	0,47	0,53	0,57	0,61	0,63	0,65	0,69	0,73	0,77	0,85	0,89

По каждому наблюдению $i=\overline{1,12}$ был построен γ-срез для точного и регрессионного результирующего значения: $[\mu_y-\alpha_y(1-\gamma),\mu_y+\beta_y(1-\gamma)]$ и $[\tilde{\mu}_y-\tilde{\alpha}_y(1-\gamma),\tilde{\mu}_y+\tilde{\beta}_y(1-\gamma)]$ соответственно (рисунок 5). Точность вычисления в зависимости от выбранного γ-среза приведена на рисунке 6, где $\tilde{\varepsilon}=\left(\sum_{i=1}^{12}\left|Y-\tilde{Y}\right|\right)/12$ при $Y=\mu_y+0{,}5(1-\gamma)(\beta_y-\alpha_y)$, $\tilde{Y}=\tilde{\mu}_y+0{,}5(1-\gamma)(\tilde{\beta}_y-\tilde{\alpha}_y)$. Выявлено, что погрешность уменьшается при увеличении значения γ.

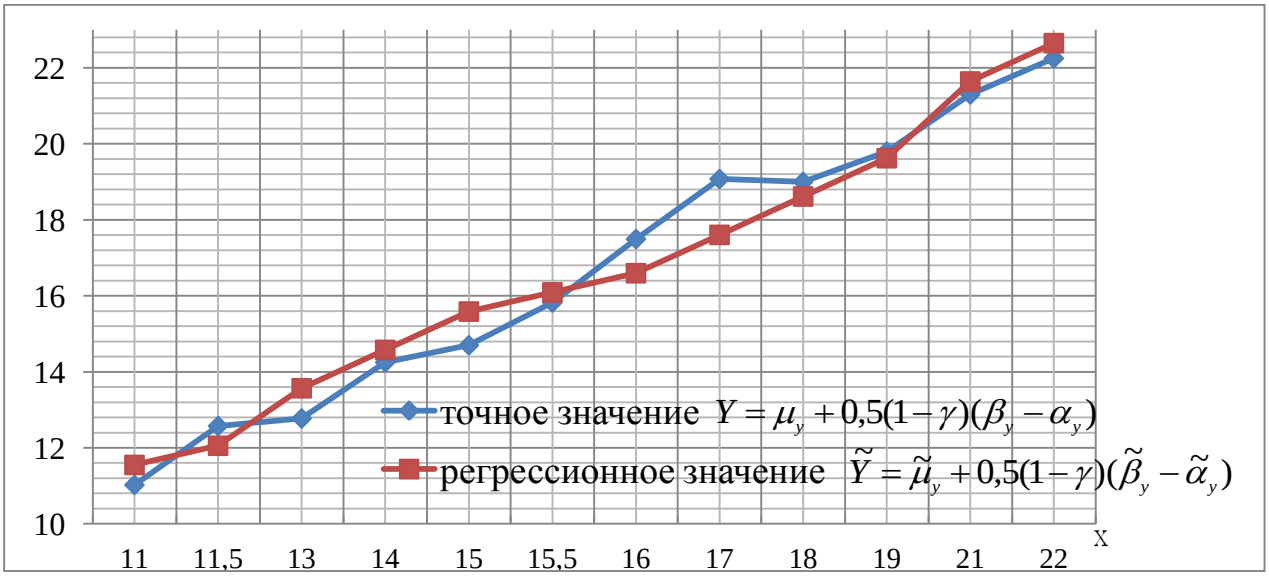

Рисунок 5– График регрессии на 0.5-срезе

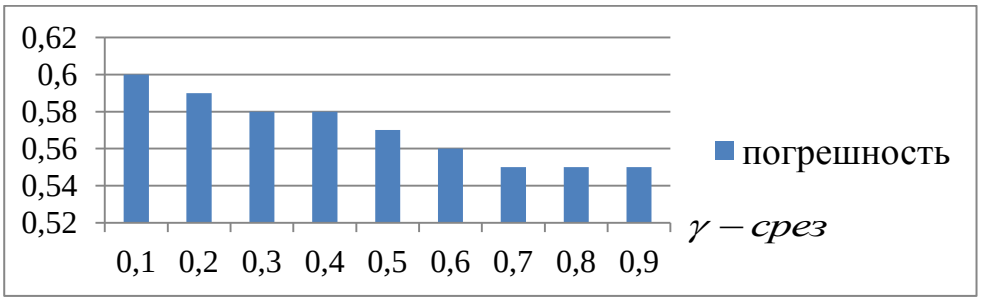

Рисунок 6 – Погрешность вычислений $\tilde{\varepsilon}$ в зависимости от γ-среза

4. Интерпретация проведенного моделирования, результаты которого были проанализированы и использованы при формировании бюджета закупки сырья.

В **заключении** излагаются основные результаты исследований и вычислительного эксперимента.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. Проанализированы существующие подходы к восстановлению закономерностей в данных и определены направления модификации методов нечеткого регрессионного моделирования для восстановления в данных, содержащих частично или полностью приближенную информацию.

2. Выявлены свойства арифметических операций над нечеткими числами *L-R*-типа и соответствующие алгебраические структуры, построены выражения для некоторых формул с нечеткими числами, которые позволяют проводить вычисления в ходе нечеткого регрессионного моделирования.

3. На основе операций над нечеткими числами предложены оценки параметров нечеткой парной и множественной линейной регрессионной модели с помощью метода наименьших квадратов, выведены формулы для нахождения точности модели и определена процедура оценки адекватности построенной модели.

4. Предложены альтернативные подходы к формированию множества существенных переменных для множественной регрессионной модели, основанные на нечетком коэффициенте корреляции, стандартизированном уравнении нечеткой множественной линейной регрессии и применении автоассоциативных нейронных сетей, принцип функционирования которых был обобщен на нечеткую информацию.

5. Разработан программный комплекс для интеллектуального анализа данных, включающий в качестве инвариантной составляющей блок нечеткой арифметики и средства для проведения нечеткого линейного регрессионного моделирования.

Публикации по теме исследования

Публикации в изданиях, рекомендованных ВАК РФ

1. Сапкина Н.В. Нечеткая множественная линейная регрессионная модель для симметричных нечетких чисел L-R-типа / Т.М. Леденева, Н.В. Сапкина // Современная экономика: проблемы и решения: науч.-практ. журнал. – Воронеж: ИПЦ ВГУ, 2011. – № 10. – С. 174-181.
2. Сапкина Н.В. Применение сети Эльмана для задачи прогнозирования изменения курса ценных бумаг / Н.В. Сапкина // Системы управления и информационные технологии. – Москва-Воронеж: ИПЦ «Научная книга», 2011. – № 2.1 (44). – С. 169-172.
3. Сапкина Н.В. Свойства операций над нечеткими числами / Н.В. Сапкина // Вестник ВГУ. Серия Системный анализ и информационные технологии. – Воронеж: ИПЦ ВГУ, 2013. – №1. – С. 23-28.
4. Сапкина Н.В. Нечеткая парная линейная регрессия и корреляция / Н.В. Сапкина // Современная экономика: проблемы и решения: науч.-практ. журнал. – Воронеж: ИПЦ ВГУ, 2013. – № 10 (46). – С. 178-189.
5. Сапкина Н.В. Нечеткая линейная множественная регрессионная модель с четкими коэффициентами. Отбор значимых переменных модели с помощью нейросетей / Н.В. Сапкина // Системы управления и информационные технологии. – Москва-Воронеж: ИПЦ «Научная книга», 2013. – №4 (54). – С. 27-30.

Свидетельства о государственной регистрации программы для ЭВМ

6. Сапкина Н.В. Отбор наиболее значимых факторов с помощью нейронной сети / Н.В. Сапкина // Свидетельство о гос. регистрации программы для ЭВМ №2013660206, РФ, 2013.

7. Сапкина Н.В. Реализация нечеткого множественного линейного регрессионного анализа / Н.В. Сапкина // Свидетельство о гос. регистрации программы для ЭВМ №2013660211, РФ, 2013.

Статьи и материалы конференций

8. Сапкина Н.В. Прогнозирование курса акции с помощью вероятностной нейронной сети и средств технического анализа / Н.В. Сапкина // Актуальные проблемы прикладной математики, информатики и механики: сб. тр. междунар. конф., Воронеж, 22-24 июня 2009 г. – Воронеж: ИПЦ ВГУ, 2009. – Ч. 2. – С. 156-161.

9. Сапкина Н.В. Применение нейронной сети Эльмана для прогнозирования курса акций / Н.В. Сапкина // Актуальные проблемы прикладной математики, информатики и механики: сб. тр. междунар. конф., Воронеж, 20-22 сент. 2010 г. – Воронеж: ИПЦ ВГУ, 2010. – С. 319-325.

10. Сапкина Н.В. Нечеткие линейные регрессионные модели. Метод наименьших квадратов для модели с четкими входами и гауссовым нечетким выходом / Н.В. Сапкина // Глобальная научная интеграция: сб. материалов междунар. науч.-практ. конф., Тамбов, 30 июня 2011 г. – Тамбов: ТМБпринт, 2011. – С. 68-71.

11. Сапкина Н.В. Метод наименьших квадратов для нечеткой линейной регрессионной модели / Н.В. Сапкина // Актуальные проблемы прикладной математики, информатики и механики: сб. тр. междунар. конф., Воронеж, 26-28 сентября 2011 г. – Воронеж: ИПЦ ВГУ, 2011. – С. 344-345.

12. Сапкина Н.В. Нечеткий парный линейный регрессионный анализ / Н.В. Сапкина, А.А. Татаринцев // Инженерия знаний. Представление знаний: состояние и перспективы: материалыВсероссийской молодежной научной школы, Воронеж, 29-30 июня 2012 г. – Воронеж: ИПЦ «Научная книга», 2012. – С. 260-261.

13. Сапкина Н.В. Нечеткий парный линейный регрессионный анализ / Н.В. Сапкина // Актуальные проблемы прикладной математики, информатики и механики: сб. тр. междунар. конф., Воронеж, 26-28 ноября 2012 г. – Воронеж: ИПЦ ВГУ, 2012.– Ч. 1. – С. 331-334.

14. Сапкина Н.В. Свойства группоида нечетких чисел LR-типа / Н.В. Сапкина // Современные методы прикладной математики, теории управления и компьютерных технологий: Сборник трудов VI Международной конференции, Воронеж, 10-16 сентября 2013г. – Воронеж: ИПЦ ВГУ, 2013. – С. 216-218.