На правах рукописи

ПОЛУПАНОВ Дмитрий Васильевич

МАТЕМАТИЧЕСКИЕ МОДЕЛИ РАНЖИРОВАНИЯ ОБЪЕКТОВ НАЛОГОВОГО КОНТРОЛЯ

Специальность 05.13.18
Математическое моделирование,
численные методы и комплексы программ

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Уфа 2007

Работа выполнена на кафедре вычислительной математики
Башкирского государственного университета
и на региональной кафедре математики и информатики филиала
Всероссийского заочного финансово-экономического института в г. Уфе

Научный руководитель: д-р тех. наук, проф.
ГОРБАТКОВ Станислав Анатольевич

Официальные оппоненты: д-р тех. наук, проф.
ЧЕРНЯХОВСКАЯ Лилия Рашитовна

канд. тех. наук, доцент
ЗОЗУЛЯ Юрий Иванович

Ведущая организация: **Институт математики с ВЦ**
Уфимского научного центра РАН

Защита состоится 25 мая 2007 г. в 10^{00} часов
на заседании диссертационного совета Д-212.288.03
в Уфимском государственном авиационном техническом университете
по адресу: 450000, г. Уфа, ул. К. Маркса, 12

С диссертацией можно ознакомиться в библиотеке университета

Автореферат разослан апреля 2007 г.

Ученый секретарь
диссертационного совета,
д-р тех. наук, проф. В.В. Миронов

ПОЛУПАНОВ Дмитрий Васильевич

МАТЕМАТИЧЕСКИЕ МОДЕЛИ РАНЖИРОВАНИЯ ОБЪЕКТОВ НАЛОГОВОГО КОНТРОЛЯ

Специальность 05.13.18
Математическое моделирование,
численные методы и комплексы программ

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы

Существующие принципы налогообложения и методики контроля правильности исчисления и уплаты налогов часто не позволяют выявить реальную налогооблагаемую базу, не в полной мере обеспечивают решение фискальных задач и реализацию принципа справедливости налоговой системы. Повышение уровня объективности и эффективности работы налоговых органов в условиях существенного искажения данных налоговых деклараций, дефицита наблюдений и т.д. требует совершенствования технологий налогового контроля с использованием современных инструментариев математического моделирования и искусственного интеллекта, например таких, как нейронные сети, вероятностные методы оценки риска.

Вопросам управления налогообложением в аспекте моделирования процессов сбора налогов и оценки добросовестности отдельных налогоплательщиков посвящены работы А.Б. Паскачева, Т.Н. Скорика, А.Б. Соколова, Д.Г. Черника. Проблемы интеллектуального управления и нейросетевого моделирования экономических объектов исследованы в трудах отечественных ученых В.И. Васильева, А.А. Ежова, Б.Г. Ильясова, Л.А. Исмагиловой, С.Т. Кусимова, С.А. Терехова, С.А. Шумского, Л.Р. Черняховской, Н.И. Юсуповой, зарубежных ученых И.С. Абу-Мустафы, Д.-Э. Бэстенса, В.-М. ван ден Берга, Д. Вуда. Теоретические и прикладные аспекты нейросетевого моделирования налогового контроля рассмотрены в ряде работ Н.Д. Бублика, Г.И. Букаева, И.И. Голичева, С.А. Горбаткова, А.Н. Романова. Общие вопросы теории нейронных сетей и нейрокомпьютинга изложены в работах А.И. Галушкина, А.Н. Горбаня, В.Л. Дунина-Барковского, Г.Г. Малинецкого, Э. Баррона, А.Г. Ивахненко, Т. Кохонена, Ф. Розенблата, С. Хайкина и других ученых нашей страны, ближнего и дальнего зарубежья.

Однако несмотря на имеющиеся разработки в области нейросетевого моделирования для стохастических объектов с сильнозашумленными данными, в частности объектов налогового контроля, методы и принципы построения эффективных, адекватных, качественных нейросетевых математических моделей (НСМ) разработаны не в полном объеме. Уровень объективности оценок в существующих информационных технологиях налогового контроля не соответствуют запросам практики и потенциальным возможностям современного математического аппарата. Учитывая вышеизложенное, актуальной научной задачей является разработка гибридных нейросетевых моделей (ГНСМ), служащих основой синтеза плана выездных проверок.

Цель работы – разработка научных основ информационной технологии ранжирования объектов налогового контроля при синтезе плана отбора налогоплательщиков для проведения выездных проверок.

Задачи исследования

1. Исследование возможности нейросетевой аппроксимации многомерных функциональных зависимостей в условиях сильного зашумления данных (и даже частичного сознательного их искажения) и дефицита наблюдений.

2. Разработка концепции построения эффективных, адекватных ГНСМ на основе общесистемных закономерностей кибернетики, разработка методов предпроцессорной обработки данных и оценки адекватности ГНСМ.

3. Разработка рабочего алгоритма ранжирования экономических объектов с сильнозашумленными данными на основе ГНСМ.

4. Построение прикладных ГНСМ ранжирования объектов налогового контроля, экспериментальная апробация и верификация ГНСМ.

Методы исследования

Работа основана на положениях и методах функционального анализа, положениях общей теории систем, методах теории нейросетевого моделирования, классических методах теории вероятности и математической статистики.

На защиту выносятся

1 Метод синтеза плана отбора налогоплательщиков для проведения выездных проверок на основе ГНСМ.

2. Метод предпроцессорной обработки данных, разработанный на основе системного подхода, который позволяет обеспечить приемлемый уровень достоверности получаемых оценок при сильном искажении базы данных (БД). Данный метод основывается на предложении об управлении качеством НСМ на ранних стадиях ее построения и включает в себя процедуры оптимальной кластеризации БД и очистки кластера от аномальных наблюдений по векторному критерию точности, устойчивости и детерминированности.

3. Вероятностный критерий ранжирования объектов налогового контроля по числовой мере искажения ими отчетной документации с внесением в него эвристической априорной информации, полученной на основе использования доверительных интервалов для отклонений между расчетными (полученными с помощью НСМ) и декларированными значениями моделируемого показателя, что позволяет повысить достоверность процедуры ранжирования.

4. Метод модифицированного обобщенного перекрестного подтверждения (МОПП) ГНСМ по финишному критерию совпадения множеств проранжированных налогоплательщиков для нескольких независимых НСМ с заданной доверительной вероятностью. Метод МОПП служит основным инструментом анализа и подтверждения адекватности ГНСМ.

5. Рабочий алгоритм ранжирования стохастических объектов с сильнозашумленными данными, который применительно к ранжированию объектов налогового контроля по числовой мере искажения ими отчетной документации служит инструментарием принятия решений о включении налогоплательщика в план проведения выездных проверок.

Научная новизна работы

1. Новизна метода синтеза плана отбора налогоплательщиков для проведения выездных проверок заключается в использовании «эталона» - производственной функции кластера налогоплательщиков, полученной с помощью ГНСМ. Это позволяет выявлять нарушения в налоговых декларациях и получать объективные оценки финансового состояния налогоплательщиков путем извлечения знаний об искаженных входных факторах и выходной величины через другие, неискаженные.

2. Новизна метода предпроцессорной обработки данных заключается в процедурах управления качеством НСМ на ранних стадиях ее построения путем многоуровневого иерархического структурирования модели. Процедура оптимальной кластеризации увязана с качеством обучения НСМ, что позволяет структурировать БД, повышая ее однородность. Процедура очистки образованных кластеров увязывает удаление аномальных наблюдений с качеством обучения НСМ. Очистка БД для построения НСМ от аномальных наблюдений по критерию точности, первоначально предложенная совместно с Г.А. Бесхлебновой [3], дополнена введением критериев устойчивости и детерминированности [18], [5]. Предложенная процедура позволяет увеличить однородность данных внутри об-

разованных кластеров. В целом предложенный метод позволяет получить НСМ с приемлемыми аппроксимативными свойствами для сложных условий моделирования (сильное зашумление БД вплоть до ее сознательного искажения, отягченное дефицитом наблюдений, неконтролируемой внутренней структурой объекта и др.).

3. Новизна критерия ранжирования объектов налогового контроля заключается в вероятностном принципе ранжирования, что позволяет учитывать эвристическую априорную информацию, предысторию и масштаб деятельности налогоплательщика. Предложенный критерий позволяет получить план выездных проверок в аспекте ожидаемых доначислений.

4. Новизна метода МОПП заключается в сравнении множеств проранжированных налогоплательщиков для независимых моделей, основанных на НСМ различных типов, отличающихся числом скрытых слоев нейросети (НС), числом нейронов в них, видом активационных функций с заданной доверительной вероятностью. Это позволяет оценить адекватность построенных ГНСМ в условиях нарушения предпосылок регрессионного анализа.

5. Новизна рабочего алгоритма ранжирования стохастических объектов с сильнозашумленными данными состоит в том, что в него введены дополнительные процедуры итерационного взаимодействия традиционных операций обучения и тестирования НС с операциями предобработки данных и обеспечения адекватности.

Практическая значимость работы

Полученные в диссертационной работе результаты могут быть использованы для решения практических задач ранжирования сложных стохастических объектов с сильнозашумленными данными. В частности, результаты ранжирования объектов налогового контроля могут служить основой производственного плана выездных проверок.

Результаты диссертационного исследования, в том числе технология математического моделирования по созданию НСМ аппроксимации производственной функции и вероятностной модели ранжирования (ВМР) объекта налогового контроля в специфических условиях, могут быть также использованы и для более широкого класса задач, не рассматриваемых в диссертации (прогнозирование экономических показателей налогоплательщика и оптимизация его финансового состояния, оценка ожидаемой суммы доначислений, ранжирование корпоративных заемщиков при предоставлении им кредитов, оптимальное бюджетирование муниципальных образований при ограничении бюджетных средств региона и др.).

Апробация работы и публикации

Основные положения диссертации докладывались на следующих научных конференциях: Международной научной конференции «Математические модели и методы их исследования», Красноярск, 1999 г.; Международной научной конференции «Моделирование, вычисления, проектирование в условиях неопределенности», Уфа, 2000 г.; Шестой Международной научно-технической конференции студентов и аспирантов «Радиоэлектроника, электротехника и энергетика», Москва, 2000 г.; Республиканской конференции студентов и аспирантов по математике, Уфа, 2000 г.; Международных научных конференциях «Континуальные логико-алгебраические и нейросетевые методы, 2000 и 2001», Ульяновск; Региональной школе-конференции для студентов, аспирантов и молодых ученых по математике и физике, Уфа, 2001 г.; Втором, Третьем, Пятом, Шестом и Седьмом Всероссийских симпозиумах по прикладной и промышленной математике (2001-206 г.г.); VIII Всероссийской конференции «Нейрокомпьютеры и их применение» НКП-2002 с международным участием, Москва, 2002 г.; VIII и XI Всероссийских научно-технической конференциях «Нейроинформатика – 2006 и 2007», Москва; V Всероссийской научно-

практической конференции «Проблемы и перспективы российской экономики», Пенза, 2006 г.; Международной научно-практической конференции «Современные направления теоретических и прикладных исследований», Одесса, 2006 г.

Основное содержание диссертации отражено в 22 опубликованных работах общим объемом 16,56 п.л. в том числе автора 8,12 п.л., из них 5 публикаций в рецензируемых журналах из списка ВАК.

Структура и объем работы

Диссертация состоит из введения, четырех глав, заключения, списка используемой литературы из 124 наименований, 2 приложений и содержит 171 страниц основного текста, 29 рисунков, 22 таблицы.

Благодарности

Автор благодарит директора филиала Всероссийского заочного финансово-экономического института в г. Уфе, д-ра экон. наук, проф. Н.Д. Бублика и д-ра физ.-мат. наук, проф. И.И. Голичева за ценные советы по обсуждению работы.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обосновывается актуальность темы исследования, сформулированы основные результаты, выносимые на защиту, с обоснованием их новизны, достоверности, теоретической ценности, практической значимости.

В **первой главе** проводится исследование возможности нейросетевой аппроксимации многомерных функциональных зависимостей в условиях искажения данных и дефицита наблюдений применительно к объектам налогового контроля. Исследуются существующие технологии налогового контроля, традиционные способы отбора налогоплательщиков для выездных налоговых проверок. Уделяется внимание статистическим методам отбора налогоплательщиков. Делается вывод о существенном влиянии субъективного фактора в существующих технологиях налогового контроля.

Ставится следующая задача ранжирования объектов налогового контроля. Пусть выделена примерно однородная группа из G объектов налогового контроля (налогоплательщиков). За ретроспективный промежуток времени T имеется БД $Z = \left\langle \vec{X}, Y \right\rangle_i, i = \overline{1, N}$; $N = G \cdot T$, составленная на основе налоговых деклараций и бухгалтерской отчетности. Требуется выбрать небольшое число существенных признаков – входных факторов $\vec{X} = (X_1, \ldots, X_n)$, таких как сумма основных средств, себестоимость, среднесписочная численность сотрудников, коммерческие расходы и др., а также выходную моделируемую величину Y, в качестве которой может выступать выручка предприятия, и построить некоторую достаточно информативную НСМ, связывающую входные и выходные величины

$$\hat{y} = F(\vec{x}, W(\vec{x}, y)), x \in \mathrm{X} \subset \Re^n, y \in \mathrm{Y} \subset \Re, \hat{y} \in \hat{\mathrm{Y}} \subset C[\Re] \ , \qquad (1)$$

где $\vec{x} = (x_1, \ldots, x_n)$ – конкретная численная реализация случайного вектора входных факторов $\vec{X}$; $y \in \mathrm{Y} \subset \Re$ – декларируемая налогоплательщиком конкретная числовая реализация наблюдаемой выходной случайной величины Y; $\hat{y}$ – эталон – расчетное значение величины Y; $\{W\}$ – множество оцениваемых синаптических весов НС; X – множество значений вектора входных факторов, Y – множество декларированных значений выходной величины, $\hat{\mathrm{Y}}$ – множество расчетных по (1) значений выходной величины.

Под «достаточной информативностью» НСМ понимаются ее дискриминантные свойства, т.е. возможность надстройки модели (1) некоторым функционалом Ψ от $\hat{y}$, который

бы позволял выявлять в сильно зашумленной и даже искаженной БД нарушителей налогового законодательства с требуемым уровнем доверительной вероятности. С дискриминантными свойствами связана основная системообразующая экономико–математическая концепция предлагаемого подхода к построению модели ранжирования налогоплательщиков, основанная на двух предложениях.

Первое предложение: нарушения в налоговой декларации эффективнее выявляются не путем автономного анализа отдельно взятого налогоплательщика, как это делается в существующих методиках, а путем сравнения производственных функций достаточно однородного кластера налогоплательщиков. Сравнительный анализ реализуется путем порождения *эталонного* (среднего для кластера) значения оценки моделируемой производственной функции или «*фона*» $\hat{y}(x)$ с помощью модели (1) и вычисления для всех объектов налогового контроля отклонений в каждом наблюдении с номером *i* на момент оценки

$$\delta_i = \hat{y}_i - y_i \,/\, y_i . \qquad (2)$$

Второе предложение состоит в вероятностном принципе ранжирования налогоплательщиков. Строится ВМР на основе вероятностного критерия

$$\psi_g = \tilde{\tilde{\delta}}_{gt}\Big|_{t=t^{TC}} \cdot P\,(\Delta_g \geq \bar{\bar{\delta}}) \cdot M_g , \qquad (3)$$

где $\tilde{\tilde{\delta}}_{gt} = \delta_{gt} + U$ – значение верхней границы доверительного интервала для отклонения δ_{gt} ($\delta_{gt} \equiv \delta_i$, в записи отклонения фиксируется номер налогоплательщика g и момент наблюдения t); $P(\Delta_g \geq \bar{\delta})$ – вероятность того события, что ожидаемое значение отклонения Δ_g моделируемой случайной величины $\hat{y}$ будет не меньше выборочного среднего $\bar{\delta}$ с учетом его смещения на полуширину доверительного интервала для δ; в момент времени t^{TC} осуществляется ранжирование налогоплательщика, т.е. это – момент проверки, соответствующий последнему кварталу подачи декларации налогоплательщиком; M_g – экспертно задаваемый коэффициент масштаба *g*-го налогоплательщика.

Во взаимодействии НСМ и ВМР строится ГНСМ ранжирования объектов налогового контроля, схематично представленная на рисунке 1.

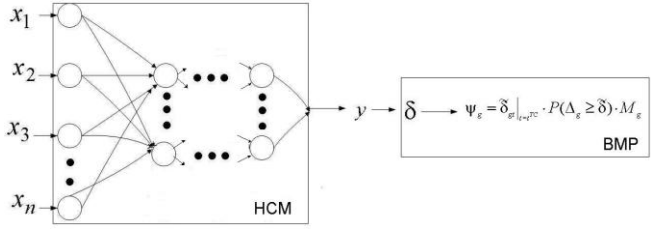


Рисунок 1. ГНСМ ранжирования объектов налогового контроля

На основе этих идей осуществлена формализованная запись ГНСМ ранжирования экономических объектов с сильнозашумленными данными в виде

$$\Theta = F_4 \circ F_3 \circ F_2 \circ F_1 (\vec{X}, \vec{Y}) , \qquad (4)$$

где $F_3 \circ F_2 \circ F_1$ – композиция операторов НСМ аппроксимации производственной функции экономического объекта, представляющая в конечном счете оператор F в формуле (1); F_4 – оператор ВМР; Θ – множество проранжированных на основе вероятностного критерия (3) экономических объектов. Вид операторов, составляющих модель (4) и сущность множества Θ раскрываются во второй главе диссертации (см. ниже формулы (7)-(9) и (16)).

Осуществлен анализ условий моделирования. Показано, что для объектов налогового контроля характерно сознательное искажение данных налоговых деклараций, изменчивость внутренней структуры налогоплательщиков, стохастическое влияние вешней среды, существенная связь входных факторов.

Исследован вопрос о влиянии взаимной стохастической зависимости компонент вектора входных факторов на качество обучения НСМ. Отмечено, что для НСМ она не является критичной, в отличие от регрессионных моделей, получаемых с помощью метода наименьших квадратов.

Исследована проблема, связанная с аппроксимацией функции многих переменных с помощью НС в специфических условиях моделирования.

Исследованы проблемы, связанные с устойчивостью НСМ по возмущению входных данных. Определена чувствительность НСМ к возмущениям входных данных при искажении обучающего множества НС. Величину, характеризующую меру интенсивности возмущения определим как: $\mu^{(k)} = \left\|A^{(k)} - A\right\| / \|A\|$. Здесь $A = \{a_{ij}\}_{i=\overline{1,N_{learn}},\ j=\overline{1,n+1}}$ – матрица наблюдений, составляющих обучающее множество НС, $a_{ij} = x_{ij}, j = \overline{1,n}$, $a_{in+1} = y_i$, $\|A\| = \sqrt{\sum_{i=1}^{N_{learn}} \sum_{j=1}^{n+1} a_{ij}}$, $\tilde{A}^{(k)} = \{\tilde{a}_{ij}^{(k)}\}_{i=\overline{1,N_{learn}},\ j=\overline{1,n+1}}$ – возмущенное обучающее множество, $\tilde{a}_{ij}^{(k)} = a_{ij} + \xi_{ij}^{(k)}, j = \overline{1,n}$, $\tilde{a}_{in+1}^{(k)} = a_{in+1} + \eta^{(k)}{}_i, i = \overline{1,N_{learn}}$. Возмущения – случайные величины, распределенные по нормальному закону: $\xi_j^{(k)} \sim N(k \cdot \bar{X}_j; k \cdot S_{X_j})$, $j = \overline{1,n}$, $\eta^{(k)} \sim N(k \cdot \bar{Y}; k \cdot S_Y)$. Параметр $k \in Q$ характеризует интенсивность возмущений, множество параметров Q определяется произвольно, например, $Q = \{0,5;1;1,5;2;2,5;3\ldots\}$.

Определим ошибку обобщения НСМ как

$$E^{(k)} = \left\|\hat{y}^{(k)} - \hat{y}\right\| / \|y\|, \quad \|y\| = \sqrt{\sum_{i=1}^{N_{test}} y_i^2} \tag{5}$$

где $\hat{y}^{(k)}$ – значение выходной величины, соответствующее мере интенсивности возмущений $\mu^{(k)}$

Определение: *Будем говорить, что отображение (1) устойчиво при возмущении обучающего множества в смысле ошибки обобщения (5), если для мер интенсивности возмущений $\mu^{(l)}, \mu^{(p)}$ существует константа $K > 0$ такая, что имеет место априорная оценка*: $\left|E^{(l)} - E^{(p)}\right| \le K\left|\mu^{(l)} - \mu^{(p)}\right|$. *При этом выполнено условие*

$$\left|\left(E^{(l+1)} - E^{(l)}\right) / \left(\mu^{(l+1)} - \mu^{(l)}\right) - \left(E^{(l)} - E^{(l-1)}\right) / \left(\mu^{(l)} - \mu^{(l-1)}\right)\right| \le \varepsilon, \varepsilon > 0, \tag{6}$$

где ε *- сколь угодно малое число.*

Утверждение: *Существует величина μ_{cr} такая, что если $\mu^{(k)} \le \mu_{cr}$, то отображение (1) устойчиво при возмущении обучающего множества, т.е. выполнено условие (6). Если $\mu^{(k)} > \mu_{cr}$, то условие (6) нарушается и НСМ не устойчива.*

Достоверность данного утверждения численно обоснована на модельном примере при варьировании меры интенсивности возмущений, числа искаженных строк и столбцов матрицы *A*. На оси абсцисс рисунка 2 указано значение $\mu^{(k)}$, на оси ординат – рассчитанное по (5) $E^{(k)}$. Из рисунка 2А следует, что при $\mu^{(k)} > 2{,}1526$ свойства устойчивости НСМ в

указанном выше смысле теряются. При варьировании доли числа искаженных строк матрицы $\tilde{A}^{(k)}$, в случае k=0,5 (рисунок 2Б) при $\mu^{(k)} > 0{,}4060$, угол наклона кривой возрастает на два порядка, модель теряет свои аппроксимативные свойства. Аналогично, при варьировании доли искаженных столбцов (рисунок 2В) по достижению $\mu^{(k)} > 0{,}2924$, угол наклона кривой возрастает на два порядка, НСМ теряет устойчивость.

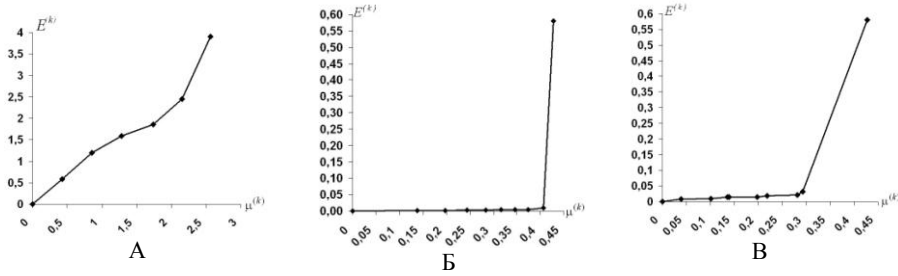


Рисунок 2. Исследование зависимости $E(\mu^{(k)})$ при варьировании меры интенсивности возмущений $\mu^{(k)}$ (А), числа искаженных строк (Б) и столбцов (В) матрицы $\tilde{A}^{(k)}$.

Во **второй главе** разрабатываются метод предпроцессорной обработки данных для построения ГНСМ ранжирования объектов налогового контроля (3) на основе системного подхода, который позволяет получить эффективные модели для сложных условий моделирования, и метод МОПП оценки адекватности ГНСМ. На базе указанных методов с использованием общесистемных закономерностей кибернетики предложена концепция построения ГНСМ для решения задач ранжирования объектов налогового контроля. Предлагаются и обосновываются 1) управление качеством НСМ на ранних стадиях ее построения путем многоэтапного структурирования модели на основе общесистемной закономерности роста и убывания энтропии; 2) использование общесистемной фоновой закономерности для повышения однородности исходной БД; 3) построение вероятностного критерия ранжирования налогоплательщиков, основанного на общесистемных закономерностях асимметрии и неполного подавления побочных дисфункций структурирования информационной системы, который позволяет получить план выездных проверок в аспекте ожидаемых доначислений; 4) метод МОПП оценки адекватности ГНСМ по финишному критерию совпадения множеств проранжированных налогоплательщиков для нескольких независимых моделей с заданной доверительной вероятностью.

Из общесистемного закона роста и убывания энтропии в открытых системах следует, что энтропия открытой системы может быть уменьшена только в том случае, если она взаимодействует с другими системами. Следовательно, при структурировании БД в НСМ нужно ввести негэнтропию (информацию) с помощью специальных способов предобработки данных, реализующих многоуровневое иерархическое структурирование модели, схема которого представлена на рисунке 3.

На трех иерархических уровнях структурирования модели реализуются специальные способы предобработки данных, повышающие однородность БД и улучшающие качество обучения НСМ.

Рисунок 3. Многоуровневое структурирование модели

Нулевой уровень структурирования модели, который является основой предложенной технологии построения НСМ, – это образование структуры из рационально взаимодействующих элементов – данных деклараций налогоплательщиков. Достигаемый эффект – получение производственной функции (1) кластера налогоплательщиков, причем с активной эксплуатацией свойств нелинейной взаимосвязи сознательно искаженных факторов.

На *первом* иерархическом уровне структурирования предложена оптимизационная итерационная процедура кластеризации исходной БД, которая, в отличие от традиционных методов кластеризации, увязана с качеством обучения НСМ. Итогом процедуры является образование в исходной БД оптимального числа достаточно однородных кластеров. Получаемый синергетический эффект данного уровня - создание предпосылок получения НСМ хорошего качества при сложных условиях моделирования.

Оператор оптимальной кластеризации имеет вид

$$F_1 : Z \to \bigcup_{q=1}^{Q^*} Z_q \,, \tag{7}$$

где $Z_q = \left\langle \vec{X}, \vec{Y} \right\rangle_i, i = \overline{1, N_q}$ – БД q-го кластера, Q^* – оптимальное число кластеров.

На *втором* иерархическом уровне предложена оптимизационная итерационная процедура очистки кластера исходных данных от аномальных наблюдений по векторному критерию точности, устойчивости и детерминированности вспомогательных НСМ (субмоделей) каждого кластера, основанная на общесистемной фоновой закономерности. В отличие от традиционных методов устранения аномальных наблюдений, никак не связанных с дальнейшим обучением моделей, предложено увязывать эти процедуры, поскольку удаляемые аномальные точки имеют разную информативность в аспекте обучения НСМ, так как БД сильно искажена.

Оператор очистки кластера вводится следующим образом

$$F_2 : Z_q \to Z_q^* \,, \tag{8}$$

где $Z_q^* = \left\langle \vec{X}, Y \right\rangle_i, i = \overline{1, N_{k^*}}$ – БД очищенного кластера, k^* – номер оптимальной итерации.

Наконец *третий*, последний уровень структурирования – это определение оптимальной архитектуры НСМ, вида активационных функций и параметров обучения. Процедуры третьего уровня завершаются построением рабочей модели на БД однородного очищенного кластера. На данном этапе можно получить эффективную НСМ только в том случае, если успешно реализованы предшествующие первый и второй уровни структурирования. Оператор рабочей НСМ следующий

$$F_3 : Z_q \to \hat{Y} \,, \tag{9}$$

где $Z_q, q = \overline{1, Q^*}$ – БД q-го кластера, $\hat{y} = f^*(\vec{x}, (W(\vec{x}, y))$ – расчетное значение выходной величины на основе рабочей НСМ, полученной на оптимальной итерации очистки q-го кластера.

Итерационная процедура оптимальной кластеризации заключается в следующем. Требуется найти оптимальное число кластеров из условия:

$$Q^*: \left[\left(\min_{d_{i,k}} \sum_{q=1}^{Q} \sum_{i,k}^{N_q} d_{i,k}^2 \right) \cap \left(\min_{Q} \max_{q} E_{\max}^{(q)}(Q, d_{i,k}) \right) \right] \tag{10}$$

при ограничениях на число наблюдений в кластере

$$(N_q / n) \geq \xi \tag{11}$$

и критическое значение ошибки обобщения

$$E_{\max}^{(q)} \geq E^*. \tag{12}$$

Критерий $E^{(q)} = \left\| \hat{y}^{(q)} - y \right\| / \left\| y \right\|$ – ошибка обобщения НСМ. Она вводится аналогично (5) и учитывает суммарный вклад взаимосвязанных факторов в обучение НСМ Критерий d – евклидовы расстояния между элементами в кластере – учитывает общность элементов по масштабу и условиям их хозяйственной деятельности.

Поисковый алгоритм решения многокритериальной задачи оптимизации строится как итерационный процесс пошагового увеличения числа кластеров. На каждом Q-ом шаге итерации при фиксированном числе кластеров Q методом k-средних образуются кластеры и, соответственно, минимизируется критерий плотности расположения элементов. Затем для каждого из образованных q кластеров строятся НСМ (субмодели) и вычисляется ошибка обобщения $E^{(q)}$ по (5). Строится кривая $E_{\max}^{(q)}$ как функция от числа кластеров Q. Итерационный процесс останавливается по двум правилам: 1) либо ошибка обобщения достигает минимума и на следующем шаге итерации $Q+1$ начинает расти; 2) либо нарушается условие (9).

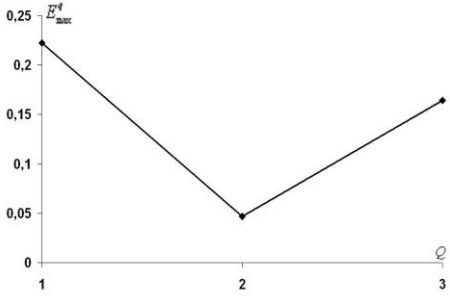


Рисунок 4. Зависимость максимальной ошибки обобщения $E_{\max}^{q}$ от номера итерации Q.

Справедливость вышеизложенного утверждения обоснована численно на примере модели зависимости выручки от шести входных факторов.

Строилась НС типа многослойный персептрон (MLP) с двумя скрытыми слоями, активационной функцией сигмоид $f(s) = (1 + \exp(-as))^{-1}, a > 0$ в них, в выходном – линейной. Значение максимальной ошибки обобщения на каждой итерации кластеризации представлено на рисунке 4. Найдено оптимальное число кластеров 2, $E_{\max}^{q} = 0.0476$.

Итерационная процедура очистки образованных кластеров от аномальных наблюдений заключается в следующем: требуется найти номер оптимальной итерации очистки кластера

$$k^*: \left(\left(\min_{k} \Phi^{(k)} \right) \cap \left(\delta_i^{(k)} \geq \varepsilon^{(k)}, i = \overline{1, N_k} \right) \right), \tag{13}$$

при ограничениях, которые вводятся аналогично (11), (12)

$$(N_k / n) \geq \xi, \tag{14}$$

$$E^{(k)} \geq E^{*}, \qquad (15)$$

Итогом процедуры является устранение аномальных наблюдений в каждом кластере, для которых относительное значение отклонения $\delta_i^{(k)} = \left|(\hat{y}^{(k)}{}_i - y_i)/y_i\right| \cdot 100\% > \varepsilon^{(k)}$, (где $\varepsilon^{(k)}$ – экспертно задаваемый верхний предел приемлемого уровня погрешности) с определением оптимальной итерации в условиях дефицита наблюдений. Итерационный процесс, который, по сути, является численным методом оптимального сглаживания данных в кластере, останавливается по правилу: а) либо нарушается условие репрезентативности выборки в данном кластере (13); б) либо достигается минимум обобщенного критерия Ф; в) либо достигается допустимый уровень ошибки обобщения – выполняется условие (14).

Обобщенный критерий Ф, представляет собой линейную свертку из трех частных критериев: $\Phi = C_1 E + C_2 S + C_3 R$. Здесь C_m – экспертно назначаемые весовые коэффициенты, характеризующие вклад в процесс оптимизации каждого частного критерия $\sum_{m=1}^{3} C_m = 1; C_m \geq 0$. Критерий $E^{(k)} = \left\|\hat{y}^{(k)} - y\right\| / \left\|y\right\|$ характеризует точность субмоделей, т.е. является ошибкой обобщения, аналогичной (5). Критерий S характеризует устойчивость субмоделей, он вводится как аналог константы Липшица $S = \left|\hat{y}_\alpha - \hat{y}_\beta\right| / \left\|\vec{x}_\alpha - \vec{x}_\beta\right\|_{R^n}$ где векторы $\vec{x}_\alpha, \vec{x}_\beta$ близки по норме в R^n, $\hat{y}_\alpha = F(\vec{x}_\alpha, W), \hat{y}_\beta = F(\vec{x}_\beta, W)$ - расчетные значения l-го компонента выходной величины (1) в точках наблюдений α, β; Ω^{test} – множество индексов $\{i\}$, которым соответствуют наблюдения, вошедшие в тестовое множество НС; α, β – конкретные значения индексов. Критерий R определен по аналогии с коэффициентом детерминации, $R = 1 - r^2_{y,\hat{y}^{(k)}}$, где $r_{y,\hat{y}^{(k)}}$ - коэффициент корреляции между декларированными и расчетными значениями выходной величины.

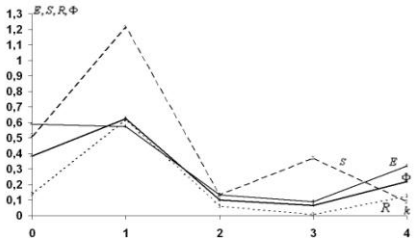


Рисунок 5. Зависимость частных критериев точности *E*, устойчивости *S* , детерминированности *R* и финишного критерия Ф от номера итерации *k*

Указанное утверждение доказано численно на примере модели зависимости выручки от восьми входных факторов. Строилась НС типа MLP с двумя скрытыми слоями, активационной функцией сигмоид $f(s) = (1 + \exp(-as))^{-1}, a > 0$ в них, в выходном – линейной. Сводная характеристика каждой итерации представлена на рисунке 5. Номер оптимальной итерации очистки кластера – 3, значение критерия $\Phi = 0.0661$.

По результатам вычисления отклонений (2) с помощью рабочей НСМ (9) строится ВМР объектов налогового контроля, оператор которой

$$F_4 : \hat{Y} \to \Theta. \qquad (16)$$

Ранжирование объектов налогового контроля заключается в присвоении каждому налогоплательщику ранга в соответствии с ψ-критерием (3), показывающего степень нарушения им налогового законодательства. Требуется найти множество номеров нало-

гоплательщиков $\Theta = \left\{ g : \Psi = \sum_{\nu=1}^{G^*} \psi_{\nu g} \to \max_g \right\}$, планируемых для проведения выездной проверки, чтобы функционал ранжирования $\Psi = \sum_{\nu=1}^{G^*} \psi_{\nu g}$ был максимальным. Это позволяет получить план выездных проверок в аспекте ожидаемых доначислений.

Рассмотрена сущность методов обобщенного перекрестного подтверждения (ОПП) оценки адекватности НСМ и МОПП оценки адекватности ГНСМ. В качестве эталона сравнения параллельных НСМ введем среднее значение отклонений: $\bar{\delta}_i = \sum_{d=1}^{D} \delta_i^d \Big/ D$, $i = \overline{1, N}$. Для НСМ типа *d* проверим выполнение неравенства допустимого уклонения от «эталона»

$$\left| (\bar{\delta}_i - \delta_i^d) / \bar{\delta}_i \right| \le \eta^* ? \qquad (17)$$

Здесь η^* – экспертно задаваемый уровень ошибки. Для каждой НСМ типа *d* рассчитывается величина $P^d = N^d / N$, где N^d –число наблюдений, удовлетворяющих условию (17). Если $P^d < P^*$, то НСМ типа *d* удовлетворяет процедуре ОПП. Здесь P^* – экспертно задаваемый уровень доверительной вероятности.

Суть МОПП заключается в сравнении планов отбора: $\Theta^d = \left\{ g : \Psi = \sum_{\nu=1}^{G^*} \psi_{g_\nu}^d \to \max_g, d = \overline{1, D^*} \right\}$ по множеству D^* параллельных моделей, уже прошедших ОПП. Если для независимых ГНСМ типов d_α и d_β G^{**} номеров налогоплательщиков из G^* возможных, отобранных в оптимальные планы Θ^{d_α} и Θ^{d_β}, $\alpha \ne \beta$, попадают в отрезок $\nu \in [1; G^*]$ независимо от порядка их следования, то считается, что процедура МОПП подтверждена с доверительной вероятностью $P^{MGCV} = G^{**} / G^*, G^{**} \le G^*$. Для D^{**} ГНСМ, прошедших МОПП, расчетное значение доверительной вероятности P^{MGCV} сравнивается с заданной доверительной вероятностью p^{**}, если $P^{MGCV} \ge p^{**}$, процедура МОПП считается выполненной.

В **третьей главе** описывается рабочий алгоритм ранжирования экономических объектов с сильнозашумленными данными на основе ГНСМ. Приводится общее описание алгоритма и составляющих его вспомогательных процедур – оптимальной кластеризации, очистки кластера от аномальных наблюдений, построения рабочей НСМ и расчета доверительного интервала, ОПП, расчета ψ-критерия (3), ранжирования объектов налогового контроля на основе ψ-критерия, МОПП, окончательного ранжирования. Логическая схема алгоритма представлена на рисунке 6.

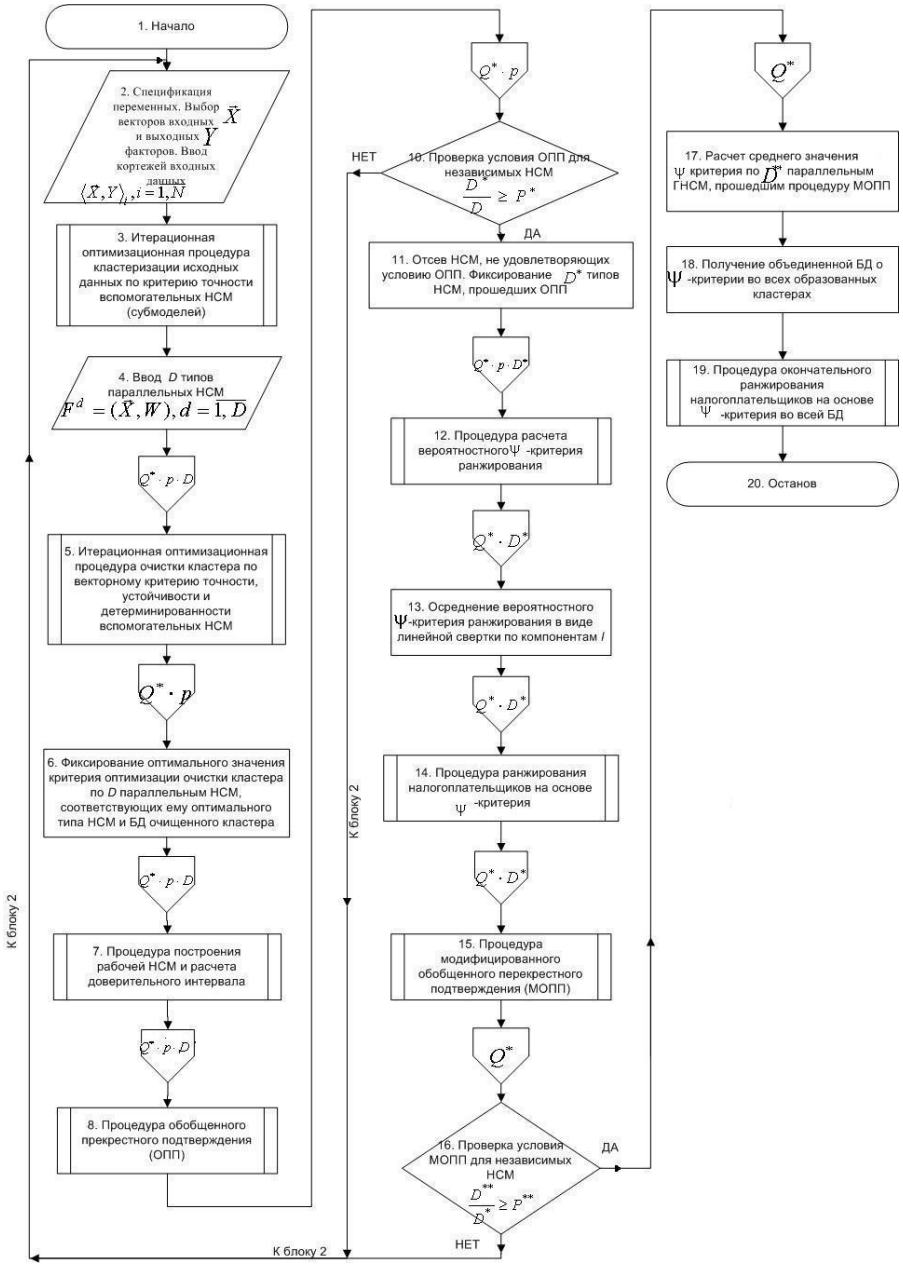
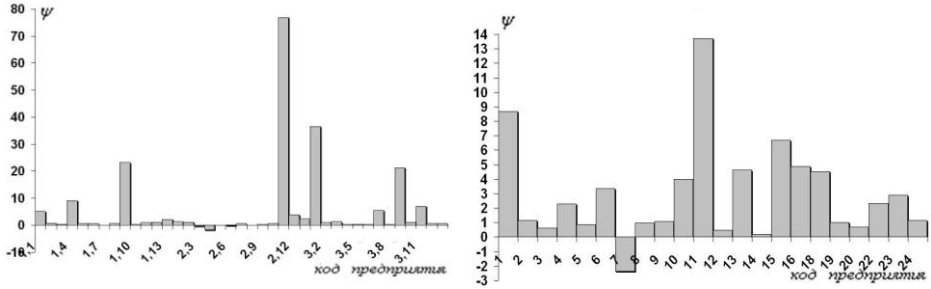


Рисунок 6. Логическая схема рабочего алгоритма ранжирования экономических объектов с сильнозашумленными данными применительно к ранжированию объектов налогового контроля

Четвертая глава посвящена решению прикладных задач ранжирования объектов налогового контроля на основе разработанной ГНСМ. Построены ГНСМ ранжирования налогоплательщиков на выборках, обозначенных как Z^{I} и Z^{II}. Выборка Z^{I} содержит 351 наблюдение. Входные факторы: X_1 - сумма основных средств, тыс. руб; X_2 - себестоимость товаров, продукции, услуг предприятия, тыс. руб; X_3 - среднесписочная численность работающих, чел.; X_4 - сумма оборотных активов, тыс. руб; X_5 - среднегодовая стоимость облагаемого налогом имущества предприятия, тыс. руб; X_6 - коммерческие расходы, тыс. руб. Выходная величина Y - выручка предприятия, тыс. руб. Выборка Z^{II} - 201 наблюдение. Входные факторы: X_1 - сумма основных средств, тыс. руб; X_2 -износ (амортизационные отчисления) за квартал, тыс. руб; X_3-оборотные активы, тыс. руб; X_4- запасы тыс. руб;, X_5-среднесписочная численность работающих, чел.; X_6-дебиторская задолженность, тыс. руб; X_7 - коммерческие расходы за квартал, тыс. руб; X_8- себестоимость реализации товаров за квартал, тыс. руб. Выходная величина Y - выручка предприятия, тыс. руб. Исходные данные взяты из монографии Г.И.Букаева, Н.Д.Бублика, С.А.Горбаткова, Р.Ф. Саттарова «*Модернизация системы налогового контроля на основе нейросетевых информационных технологий*», М.: Наука, 2001. Результаты ранжирования на этих выборках представлены на рисунке 7, где на оси абсцисс обозначены коды предприятий-налогоплательщиков в выборке, на оси ординат – значение ψ-критерия (3).



А Б

Рисунок 7. Результаты ранжирования налогоплательщиков на выборках Z^{I} (А) и Z^{II} (Б).

Проведена верификация ГНСМ на шести натурных объектах, результаты которой отражены в таблице 1. Здесь использованы реальные исходные данные, состав факторов и результаты моделирования НСМ выручки из работы [14].

Таблица 1. Результаты верификации ГНСМ по поверочному эксперименту

Код предприятия	Y_o	$\hat{Y}$	$Y_{вп}$	Сумма доначислений по решению ИФНС	$\delta_{вп}\cdot 100\%$	$\delta\cdot 100\%$
1	858,7	1940,6	4293,5	3434,8	400%	126%
2	113,5	107,8	113,5	0	0,00%	5%
3	328613	656613	527869	199256	60,64%	99,81%
4	7328	7865,3	7328	0	0,00%	7,3%
5	5591	5216,158	5591	0	0,00%	-6,7%
6	907	825,9	909	2	0,22%	-8,9%

В таблице 1: $Y_д$ – значение выходной величины, декларированное плательщиком; $Y_{вп}$ – значение выходной величины, уточненное в ходе выездной проверки с учетом доначислении; $\hat{Y}$ – осредненное по шести ГНСМ расчетное значение выходной величины; $\delta_{вп} = (Y_{вп} - Y_д)/Y_д$ – относительное отклонение между декларированным и скорректированным в ходе выездной проверки значением выходной величины; δ – отклонение, определяемое по (2). Предприятия с кодом 1 и 3 были классифицированы как «нарушитель». Летом 2002 г. на этих предприятиях были организованы выездные проверки, подтвердившие данные моделирования. Таким образом, модель достоверно распознает как нарушителей, искажающих документацию, так и законопослушных налогоплательщиков.

Независимым подтверждением адекватности ГНСМ служат результаты сравнения полученного на ее основе плана отбора налогоплательщиков для 18 проверяющих бригад с планом отбора, полученным по альтернативной модели непараметрического сглаживания Estimation Tax (Голичев И.И. Вариков А.А. Свидетельство № 2006616133 об официциальной регистрации программы для ЭВМ. *Аппроксимация регрессионной зависимости*. М.: РосПатент, 2006). В таблице 2 приведены коды налогоплательщиков, включенных в планы отбора по обеим моделям. Совпадения обозначены заливкой.

Таблица 2. Сравнение ГНСМ с альтернативной моделью отбора

ГНСМ	41	35	26	43	66	68	11	27	40	73	18	25	71	52	46	57	16
Estimation Tax	41	60	21	35	11	26	43	18	40	73	68	27	71	46	25	66	78

Как следует из таблицы 2, по каждой модели совпадают 15 объектов налогового контроля из 18, т.е. модели взаимно подтверждают друг друга на 83 %.

В **заключении** подводятся основные итоги выполнения диссертации**.**

В **приложениях** приводятся исходные данные, использованные для построения ГНСМ.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

1. Разработан метод синтеза плана отбора налогоплательщиков для проведения выездных проверок, основанный на использовании полученного с помощью ГНСМ «эталона» – оценки производственной функции кластера налогоплательщиков.

2. Предложен оригинальный метод предпроцессорной обработки данных, реализованный в разработке специальных процедур, без использования которых построение адекватных моделей с приемлемыми аппроксимативными свойствами не представляется возможным. Предложены процедуры оптимальной кластеризации и оптимальной очистки кластера от аномальных наблюдений на первом и втором иерархических уровнях структурирования, повышающие однородность БД.

3. Разработан вероятностный критерий ранжирования объектов налогового контроля по числовой мере искажения ими отчетной документации с внесением в него эвристической априорной информации, полученной на основе использования доверительных интервалов для отклонений между расчетными (полученными с помощью НСМ) и декларированными значениями моделируемого показателя.

4. Разработан метод МОПП ГНСМ по финишному критерию совпадения множества проранжированных налогоплательщиков для нескольких независимых моделей с заданной доверительной вероятностью. Данный метод служит основным инструментом анализа и подтверждения адекватности ГНСМ.

5. Разработан алгоритм ранжирования экономических объектов с зашумленными данными на базе ГНСМ.

6. Теоретические предложения по разработке технологии ранжирования объектов налогового контроля апробированы путем построения ГНСМ на реальных числовых данных. Осуществлена проверка адекватности НСМ с помощью процедуры ОПП и ГНСМ с помощью процедуры МОПП. Положительные результаты верификации модели на натурных объектах (таблица 1), показавшие правильное распознавание, как «нарушителей», так и законопослушных налогоплательщиков, являются доказательством пригодности разработанной ГНСМ для ее использования при составлении производственного плана проведения выездных проверок. Сравнение ГНСМ с независимой моделью непараметрического сглаживания (таблица 2), давшее совпадение 83%, является подтверждением их взаимной адекватности.

СПИСОК ПУБЛИКАЦИЙ

В рецензируемых журналах из списка ВАК

1. Повышение устойчивости нейросетевых моделей налогового контроля с использованием общесистемных закономерностей. / С.А. Горбатков, **Д.В. Полупанов**, А.М. Солнцев и др. // Обозрение прикладной и промышленной математики. 2004. Т.11. Вып. 4. С. 786–787.

2. Совершенствование нейросетевой математической модели налогового контроля на основе общесистемных закономерностей кибернетики. / С.А. Горбатков, **Д.В. Полупанов** // Нейрокомпьютеры: разработка, применение. 2005. № 3. С. 43–52.

3. Построение нейросетевых математических моделей в технических и экономических системах в условиях искажения входных данных / С.А. Горбатков, **Д.В. Полупанов**, Г.А. Бесхлебнова и др. // Обозрение прикладной и промышленной математики. 2005. Т.12. Вып. 2. С. 337–338.

4. Устойчивость нейросетевого отображения на обучающем множестве в смысле ошибки обобщения. / С.А. Горбатков, **Д.В. Полупанов** // Нейрокомпьютеры: разработка, применение. 2005. № 12. С. 25–34.

5. Совершенствование нейросетевой математической модели налогового контроля на основе оптимизационной процедуры очистки кластера по векторному критерию точности и устойчивости / С.А. Горбатков, **Д.В. Полупанов** // Нейрокомпьютеры: разработка, применение. 2006. № 3. С. 69–74.

В других изданиях

6. Совершенствование региональной системы налогового контроля и управления на основе нейросетевых информационных технологий. / Н.Д. Бублик, С.А. Горбатков, **Д.В. Полупанов** и др. Уфа: Башкирский территориальный институт профессиональных бухгалтеров, 2000. 64 с.

7. Математическое моделирование финансовых показателей сложных экономических объектов на основе нейросетевых технологий / **Д.В. Полупанов** // Радиоэлектроника, электротехника и энергетика. Шестая международная научно-техническая конференция студентов и аспирантов. М.: Издательство МЭИ, 2000. С. 318–319.

8. Математическая нейросетевая модель оценки финансовых показателей объектов налогообложения и разработка плана документальных налоговых проверок на ее основе / **Д.В. Полупанов** // Республиканская конференция студентов и аспирантов по математике. Уфа, БашГУ, 2000. С. 196–197.

9. Теорема существования элемента наилучшего приближения в задаче обучения нейронных сетей / **Д.В. Полупанов** // Аспирант и соискатель. 2001. № 5(6). С. 177–179.

10. Инструментарий нейросетевого моделирования. / С.А. Горбатков, **Д.В. Полупанов** // Г.И.Букаев, Н.Д.Бублик, С.А. Горбатков, Р.Ф. Саттаров Модернизация региональной системы налогового контроля и управления на основе нейросетевых информационных технологий. М.: Наука, 2001. С. 187–221

11. Постановка задачи моделирования выручки. Выбор входных факторов и выходного показателя. Образование кластеров. / С.А. Горбатков, **Д.В. Полупанов,** Б.Г. Колбин // Г.И.Букаев,

Н.Д.Бублик, С.А. Горбатков, Р.Ф. Саттаров Модернизация региональной системы налогового контроля и управления на основе нейросетевых информационных технологий. М.: Наука, 2001. С. 222– 225.

12. Теорема устойчивости нейросетевого отображения по возмущению начальных данных на тестовом множестве / **Д.В. Полупанов** // Нейрокомпьютеры и их применение НКП – 2002: труды VIII Всероссийской конференции с международным участием / Под ред. проф. А.И. Галушкина. М.: Ин-т проблем управления им. В.А. Трапезникова РАН, 2002. С. 1019–1022.

13. Апробация концепции вложенных математических моделей. / С.А. Горбатков, **Д.В. Полупанов**, Р.Р. Сиразев // Н.Д.Бублик, И.И. Голичев, С.А. Горбатков, А.В. Смирнов. Теоретические основы разработки технологии налогового контроля и управления. Уфа: РИО БашГУ, 2004. С. 190–193.

14. Верификация нейросетевой модели на основе натурных экспериментов. / С.А. Горбатков, Н.Т.Габдрахманова, **Д.В. Полупанов** // Н.Д.Бублик, И.И. Голичев, С.А. Горбатков, А.В. Смирнов. Теоретические основы разработки технологии налогового контроля и управления. Уфа: РИО БашГУ, 2004. С. 209–212.

15. Реализация принципа комбинации различных методов для разработки модели оптимизации плана выездных проверок в СНКУ / С.А. Горбатков, **Д.В. Полупанов** // Н.Д.Бублик, И.И. Голичев, С.А. Горбатков, А.В. Смирнов. Теоретические основы разработки технологии налогового контроля и управления. Уфа: РИО БашГУ, 2004. С. 213–219.

16. Построение оптимального плана отбора для выездных налоговых проверок предприятий сферы гостиничного бизнеса с помощью вероятностного критерия. / **Д.В. Полупанов** // Н.Д.Бублик, И.И. Голичев, С.А. Горбатков, А.В Смирнов. Теоретические основы разработки технологии налогового контроля и управления. Уфа: РИО БашГУ, 2004. С. 315–318.

17. К вопросу обеспечения адекватности гибридной нейросетевой модели налогового контроля / **Д.В. Полупанов** // Информационные технологии моделирования и управления. 2005. №6. С. 812–820.

18. Об одном методе предобработки сильнозашумленных данных при построении нейросетевой модели налогового контроля / **Д.В. Полупанов** // Информационные технологии моделирования и управления. 2005. №6. С. 821–827.

19. Алгоритм синтеза оптимального плана отбора налогоплательщиков для проведения выездных проверок на основе гибридной нейросетевой математической модели / С.А. Горбатков, **Д.В. Полупанов,** А.М. Солнцев // Сборник научных трудов по материалам научно-практической конференции «Современные направления теоретических и прикладных исследований». Т.5. Экономика. Одесса: Черноморье, 2006. С. 21– 26.

20. Компьютерная технология тематических выездных налоговых проверок на основе нейросетевого моделирования / С.А. Горбатков, **Д.В. Полупанов,** А.М. Солнцев // Сборник научных трудов по материалам научно-практической конференции «Современные направления теоретических и прикладных исследований». Т.5.Экономика. Одесса: Черноморье, 2006. С. 26– 30.

21. Процедура оптимальной кластеризации исходных данных при построении нейросетевой модели налогового контроля / **Д.В. Полупанов** // Проблемы и перспективы российской экономики: сборник статей V Всероссийской научно-практической конференции. Пенза, НОУ «Приволжский Дом знаний», 2006. С. 141–144.

22. Рабочий алгоритм ранжирования экономических объектов с сильнозашумленными данными на основе гибридной нейросетевой математической модели. / С.А. Горбатков **Д.В. Полупанов** // Свидетельство об отраслевой регистрации разработки в отраслевом фонде алгоритмов и программ № 6398 от 16.06.2006. Номер государственной регистрации в Национальном информационном фонде неопубликованных документов: 50200600974 от 19.06.2006.

Диссертант _____ Полупанов Д.В.