

**КОНСТАНТИНОВА**

**Екатерина Даниловна**

**Статистическое моделирование многофакторного воздействия на живые системы с бинарным откликом при наличии корреляций между факторами**

Специальность 05.13.18 – Математическое моделирование, численные методы и комплексы программ

**АВТОРЕФЕРАТ**

диссертации на соискание ученой степени  
кандидата физико-математических наук

Екатеринбург – 2012

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институте промышленной экологии Уральского отделения Российской академии наук

Научный руководитель: доктор физико-математических наук, профессор  
Вараксин Анатолий Николаевич

Официальные оппоненты: доктор технических наук, профессор  
Гольдштейн Сергей Львович,  
доктор физико-математических наук  
Кацнельсон Леонид Борисович

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт математики и механики Уральского отделения РАН, г. Екатеринбург

Защита состоится 21 февраля 2012 г. в 15:00 часов на заседании диссертационного совета Д 212.285.13 при ФГАОУ ВПО «Уральский федеральный университет имени первого Президента России Б.Н. Ельцина» по адресу: 620002, г. Екатеринбург, ул. Мира, 19, аудитория I главного учебного корпуса (зал ученого совета).

С диссертацией можно ознакомиться в читальном зале библиотеки ФГАОУ ВПО «Уральский федеральный университет имени первого Президента России Б.Н. Ельцина».

Отзыв на автореферат в одном экземпляре, заверенный гербовой печатью, прошу направить по адресу: 620002, г. Екатеринбург, ул. Мира, 19, ФГАОУ ВПО «Уральский федеральный университет имени первого Президента России Б.Н. Ельцина», ученому секретарю университета.

Автореферат разослан "19" января 2012 года.

Ученый секретарь диссертационного совета Д 212.285.13,  
кандидат физ.-мат. наук, профессор



Рогович В.И.

## **Общая характеристика работы**

**Актуальность темы** Одна из проблем современного анализа данных – поиск ведущих факторов, определяющих поведение системы. Актуальной и практически значимой является задача определения ведущих факторов и их *комплексов*, оказывающих максимальное влияние на живые системы, например, влияние комплекса факторов среды обитания на природные популяции животных и растений, факторов риска на здоровье населения.

Одним из инструментов для количественной оценки взаимосвязей в сложной системе (например, в системе «среда обитания – здоровье населения») являются методы многофакторного статистического анализа, которые позволяют учитывать одновременное влияние на систему большого числа факторов. Такие методы позволяют разрабатывать новые методики и алгоритмы построения новых многофакторных моделей системы и, на их основе, *интерпретировать поведение системы* (актуальная задача современной науки). Важным условием эффективной интерпретации поведения системы является применение *предметно-ориентированного* подхода, все этапы которого поддаются осмыслению специалистом в предметной области и дают результаты, важные для академической и практической науки (например, для экологии человека и биологии). Только в этом случае результатом моделирования являются *новые знания о системе*, а не набор чисел, не имеющих реального смысла. Разработка такого предметно-ориентированного подхода является актуальной задачей, решение которой позволит повысить эффективность управления сложными системами (например, управления здоровьем населения в связи с воздействием комплекса факторов окружающей среды).

**Цель работы** Разработка методических подходов к статистическому моделированию многофакторного воздействия на живую систему при наличии корреляций между факторами; применение методологии моделирования для описания воздействия комплекса факторов риска на здоровье населения.

### **Основные задачи работы**

1. Разработка методов корректировки однофакторных эффектов, искаженных взаимосвязями между факторами.
2. Разработка методики анализа двухфакторных эффектов с оценкой степени неаддитивности.
3. Разработка методики выявления комплекса факторов (число факторов больше двух), оказывающих наибольшее влияние на систему.

4. Применение разработанных методик для изучения взаимосвязей между факторами риска и здоровьем детей Екатеринбурга.

#### **Научная новизна исследования**

1. Для систем с бинарным откликом и категоризованными факторами разработана новая комплексная методология моделирования эффектов разной размерности (от однофакторных до 3-4 факторных) с явным учетом взаимосвязей между факторами, определяющими поведение системы.

2. На основе идеологии иерархической классификации (дерева классификации) предложена новая методика построения и анализа многофакторных моделей взаимосвязей факторов риска и здоровья населения.

3. С использованием разработанной методологии впервые проведено комплексное исследование взаимосвязей показателей здоровья детей-дошкольников Екатеринбурга с набором экологических и социальных факторов риска потери здоровья; впервые определены сочетания 3-4 факторов риска, оказывающих наиболее неблагоприятное влияние на состояние детей.

4. Разработаны и протестированы вычислительные методы анализа взаимосвязей факторов риска и здоровья населения. На их основе разработан комплекс программ, включающий:

- программу автоматического анализа двухфакторных эффектов для изучения эффектов неаддитивности;
- программы создания выборок методом «случай-контроль» с различными вариантами создания копий;
- программу пошагового полуавтоматического построения леса деревьев классификации.

**Практическая значимость работы** Методика построения и анализа многофакторных статистических моделей, описывающих взаимосвязи произвольного бинарного отклика с комплексом категоризованных факторов, используется в учебном процессе в Уральском федеральном университете при чтении курсов лекций «Моделирование» и «Методы обработки биомедицинских данных» (имеется акт внедрения).

Факторы риска и их сочетания, оказывающие максимальное влияние на распространенность заболеваний у детей Екатеринбурга, найденные в результате комплексного анализа, используются для разработки научно-обоснованных программ по сохранению и восстановлению здоровья детей (управление здоровьем).

Результаты переданы специалистам Екатеринбургского Центра детской экопатологии и используется в практике работ Центра (имеется акт внедрения).

Работа выполнена при поддержке РФФИ (грант № 07-04-96120) и Президиума РАН (проект ФМ-Н № 09-П-2-1027). В настоящее время результаты работы используются при выполнении междисциплинарных исследований УрО РАН (проект 12-М-24-2016).

### **Положения, выносимые на защиту**

1. Предложенный вариант пошагового построения деревьев классификации позволяет получить эффективные и наглядные решающие правила для разделения объектов на несколько классов.
2. Искажения эффектов «низкой размерности», обусловленные коррелированностью факторов, требуют корректировки. Корректировка может быть выполнена предложенными в диссертации различными методами, среди которых наилучшими свойствами обладает «метод маргинальных частот». Широко используемый в экологии человека и биологии метод логистической регрессии в ряде случаев дает неудовлетворительные результаты.
3. Реально наблюдаемые в г. Екатеринбурге уровни загрязнения среды обитания человека *в сочетании* с социально-экономическими факторами риска предметно и статистически значимо повышают распространенность заболеваний органов дыхания, системы кровообращения, болезней костно-мышечной системы и соединительной ткани, расстройств поведения у детей дошкольного возраста.
4. При совместном действии комплекса факторов риска на детей Екатеринбурга имеют место сильные сверхаддитивные эффекты.

**Личный вклад автора** Вошедшие в диссертацию результаты получены автором совместно с научным руководителем, профессором А.Н. Вараксиным. Диссертант провел системный анализ взаимосвязей показателей здоровья детского населения с факторами риска на основе идеологии иерархической классификации, выявил комплексы факторов риска, наименее благоприятные для здоровья детей, разработал методы коррекции эффектов, искаженных коррелированностью факторов.

**Реализация и апробация работы** Основные положения диссертационной работы были представлены на Всероссийской научной конференции «Влияние загрязнения окружающей среды на здоровье человека», Новосибирск, 2002; X Международном экологическом симпозиуме «Урал атомный, Урал промышленный», Екатеринбург, 2002; научно-практической конференции «Здоровье детей и экология»,

Екатеринбург, 2003; Всероссийской научно-практической конференции «Современные технологии исследований в гигиене и экологии», Санкт-Петербург, 2004; XI Всероссийском конгрессе «Экология и здоровье человека», Самара, 2006; Пленуме научного совета по экологии человека и гигиене окружающей среды, Москва, 2006; 3-й Международной научно-практической конференции «Составляющие научно-технического прогресса», Тамбов, 2007; 2-ом Международном экологическом Форуме, Санкт-Петербург, 2008; 5-ой международной конференции «Экологические и гидрометеорологические проблемы больших городов и промышленных зон». Санкт-Петербург, 2009; 23<sup>rd</sup> annual Conference of International Society for Environmental Epidemiology. Barcelona (Spain), 13-16 September 2011.

**Публикации** Основное содержание диссертации представлено в 20 публикациях, из них 7 в журналах из списка ВАК.

**Объем и структура работы** Диссертация состоит из введения, четырех глав, выводов, списка литературы, содержит 125 страниц основного текста, 33 таблицы, 36 рисунков и одно приложение. Список литературы включает 156 источников и содержит 16 страниц.

### **Основное содержание работы**

**Во введении** обоснована актуальность проблемы, сформулированы цель и основные задачи исследования: построение и анализ статистических моделей многофакторного воздействия на живую систему в случае категоризованных факторов и бинарного отклика системы на воздействие; такие ограничения (бинарный отклик и категоризованные факторы) существенно сужают круг возможных методов статистического моделирования. Наличие корреляций между факторами резко усложняет получение и интерпретацию результатов, что приводит к необходимости разработки новых методов и подходов.

**Первая глава** диссертации представляет обзор исследований по статистическому моделированию взаимосвязей между состоянием живой системы и набором определяющих поведение системы факторов, рассмотрены существующие на сегодняшний день в этой области методы и направления (для произвольных систем в случае бинарного отклика и категоризованных факторов). В качестве основного примера приложения многофакторного моделирования для живой системы рассматривается человеческая популяция, для которой актуальной проблемой является построение моделей взаимосвязей между здоровьем населения

и факторами риска потери здоровья, их анализ и предметная интерпретация результатов.

Для эффектов низкой размерности (одно- и двухфакторные эффекты) основной проблемой является их искажение вследствие коррелированности факторов. В диссертации показано, что наиболее распространенный метод коррекции однофакторных эффектов для систем с бинарным откликом – метод логистической регрессии – не всегда дает удовлетворительные результаты; отсюда вытекает необходимость разработки новых методов.

Основным инструментом для количественной оценки и анализа многофакторных взаимосвязей в системе с бинарным откликом и набором категоризованных факторов среди методов прикладной математической статистики являются методы классификации (дискриминантного анализа), такие как логистическая регрессия, распознавание образов, нейронные сети и т.п. Отдельные этапы моделирования систем с бинарным откликом и категоризованными факторами могут быть выполнены методами, аналогичными дисперсионному анализу. Главными недостатками перечисленных методов являются: 1) необходимость построения решающего (классификационного) правила высокого качества; для живых систем, в частности, для человеческих популяций, при решении многих важных практических задач это в принципе неосуществимо. 2) многие методы классификации (распознавание образов, нейронные сети и др.) являются моделями «черного ящика», которые не допускают предметной интерпретации. 3) у моделей классификации, таких как метод Фишера, логистическая регрессия и аналогов моделей дисперсионного анализа, при наличии взаимосвязей между факторами теряется предметный смысл коэффициентов модели; в нашем подходе предметная трактовка результатов является одной из основных целей моделирования.

В диссертации показано, что плодотворная идея, которая позволяет решить поставленную задачу, содержится в методе деревьев классификации (идея последовательного иерархического построения решающего правила). Методы иерархической классификации позволяют, после определенной модификации, построить *предметно-ориентированные* модели, понятные не только математику, но и специалисту в предметной области (биологу, экологу, эпидемиологу). Решающее правило предметно-ориентированной модели позволяет специалисту объяснить полученные результаты моделирования, получить новые знания о системе, сформулировать практические рекомендации по целенаправленному

воздействию на систему (например, рекомендации, снижающие заболеваемость населения, подвергающего воздействию комплекса факторов риска).

**Во второй главе** решается проблема определения однофакторных эффектов, искаженных наличием взаимосвязей между факторами. Пусть имеется набор категоризованных факторов  $\Phi P_1, \Phi P_2, \dots$ , оказывающих влияние на отклик системы  $W$ . Каждый объект исследования может находиться в двух состояниях: 0 и 1 (бинарный отклик). В статистических исследованиях реально наблюдаемый отклик  $W$  – это доля объектов в состоянии 1 (бинарный отклик, усредненный по некоторому множеству объектов). Например, в задачах экологии человека, при изучении влияния факторов риска на здоровье, человек может находиться в двух состояниях: 0 (здоров) и 1 (болен). В статистических исследованиях анализируют показатель  $W$  – распространенность заболевания (доля больных среди исследованных) на разных уровнях факторов и на сочетаниях уровней комплекса факторов риска потери здоровья. Факторы разделяют изучаемые объекты на группы по уровням факторов, а отклик принимает значения  $W_{ijl\dots}$ , где каждый индекс соответствует определенному фактору.

Под однофакторным эффектом (бинарного) фактора  $\Phi P$  понимаем величину

$$\Delta W_1 (\Phi P) = W_1 (\Phi P=1) - W_0 (\Phi P=0), \quad (1)$$

где  $W_1$  и  $W_0$  – доля объектов исследования в состоянии 1 при  $\Phi P=1$  и  $\Phi P=0$ ; в примере с заболеваемостью  $\Delta W_1 (\Phi P)$  – это увеличение распространенности заболевания при наличии фактора риска ( $\Phi P=1$ ) по сравнению с его отсутствием ( $\Phi P=0$ ).

При наличии взаимосвязанных факторов  $\Phi P_1, \Phi P_2, \dots$  и т.д. однофакторный эффект (1) фактора  $\Phi P_1$  отражает, в той или иной мере, одновременное действие на систему всех перечисленных факторов. Для практических приложений оценка и анализ «истинных» (полученных для независимых факторов) однофакторных эффектов имеет особое значение. При наличии взаимосвязанных факторов получить точное значение «истинного» однофакторного эффекта невозможно в принципе; поэтому, речь идет о разработке численных методов *коррекции* эффектов, искаженных взаимосвязями между факторами.

При анализе ситуаций с коррелированными факторами желательно иметь наглядную и понятную специалисту в предметной области картину всех связей (какие факторы и как сильно связаны друг с другом). Для этих целей в диссертации использован метод корреляционных плеяд: между всеми парами факторов



рассчитывалась мера Крамера  $V$  и строилась плеяда, в которой отражены связи с *максимальными* значениями  $V$ . В диссертации приведены примеры построения двух плеяд: 1-связи между уровнями загрязнения атмосферного воздуха Санкт-Петербурга различными токсикантами (первичные данные А.П. Щербо и сотр., 2002); 2-связи между социально-экономическими показателями семей г.Екатеринбурга, имеющих детей дошкольного возраста (Константинова, Вараксин, 2002-2007). В диссертации проведено предметное обсуждение обеих плеяд. Вторая плеяда использована в главе 5 для проведения комплексного анализа различных факторов риска на здоровье детей Екатеринбурга.

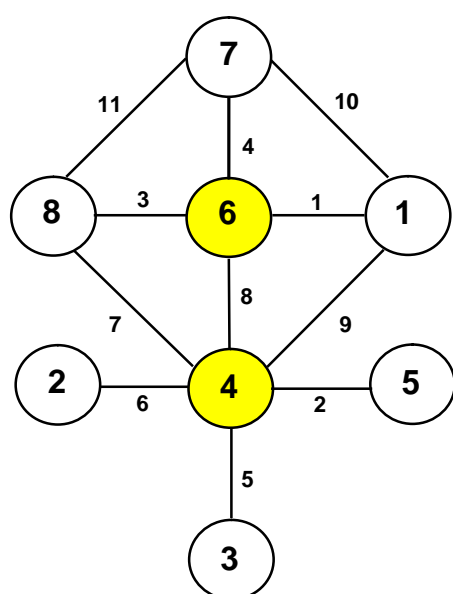


Рис. 1. Корреляционная плеяда

Факторы риска на рис.:

- 1 – тип питьевой воды
- 2 – санитарное состояние квартиры
- 3 – семья полная / неполная
- 4 – уровень материальной обеспеченности семьи
- 5 – психологический климат семьи
- 6 – уровень образования матери
- 7 – курение матери (есть, нет)
- 8 – физическая активность ребенка

На рисунке кружками обозначены факторы (цифра в кружке – номер фактора), а линиями – связи между факторами; номера у линий показывают ранг (силу) связи. Самая сильная связь (ранг 1) наблюдается между ФР «Образование матери» и «Тип питьевой воды»; как и следовало ожидать, в семьях с более высоким образованием матери чаще употребляют чистую или фильтрованную воду. Мы обнаружили два «системообразующих» (имеющих наибольшее количество связей) фактора: уровень образования матери (фактор № 6) и уровень материальной обеспеченности семьи (№ 4).

В диссертации подробно описаны предложенные нами новые численные методы коррекции однофакторных эффектов: метод линейной коррекции (безусловная коррекция), метод маргинальных частот (условная коррекция), метод

«Случай-контроль» с подбором копий; для сравнения приведены результаты коррекции, полученные известным методом логистической регрессии.

В диссертации приведены результаты расчетов однофакторных эффектов и их коррекция на *конкретных* примерах из области биологии (физиологические показатели мелких грызунов Уральского региона) и экологии человека (заболеваемость детей Екатеринбурга). Показано, что взаимосвязи между факторами значимо искажают однофакторные эффекты основных исследуемых факторов, так что их корректировка является необходимым этапом работы. Результаты примера из области экологии человека приведены в табл. 1.

Таблица 1. Результаты сравнительного анализа методов коррекции однофакторного эффекта на примере одновременного действия на распространенность болезней системы кровообращения основного фактора «Физическая активность ребенка» и сопутствующего фактора «Уровень образования матери».

№№	Метод оценки эффекта	Эффект $\Delta W1$ , %	Отношение шансов OR
1	Однофакторный эффект фактора «Физическая активность» («истинный»)	?	?
2	Однофакторный эффект фактора «Физическая активность», искаженный связью с фактором «Уровень образования матери»	5,85	1,63
3	Логистическая регрессия, эффект фактора «Физическая активность», искаженный связью с фактором «Уровень образования матери»	---	1,63
4	Однофакторный эффект фактора «Физическая активность», скорректированный методом линейной коррекции	3,94 ↓	---
5	Однофакторный эффект фактора «Физическая активность», скорректированный методом маргинальных частот	4,61 ↓	1,46 ↓
6	Логистическая регрессия, эффект фактора «Физическая активность», скорректированный на влияние фактора «Уровень образования матери»	---	1,45 ↓
7	Эффект, скорректированный методом «Случай-контроль» (подбор копий по сопутствующему фактору «Уровень образования матери»)	6,00 ↑	1,64 ↑

Величина однофакторного эффекта, скорректированного методом линейной коррекции и методом маргинальных частот ниже, чем нескорректированный эффект; уменьшение эффекта (снижение показателя «Отношение шансов») дает коррективку методом логистической регрессии. Уменьшение эффекта после коррективки ожидаемо, поскольку предварительные экспертные оценки говорят в его пользу. В то же время метод «Случай-контроль» даёт завышенное, по сравнению с исходным, значение скорректированного однофакторного эффекта фактора «Физическая активность». Возникает закономерный вопрос: какое из полученных разными методами значений скорректированного однофакторного эффекта ближе всего к истинному? Оставаясь в рамках имеющихся экспериментальных данных, ответить на него невозможно. Для получения ответа на этот вопрос была поставлена серия модельных численных экспериментов, в которых «истинный» эффект заранее известен (задан).

Результаты одного из таких экспериментов приведены в табл. 2. По результатам всех модельных экспериментов наиболее эффективным признан метод маргинальных частот, суть которого (на примере двух бинарных факторов) заключается в следующем.

Пусть ФР1 считается основным фактором, влияние которого надо изучить, а ФР2 является фактором, «мешающим» изучению влияния ФР1. Задача заключается в том, чтобы выявить влияние ФР1 на  $W$  в более «чистом» виде, устранив мешающее влияние ФР2. Используем обозначения:  $W_{ij} = W(\Phi P1 = i, \Phi P2 = j)$  - распространенность «признака»,  $n_{ij}$  - численность объектов на уровнях  $(i, j)$ . Если ФР1 - основной фактор, требуется еще ввести обозначение  $n_{+j} = n_{0j} + n_{1j}$  - численность объектов на уровнях фактора ФР2 без деления на уровни ФР1. Для коррективки искажающего эффекта фактора ФР2 необходимо, чтобы соотношение распространенностей второго фактора ФР2 на уровне ФР1=0 (это числа  $n_{00}$  и  $n_{01}$ ) и на уровне ФР1=1 (числа  $n_{10}$  и  $n_{11}$ ) было такое же, как для всех объектов вместе на уровнях ФР2 без деления на уровни ФР1 ( $n_{+0}$  и  $n_{+1}$ ). Тогда нескорректированный однофакторный эффект фактора ФР1, введенный ранее соотношением (1), выражается через  $W_{ij}$  как

$$\Delta W1(\Phi P1) = W_1(\Phi P1 = 1) - W_0(\Phi P1 = 0) = \frac{W_{10}n_{10} + W_{11}n_{11}}{n_{10} + n_{11}} - \frac{W_{00}n_{00} + W_{01}n_{01}}{n_{00} + n_{01}},$$

а эффект, скорректированный (*adjusted*) на влияние ФР2, как

$$\Delta W1(\Phi P1)_{adj} = W_1(\Phi P1)_{adj} - W_0(\Phi P1)_{adj} = \frac{W_{10}n_{+0} + W_{11}n_{+1} - W_{00}n_{+0} - W_{01}n_{+1}}{N} \quad (2)$$

Частоты  $n_{+0}$  и  $n_{+1}$  называются маргинальными, в связи с чем описанную методику коррективки предлагаем называть «методом маргинальных частот».

Таблица 2. Результаты сравнительного анализа методов коррекции на модельном примере.

№№	Метод оценки эффекта фактора ФР1	Эффект $\Delta W1$ , доля	Отношение шансов OR
1	Однофакторный эффект фактора ФР1 («истинный»)	0,077	2,408
2		0,143	2,428
3	Логистическая регрессия, эффект фактора ФР1 «истинный»	-----	2,420
4	Однофакторный эффект фактора ФР1, искаженный связью с ФР2	0,064	1,979
5		0,096	1,899
6	Логистическая регрессия, эффект ФР1 искаженный связью с ФР2	-----	1,931
7	Однофакторный эффект ФР1, скорректированный методом линейной коррекции	0,080	-----
8		0,110	-----
9	Однофакторный эффект ФР1, скорректированный с помощью метода маргинальных частот	0,077	2,403
10		0,142	2,405
11	Логистическая регрессия, эффект ФР1, скорректированный на ФР2	-----	2,352
12	Эффект ФР1, скорректированный методом «Случай-контроль» (подбор копий по сопутствующему фактору ФР2)	0,045	2,394
13		0,170	3,836

В примере табл. 2 рассмотрены два трехуровневых фактора (уровни 0,1,2), влияющих на бинарный отклик. В строках 1-2 показаны однофакторные эффекты фактора ФР1, рассчитанные между уровнями 0-1 и 1-2 (величины  $\Delta W1$  рассчитаны по формуле (1); для расчета отношения шансов OR использована четырехпольная таблица); термином «истинный» в данной таблице назван эффект фактора ФР1,

полученный в условиях его независимости от фактора ФР2. В строке 3 приведено отношение шансов, полученное методом логистической регрессии. Логистическая регрессия дает только одно значение отношения шансов (независимо от числа уровней фактора ФР1), которое оказалось близко (как и должно быть) к показанному в строках 1-2 (в строках 1-2 условия модельного эксперимента подбирались таким образом, чтобы два значения отношения шансов OR были близки).

После выполнения этих расчетов были проведены изменения первичных данных таким образом, чтобы факторы ФР1 и ФР2 стали статистически взаимосвязанными. В строках 4-6 показаны эффекты фактора ФР1, искаженные его взаимосвязью с фактором ФР2. Видно, что искажения достаточно заметные с предметной точки зрения и статистически значимы (истинные и искаженные эффекты различаются значимо,  $p < 0,05$ ). С предметной точки зрения отметим, что результатом искажения стало *уменьшение* эффекта (снижение  $\Delta W1$  и OR).

В строках 7-13 табл. 2 показаны результаты коррекции однофакторных эффектов. Здесь и во всех других модельных экспериментах наилучшие результаты показал метод маргинальных частот (эффекты, скорректированные этим методом, оказались очень близки к «истинным»). В некоторых ситуациях хорошие результаты дают методы линейной коррекции и логистической регрессии. Было показано, что указанные методы «работают» в случаях, когда отклик системы на второй фактор ФР2 является монотонным ( $W$  монотонно увеличивается или уменьшается при изменении уровней фактора ФР2). В случае немонотонной зависимости  $W$  от ФР2 методы линейной коррекции и логистической регрессии могут давать неверные (даже по знаку!) величины поправок. Метод «Случай-контроль» с подбором копий по фактору ФР2 в большинстве случаев дает неверные результаты (несмотря на его *a priori* кажущуюся очевидную применимость). Причины неверной работы данного метода – появление эффектов смещения и переуравновешивания, которые неизбежно появляются, когда метод применяется на этапе анализа данных, а не на этапе планирования эксперимента.

**В третьей главе** описаны методы работы с двухфакторными эффектами. Допустим, имеется два бинарных фактора ФР1 и ФР2. По сравнению с однофакторным анализом, где величина эффекта  $\Delta W$  оценивалась по формуле (1), в двухфакторном анализе имеем четыре однофакторных и один двухфакторный  $\Delta W2(\text{ФР1}+\text{ФР2})$  эффект. Однофакторные эффекты определяются как  $\Delta W1(\text{ФР1}, \text{ФР2}=0)=W_{10}-W_{00}$ ,  $\Delta W1(\text{ФР1}, \text{ФР2}=1)=W_{11}-W_{01}$ ,  $\Delta W1(\text{ФР2}, \text{ФР1}=0)=W_{01}-$

$W_{00}$ ,  $\Delta W1(\Phi P2, \Phi P1=1)=W_{11}-W_{10}$ , двухфакторный – как  $\Delta W2(\Phi P1+\Phi P2)=W_{11}-W_{00}$ , где  $W_{11}$  – распространенность изучаемой патологии в группе индивидуумов, на которые действуют оба фактора риска одновременно,  $W_{10}$  и  $W_{01}$  – распространенности в условиях действия одного (первого или второго) фактора риска при отсутствии другого (второго или первого) фактора риска, распространенность  $W_{00}$  определяется для группы индивидуумов при отсутствии обоих факторов риска.

При работе с двухфакторными эффектами возникает ряд вычислительных трудностей: если однофакторный анализ можно провести для каждого фактора даже при наличии очень большого исходного перечня факторов, то при анализе *комплексов* факторов число возможных комбинаций быстро растет с ростом числа факторов в комплексе (время анализа увеличивается). Для сокращения времени анализа на языке Visual Basic в среде Statistica for Windows была создана программа автоматического перебора пар факторов. Эта программа для каждой пары факторов (для всех пар) и выбранного отклика рассчитывает величину двухфакторного эффекта  $\Delta W2(\Phi P1+\Phi P2)=(W_{11}-W_{00})$  и перекрестного члена  $(W_{11}-W_{00})-(W_{10}-W_{00})-(W_{01}-W_{00})$ . Значения двухфакторного и перекрестного эффектов записываются в базу данных, в которой, по окончании полного перебора всех пар факторов, можно провести автоматическую сортировку записей с целью отбора наиболее значимых эффектов. В результате для дальнейшего ручного (экспертного) анализа остается небольшое число наиболее значимых пар факторов.

Для науки и практики наиболее интересны так называемые неаддитивные эффекты действия комплекса факторов на отклик. В случае двух факторов определение аддитивности можно выразить соотношением

$$(W_{11} - W_{00}) = (W_{10} - W_{00}) + (W_{01} - W_{00}). \quad (3)$$

В диссертации показано, что формула (3) является обобщением соотношения классического дисперсионного анализа, которое (в стандартных обозначениях) имеет вид  $(\bar{y}_{ij} - \bar{y}_{i+} - \bar{y}_{+j} + \bar{y}) = 0$ . Формула (3) получается в предположении, что аналогом среднего значения  $\bar{Y}_{ij}$  в дисперсионном анализе является распространенность признака  $W_{ij}$ .

Нарушение равенства условия аддитивности (3) может выражаться в предметно-значимых и понятных специалисту эффектах. В гл. 3 приведены некоторые *примеры* обнаруженных нами двухфакторных эффектов влияния факторов различной природы на здоровье детей Екатеринбурга – выполнения и нарушения условия аддитивности (3) и других предметно-значимых эффектов.

Полный анализ всех двухфакторных эффектов для двух наиболее распространенных заболеваний у детей Екатеринбурга представлен в главе 5.

**В четвертой главе** изложена методология анализа многофакторных (число факторов больше двух) эффектов, позволяющая разрабатывать предметно-ориентированные статистические модели. Основная задача, которая ставится при изучении многофакторных эффектов – нахождение таких комбинаций факторов и их уровней, которые приводят к резкому повышению (или понижению) отклика  $W$  по сравнению с одно- и двухфакторными воздействиями. Какие здесь возникают трудности ?

- большое число многофакторных комбинаций;
- итогом многофакторного анализа должна стать формулировка *простого и понятного* специалисту в предметной области (биологу, эпидемиологу) решающего правила, по которому определяются классы объектов с низкой и высокой распространенностью отклика (например, заболевания).
- построение такого решающего правила в принципе не может быть полностью алгоритмизировано и является результатом экспертной работы специалиста по анализу данных.

Решающее правило служит для специалиста в предметной области основой для разработки управляющих мероприятий. В примере решения задачи взаимосвязи ФР и здоровья населения такое правило используется для разработки мероприятий по снижению заболеваемости детей (управление здоровьем населения).

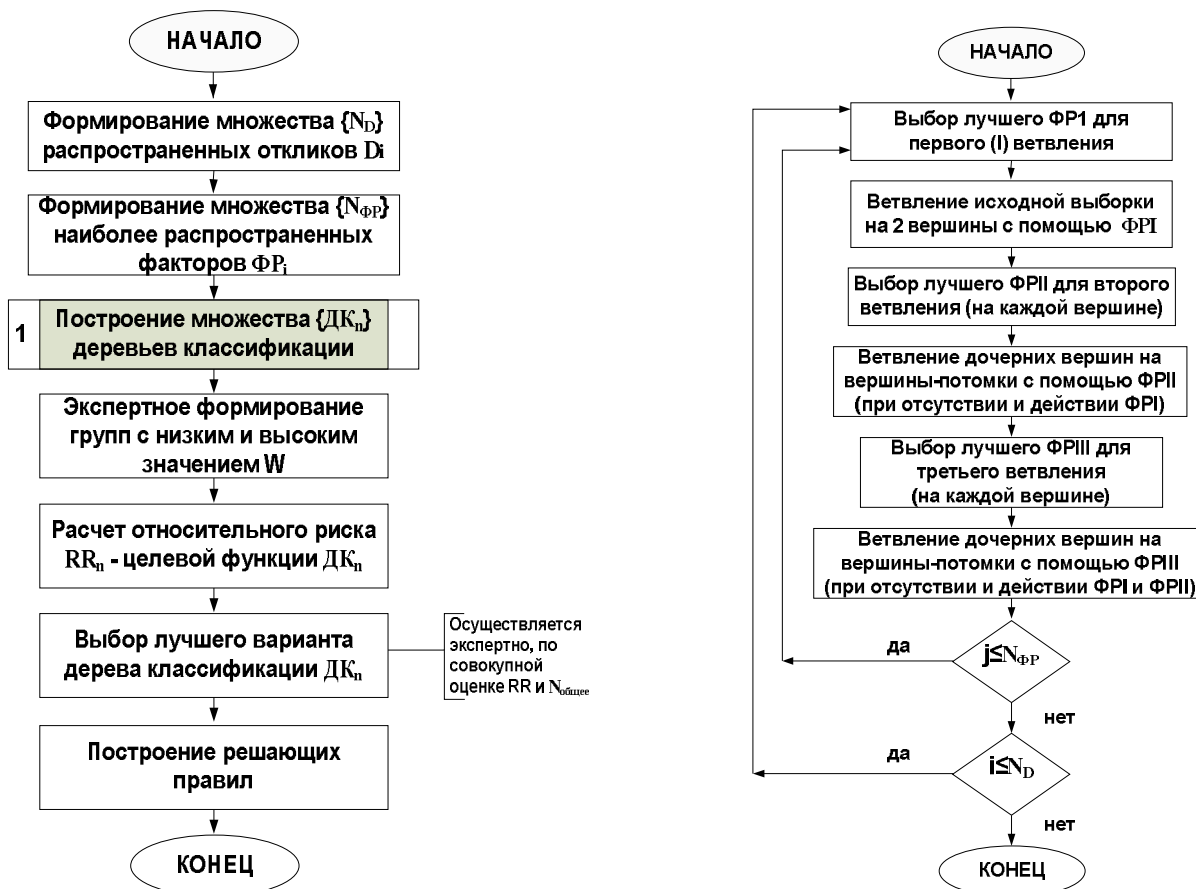
Предлагаемая методика построения решающего правила основана на идее последовательного иерархического построения дерева классификации (ДК) для двух классов объектов (например, больные и здоровые), где предикторами являются, например, факторы риска (ФР) возникновения болезни. Что касается конкретной программной реализации метода, нам пришлось отказаться от версии ДК, имеющейся в известном компьютерном пакете Statistica for Windows. Тому имеется две причины:

- в пакете Statistica разделение вершины дерева на ветви производится по критериям Джини и хи-квадрат; по нашему мнению, лучшим вариантом для решения нашей задачи является разделение по величине однофакторного эффекта  $\Delta W_1$  (1);
- версия компьютерного пакета дает единственный вариант построения дерева, соответствующий экстремуму критерия Джини или хи-квадрат; мы предлагаем исследовать несколько вариантов одного ветвления, используя несколько факторов с максимальными  $\Delta W_1$ .

В результате нами предложен алгоритм построения решающего правила, разделяющего объекты двух классов (например, детей с низкой и высокой распространенностью заболевания), реализуемый в 5 этапов.

1. Построение набора ДК (леса деревьев классификации).
2. Формирование двух классов объектов для каждого дерева.
3. Анализ «качества» деления объектов на классы с высокой и низкой распространенностью отклика.
4. Формулировка предметно-ориентированного решающего правила.
5. Анализ промежуточных и отсеченных терминальных вершин; возможность компенсации одних факторов риска другими.

Алгоритм построения решающего правила в формализованном виде представлен ниже.



Пример построения дерева показан на рис. 1.



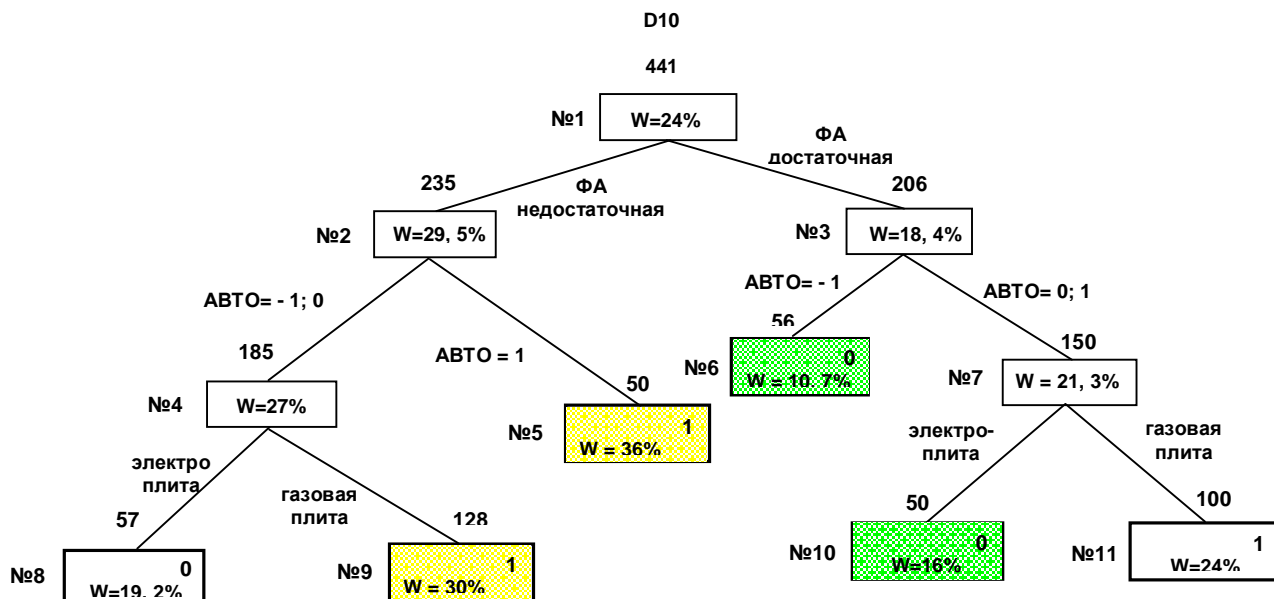


Рис. 1. Дерево классификации для заболеваний верхних дыхательных путей (класс патологий D10)

Очевидно, что решающее правило, сформулированное на основе дерева классификации, понятно специалисту в предметной области (экологическая медицина) и позволяет выработать рекомендации, выполнение которых может снизить заболеваемость детей.

**В пятой главе** описаны результаты комплексного исследования влияния факторов риска различной природы (загрязнение окружающей среды, социально-экономические факторы) на здоровье детей-дошкольников Екатеринбурга. Глава содержит результаты анализа всех наиболее значимых эффектов (одно-, двух и многофакторных) для двух наиболее распространенных классов заболеваний: заболевания органов дыхания и системы кровообращения. Для указанных классов заболеваний приведены результаты анализа (либо даны комментарии) 10 наиболее значимых однофакторных и 20 двухфакторных эффектов.

Анализ двухфакторных эффектов позволил обнаружить сильные эффекты неаддитивности. Рассмотрим пример *значимого эффекта усиления действия факторов риска в паре по сравнению с однофакторными эффектами*. Так, для заболеваний органов дыхания наибольший негативный эффект (максимум  $W_{11}$ ) дает сочетание факторов риска «Газовая плита» и «Неудовлетворительное санитарное состояние квартиры»:  $W_{11}=45,0\%$  , рис. 2.

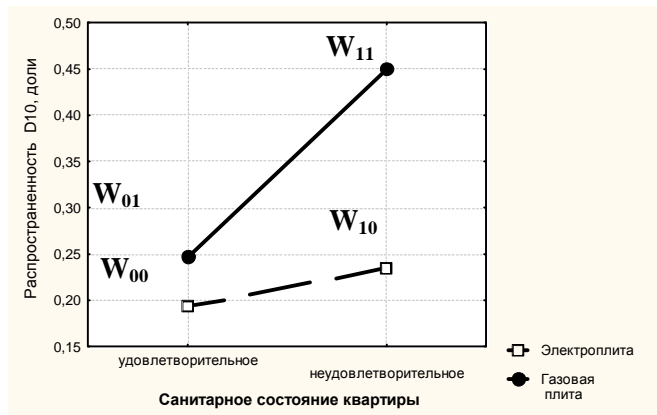


Рис. 2. Двухфакторные эффекты для заболеваний органов дыхания, факторы риска «Неудовлетворительное санитарное состояние квартиры» и «Газовая плита»

Сделаем оценку эффекта неаддитивности. Величина двухфакторного эффекта, стоящая в левой части формулы условия аддитивности (3), равна  $\Delta W_2(\Phi P1 + \Phi P2) = (W_{11} - W_{00}) = 45,0 - 19,4 = 25,6\%$ . Сумма однофакторных эффектов в правой части формулы условия аддитивности равна  $\Delta W_1(\Phi P1) + \Delta W_1(\Phi P2) = (W_{10} - W_{00}) + (W_{01} - W_{00}) = (23,5 - 19,4) + (24,4 - 19,4) = 9,3\%$ . Различия между  $\Delta W_2(\Phi P1 + \Phi P2)$  и  $\Delta W_1(\Phi P1) + \Delta W_1(\Phi P2)$  статистически значимы,  $p < 0,01$ . С практической точки зрения это означает, что совместное действие этих ФР может увеличивать заболеваемость значительно сильнее, чем можно было ожидать на основе однофакторных эффектов. Такие сочетания факторов риска наиболее опасны, следовательно, их необходимо устранять в первую очередь. Рассмотренный пример значимого усиления действия факторов риска в паре по сравнению с однофакторными эффектами является примером эффектов более чем аддитивных, или **синергизма** совместного действия пары ФР.

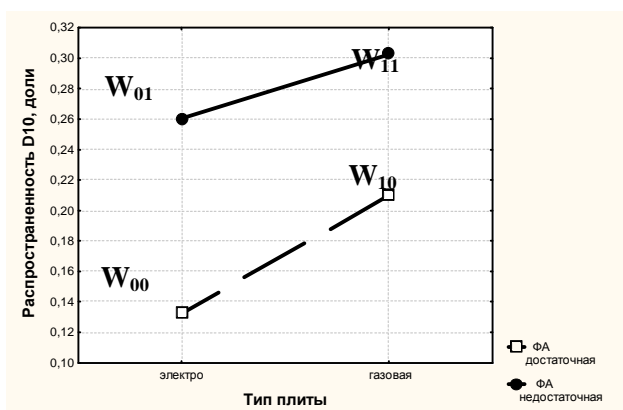


Рис. 3. Двухфакторные эффекты для заболеваний органов дыхания, факторы риска – «Газовая плита» и «Недостаточная физическая активность ребенка»

Анализируя эффект совместного влияния физической активности ребенка ФР1 и типа плиты ФР2 (рис. 3), можно сказать, что тип плиты не является статистически значимым фактором риска (на обеих градациях фактора «Физическая активность»:

$W_{00}$  статистически значимо не отличается от  $W_{10}$  ( $p=0,09$ ), а  $W_{01}$  статистически значимо не отличается от  $W_{11}$  ( $p=0,26$ ). В то же время, фактор «Физическая активность» статистически значимо влияет на распространенность D10 при любом типе плиты:  $W_{00}$  статистически значимо отличается от  $W_{01}$  ( $p=0,030$ ), а  $W_{10}$  от  $W_{11}$  ( $p=0,035$ ). Отметим, что двухфакторный эффект  $\Delta W2(\Phi P1+\Phi P2)=W_{11}(\Phi P1=1;\Phi P2=1) - W_{00}(\Phi P1=0;\Phi P2=0) = 30,2-13,2=17,0\%$  существенно превосходит любой из однофакторных эффектов  $\Delta W1(\Phi P1)=W_1(\Phi P1=1) - W_0(\Phi P1=0) = 10,5\%$  и  $\Delta W1(\Phi P2)=W_1(\Phi P2=1) - W_0(\Phi P2=0) = 6,2\%$ ; при этом условие аддитивности (3) не нарушается. Это пример *аддитивности* действия факторов. Противоположная ситуация имеет место для факторов «Физическая активность ребенка» и «Уровень образования матери» (рис.4). Визуальный анализ показывает, что обеспечение ребенку достаточного уровня физической активности приводит к резкому уменьшению распространенности D13 только у детей, матери которых имеют высшее образование. Для детей, матери которых не имеют достаточного уровня образования (только среднее), уровень физической активности не имеет значения ( $W_{01}$  и  $W_{11}$  очень высокие и статистически не различаются,  $p=0,47$ ). Налицо эффект менее, чем аддитивный. В токсикологии такие эффекты называются антагонизмом; здесь скорее уместен термин «насыщение».

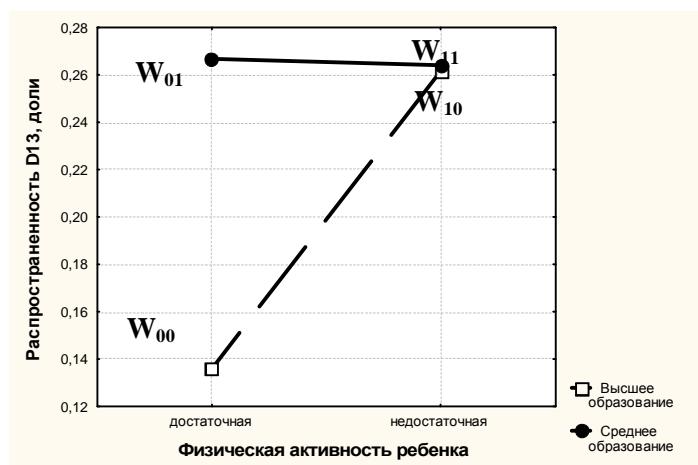


Рис. 4. Двухфакторные эффекты для болезней костно-мышечной системы и соединительной ткани, факторы риска «Недостаточный уровень образования матери», «Недостаточная физическая активность ребенка».

Для многофакторных моделей с числом факторов больше двух, в диссертации приведены примеры решающих правил, построенные методами деревьев классификации, на основе которых могут быть выработаны решения по управлению здоровьем населения. Разработана методика наиболее эффективного нахождения факторов, способных компенсировать неблагоприятное действие экологических (неустраняемых) факторов риска; такими факторами, частично или полностью модифицируемыми, являются социально-экономические факторы семьи.

В качестве примера комплексного анализа многофакторных эффектов приведены результаты анализа 12 деревьев классификации, построенных для заболеваний верхних дыхательных путей (класс D10). Показано, что для этого класса заболеваний «системообразующими» являются факторы: Загрязнение атмосферного воздуха, Тип плиты в квартире (электрическая/газовая) и Уровень физической активности (ФА) ребенка. Именно эти факторы встречаются *во всех* 12 построенных деревьях; именно эти факторы вместе дают наибольшую распространенность заболевания  $W$ , а их одновременное отсутствие – наименьшую  $W$ . Более того, эти факторы демонстрируют возможность взаимной компенсации. Например, в одном из вариантов дерева классификации наличие достаточного уровня физической активности ребенка снижает негативное действие двух других факторов риска – загрязненного атмосферного воздуха и газовой плиты в квартире. В другом варианте дерева классификации электроплита, установленная вместо газовой плиты, снижает (компенсирует) негативное влияние холода и запыленности в квартире ребенка и недостаточный уровень его физической активности. Если учесть, что уровень физической активности тесно связан с образованием матери и материальной обеспеченностью семьи, получаем комплекс социальных и экологических факторов, образующих некую «паутину» причинности для заболеваний органов дыхания.

### **Основные результаты исследования**

1. На основе идеологии иерархической классификации разработана новая методика статистического моделирования многофакторного воздействия на систему при наличии корреляций между факторами, позволяющая интерпретировать результаты исследований, в частности, в области экологии человека и биологии.
2. Разработан пошаговый алгоритм метода «Деревья классификации» – новая версия известного метода классификации применительно к решению задач экологии человека и биологии.
3. Проведены комплексные исследования влияния факторов риска различной природы (загрязнение окружающей среды, социально-экономические факторы) на здоровье детей-дошкольников Екатеринбурга с применением новых технологий математического моделирования.
4. Предложен ряд новых методов коррекции эффектов низкой размерности (в первую очередь, однофакторных эффектов), искаженных взаимосвязями между

факторами: численный метод линейной коррекции (безусловная коррекция), метод, основанный на подборе копий в идеологии «Случай-контроль», коррекция методом маргинальных частот (условная коррекция). Выполненные модельные расчеты показали, что наилучшими корректирующими свойствами обладает метод маргинальных частот. Показано, что широко используемый метод логистической регрессии, будучи примененный для коррекции однофакторных эффектов, дает неудовлетворительные результаты. Показана предметная значимость коррекции в задачах экологии человека и биологии.

5. Для построения и анализа моделей создан комплекс программ, включающий следующие программы:

- программа автоматического анализа двухфакторных эффектов методом полного перебора с выбором наиболее значимых парных эффектов и перекрестных членов;
- программа создания копий в методе «Случай-контроль» с ориентацией на группу с наименьшей численностью объектов исследования и с сохранением параметров исходной выборки;
- программа, реализующая пошаговый полуавтоматический алгоритм построения леса деревьев классификации.

6. Проведено комплексное исследование влияния факторов риска различной природы (загрязнение окружающей среды, социально-экономические факторы) на здоровье детей-дошкольников Екатеринбурга. Найдены факторы и их комплексы, оказывающие наибольшее негативное влияние на распространенность заболеваний органов дыхания и системы кровообращения. Разработана методика нахождения социально-экономических мер, позволяющих компенсировать неблагоприятное действие загрязнения окружающей среды на здоровье детей. Например, выдвинута гипотеза, что негативное действие загрязнения атмосферного воздуха на детей г.Екатеринбурга может быть компенсировано, в ряде случаев, сменой газовой плиты в квартире на электрическую или увеличением физической активности ребенка.

7. Методология комплексного анализа используется в курсах лекций для студентов Уральского федерального университета (имеется акт внедрения). Результаты комплексного анализа внедрены и используются в практике работ Центра детской экопатологии, г.Екатеринбург (имеется акт внедрения).

## **Основные работы, опубликованные по теме диссертации**

### *Публикации в журналах из списка ВАК:*

1. Константинова Е.Д., Вараксин А.Н. Системный подход в изучении влияния комплекса факторов риска на показатели здоровья детей // Информатика и системы управления. 2010. № 2(24). С.186-189.
2. Константинова Е.Д., Вараксин А.Н. Метод «Деревья классификации» в задачах оценки комплексного влияния факторов риска на здоровье детей // Экологические системы и приборы (Москва). 2009. № 10. С.51-54.
3. Константинова Е.Д., Вараксин А.Н. Разработка методики нахождения факторов, компенсирующих неблагоприятное действие загрязнения окружающей среды на здоровье человека // Экологические системы и приборы (Москва). 2010, № 5. С.35-38.
4. Вараксин А.Н., Константинова Е.Д. Эффекты взаимной коррелированности факторов риска при изучении связей «Здоровье населения – факторы риска» // Экологические системы и приборы (Москва). 2009. № 2. С.9-13.
5. Вараксин А.Н., Живодеров А.А., Константинова Е.Д., Жовнер И.В. Применение метода корреляционных плеяд в задачах медико-экологического мониторинга // Экологические системы и приборы (Москва). 2009. № 5. С.51-54.
6. Антонов К.Л., Константинова Е.Д., Вараксин А.Н. Воздействие выбросов автотранспорта на здоровье детей Екатеринбурга // Гигиена и санитария (Москва). 2007. №5. С. 28-32.
7. Константинова Е.Д., Вараксин А.Н., Живодеров А.А., Жовнер И.В. Эколого-социальные факторы и здоровье детей промышленного центра // Уральский медицинский журнал (Екатеринбург). 2007. №11(39). С. 48-52.

### *Остальные публикации*

8. Константинова Е.Д., Вараксин А.Н. Методология системного анализа взаимосвязей между факторами риска и здоровьем населения в задаче устойчивого развития // Международный журнал. Устойчивое развитие: наука и практика (Дубна). 2010. №2(5) ст.3. С. 68-85.
9. Константинова Е.Д., Антонов К.Л., Вараксин А.Н., Чуканов В.Н. Алгоритм анализа влияния факторов окружающей среды на распространенность болезней у детей // Материалы Всероссийской научно-практической конференции, Санкт-Петербург, Военно-медицинская академия, 2004. С.24-25.

10. Константинова Е.Д., Антонов К.Л., Вараксин А.Н. Влияние выбросов автотранспорта на здоровье детей промышленного города // Сборник материалов Пленума научного совета по экологии человека и гигиене окружающей среды, Москва, НИИ экологии человека и гигиены окружающей среды им. А.Н. Сысина, 2006.
11. Константинова Е.Д. Качество питьевой воды и здоровье детей-дошкольников крупного города // Вестник Российской военно-медицинской академии. 2008, № 3(23), приложение 2, часть 2, с.448-449.
12. Константинова Е.Д., Вараксин А.Н., Антонов К.Л. Влияние выбросов автотранспорта на здоровье детей (болезни костно-мышечной системы и психические расстройства и расстройства поведения) // Там же, часть 1, с.134-135.
13. Константинова Е.Д., Вараксин А.Н. Применение метода деревьев классификации при анализе связей «Факторы среды обитания – здоровье населения // Материалы V международной конференции «Экологические и гидрометеорологические проблемы больших городов и промышленных зон». Санкт-Петербург, 7-9 июля 2009 г. СПб.: Крисмас+, 2009. С.104-105.
14. Константинова Е.Д. Взаимосвязанность факторов риска при оценке комплексного влияния окружающей среды на здоровье детей // там же. С.102-104.
15. Konstantinova E.D., Varaksin A.N. Elaboration and application of a new hierarchical classification algorithm in epidemiological research // 23<sup>rd</sup> annual Conference of International Society for Environmental Epidemiology. Barcelona (Spain), 13-16 September 2011. Abstract № 00389.

---

Подписано в печать 9.01.2012	Формат 60×84 <sup>1</sup> / <sub>16</sub>	Бумага писчая
Плоская печать	Тираж 100	Заказ

---

Ризография НИЧ УрФУ  
620002, г. Екатеринбург, ул. Мира, 19