Электронное научное издание
«Международный электронный журнал. Устойчивое развитие: наука и практика»
www.yrazvitie.ru вып. 2 (5), 2010, ст. 3

УДК 004.02, 504.75.05

МЕТОДОЛОГИЯ СИСТЕМНОГО АНАЛИЗА ВЗАИМОСВЯЗЕЙ МЕЖДУ ФАКТОРАМИ РИСКА И ЗДОРОВЬЕМ НАСЕЛЕНИЯ В ЗАДАЧЕ УСТОЙЧИВОГО РАЗВИТИЯ

Константинова Екатерина Даниловна, научный сотрудник Института промышленной экологии Уральского отделения РАН, лаборатория математического моделирования в экологии и медицине Института промышленной экологии Уральского отделения РАН (Россия, Екатеринбург).

Вараксин Анатолий Николаевич, доктор физико-математических наук, профессор, зав. лабораторией математического моделирования в экологии и медицине Института промышленной экологии Уральского отделения РАН (Россия, Екатеринбург).

Аннотация

Целью настоящей работы является разработка методологии анализа многофакторного воздействия на систему при наличии корреляций между факторами. В качестве примера проведен анализ влияния комплекса взаимосвязанных факторов риска на здоровье населения (дети дошкольного возраста). Многофакторный анализ базируется на последовательном рассмотрении одно-, двух- и многофакторных воздействий. При количестве факторов больше двух для нахождения комплекса факторов, оказывающих наибольшее влияние на систему, предложено использовать идею иерархической классификации, в рамках которой разработан и апробирован оригинальный алгоритм. Предложена методика поиска факторов, позволяющих компенсировать (уменьшать) негативное влияние на здоровье населения, обусловленное загрязнением окружающей среды.

КЛЮЧЕВЫЕ СЛОВА: многофакторный отклик, коррелированные факторы, нахождение главных факторов, иерархическая классификация, здоровье населения, компенсация негативного влияния факторов риска.

THE METHODOLOGY OF THE SYSTEM ANALYSIS OF THE INTERCONNECTIONS BETWEEN RISK FACTORS AND THE HEALTH OF POPULATION IN THE SUSTAINABLE DEVELOPMENT ISSUE

Ekaterina Danilovna Konstantinova, research worker at the Institute of industrial ecology of the Ural Department of Russian Academy of Science, the laboratory of the mathematic modeling in ecology and medicine of Institute of Industrial ecology of the Ural branch of the Russian Academy of Sciences (Russia, Ekaterinburg).

Anatoly Nikolaevich Varaksin, the doctor of physical and mathematical sciences, the professor, the manager of the laboratory of the mathematic modeling in ecology and medicine of Institute of Industrial ecology of the Ural branch of the Russian Academy of Sciences (Russia, Ekaterinburg).

ABSTRACT

The aim of the present work is to develop the methodology of the analysis of multiple-factor influence on the system in the presence of the correlation between the factors. As an example the analysis of the influence of a complex of interrelated risk factors on the health of population (preschool age children) is held. Multiple-factor analysis is based on the successive consideration of one-, two- and multiple-factor influences. In the presence of more than two factors to find a complex of factors having the most influential effect on the system, it is suggested to use the idea of hierarchical classification within the framework of which there has been developed and tested an original algorithm. There have been suggested certain methods of the search for factors that allow to diminish the negative influence on the health of population caused by the pollution of the environment.

KEY WORDS: multiple-factor response, correlated factors, the finding of the main factors, the hierarchical classification, the health of population, the compensation of the negative influence of the risk factors.

В рамках современной парадигмы принята триединая концепция устойчивого эколого-социально-экономического развития.

Электронное научное издание
«Международный электронный журнал. Устойчивое развитие: наука и практика»
www.yrazvitie.ru вып. 2 (5), 2010, ст. 3

Первая из них - экономическая (*Economic*) составляющая представляет собой общепринятую форму капитала, используемую в экономическом анализе, и включает в себя здания, промышленное оборудование и машины, компоненты инфраструктуры и т.п., т.е. все то, что создано человеком.

Вторая составляющая – социальная (*Social*) - ориентирована на человека и направлена на сохранение стабильности социальных и культурных систем. Это «располагаемый человеком запас знаний, навыков, способностей и других атрибутов». Очевидно, что интеллект, знания не могут быть оторваны от своего носителя – человека. Таким образом, в современных условиях на первое место в стратегии устойчивого развития общества должны выступать целевые установки, связанные с возрастанием «человеческого капитала». В первую очередь речь должна идти об укреплении здоровья населения, увеличении продолжительности активной жизни (как одного из главных индикаторов здоровья), повышении уровня образования, развитии науки, т.е. в целом - повышении качества жизни.

Третья составляющая - окружающая среда *(Environment).* Оно включает помимо традиционно рассматривающихся природных ресурсов также и созданную человеком городскую среду со всеми присущими городу факторами риска.

Соотношение между аспектами (компонентами) триединой концепции может (и будет) изменяться со временем. Так, доля природных ресурсов с течением времени, скорее всего, начнет уменьшаться. При этом резкое смещение центра тяжести в какую-либо сторону уводит систему в зону неустойчивого развития, при этом возможности замещения компонент из разных составляющих крайне ограничены, если мы хотим оставаться в этой зоне.

В настоящей работе проводится анализ поведения системы, подвергающейся воздействию большого числа факторов различной природы (экономических, социальных, факторов окружающей среды). В качестве примера рассмотрена задача о здоровье населения, подвергающегося воздействию факторов риска различной природы. В общем вопрос ставится следующим образом: как найти наиболее важные факторы, контроль за которыми обеспечивает устойчивое развитие, как снизить негативный эффект факторов риска, которых невозможно избежать [Сорокин, 2009]?

Общая постановка задачи

Пусть имеется набор факторов ФР1, ФР2, ….., оказывающих влияние на отклик W. В нашем классе задач W – доля объектов исследования, обладающих некоторым заданным свойством. Факторы разделяют изучаемые объекты на группы по уровням факторов; например, если фактор ФР1 принимает два значения ФР1=0 и ФР1=1, тогда отклик W

Электронное научное издание
«Международный электронный журнал. Устойчивое развитие: наука и практика»
www.yrazvitie.ru вып. 2 (5), 2010, ст. 3

принимает два значения W(ФР1=0)=W_0 и W(ФР1=1)=W_1. При рассмотрении комплекса из нескольких факторов (два фактора и более) изучаемые объекты разделяются на множество подгрупп, определяемых набором уровней всех факторов комплекса, а отклик принимает значения W_{ijl} , где каждый индекс соответствует определенному фактору (количество индексов равно числу факторов в комплексе).

Примером отклика W описанного типа может служить заболеваемость населения (доля лиц, имеющих изучаемое заболевание W=n/N, где n-число лиц с заболеванием среди полного числа N); примерами факторов, оказывающих влияние на отклик (здоровье населения), являются загрязнение окружающей среды (обычно представляется в виде категорий «Сильное», «Умеренное», «Слабое»), курение (курит/не курит), образование (высшее, среднее специальное, среднее) и др.[Константинова, Вараксин, 2007г.].

Задача ставится следующим образом: найти комплекс факторов и конкретные комбинации уровней факторов, таких, чтобы значения отклика $W_{ijl...}$ были экстремальны (максимальны либо минимальны) при максимально возможной численности групп; возможны варианты условий экстремальности. Пример постановки задачи: найти *комплекс* факторов, оказывающих наибольшее влияние на здоровье населения. Варианты постановки задачи: найти комплекс факторов, оказывающих наибольшее *негативное* влияние на здоровье; найти факторы, позволяющие компенсировать негативное влияние на здоровье заданного комплекса факторов риска.

Задача о здоровье населения (и аналогичные задачи) обладает определенной спецификой.

a) Необходима количественная оценка эффекта комплекса факторов, выполненная для конкретной популяции. Часто встречающиеся в литературе утверждения типа «Загрязнение окружающей среды обуславливает 30% всех заболеваний» без указания популяции, без указания типа и уровней загрязнения – не годятся в принципе. Нужна методология, применимая для конкретных условий; при этом, методология должна допускать применение к любой другой популяции без изменения принципов (работать «в общем случае»).

b) Для того, чтобы методология работала «в общем случае» нужен комплексный (системный) подход, так, чтобы ответ был не абсолютно частный (который может измениться при малейшем изменении первичных данных), а устойчивый к возможным (реально допустимым) изменениям условий существования популяции. Нужен подход, который давал бы не просто числа, но позволял бы выдвигать и проверять предметные гипотезы (примеры приведены ниже).

Электронное научное издание
«Международный электронный журнал. Устойчивое развитие: наука и практика»
www.yrazvitie.ru вып. 2 (5), 2010, ст. 3

c) Кроме требований конкретности и устойчивости, сформулированных выше, построенные модели должны удовлетворять требованиям понятности (интерпретируемости), практической реализуемости и полезности: результаты должны быть понятны специалисту в предметной области (медику-исследователю и медику- практикующему врачу), их можно реализовать на практике с очевидной пользой для пациентов.

Факторы

В предлагаемом подходе факторы являются (с точки зрения математики) категоризированными переменными, имеющими несколько градаций (две-три). Примеры таких факторов для задачи о здоровье населения приведены выше. Существующие подходы к оценке влияния комплекса факторов на отклик W (многофакторный отклик, многофакторный анализ) базируются, чаще всего, на предположении о независимости факторов (фактор ФР1 изменяется независимо от фактора ФР2 и также все остальные), либо на неявном предположении, что взаимосвязь между факторами не оказывает существенного влияния на результат. Основное отличие нашего подхода от имеющихся – явная декларация наличия значимых корреляций между факторами и необходимость учета этих корреляций при определении отклика.

Статистические взаимосвязи между категоризированными факторами с небольшим числом градаций устанавливаются с помощью таблиц сопряженности признаков. Статистическая значимость связи определяется критерием хи-квадрат, сила связи – мерой Крамера (или другими аналогичными мерами). После определения связей между всеми парами факторов, необходимо дать наглядное (графическое) представление всех связей в виде, например, графа корреляционных плеяд [Терентьев, 1977г., Миркин, 1985г.].

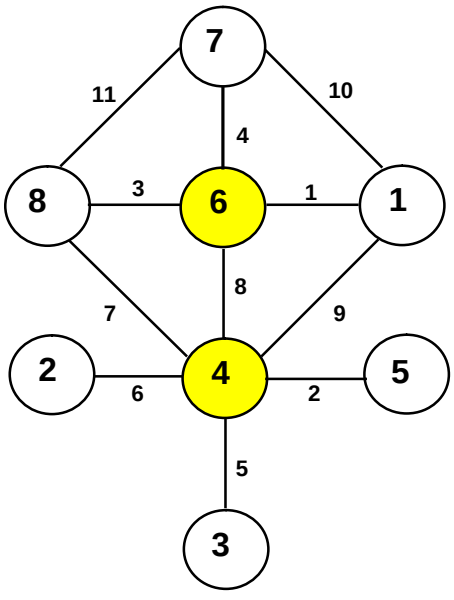

Рис. 1. Граф корреляционной плеяды

В качестве примера приведем результаты исследования взаимосвязей между факторами в задаче о здоровье населения. В работе исследованы 441 семья детей-дошкольников Екатеринбурга, посещающие детские дошкольные учреждения. Факторами, которые оказывают возможное влияние на здоровье детей (W – распространенность какой-либо патологии),

Электронное научное издание
«Международный электронный журнал. Устойчивое развитие: наука и практика»
www.yrazvitie.ru вып. 2 (5), 2010, ст. 3

являются факторы семьи. Графическое представление связей между этими факторами («портрет» семьи в координатах семейных факторов) показано на рис. 1 [Вараксин, Живодеров, 2009г.].

Факторы риска на рис. 1.:

1 – тип питьевой воды

2 – санитарное состояние квартиры

3 – семья полная / неполная

4 – материальная обеспеченность семьи

5 – психологический климат семьи

6 – образование матери

7 – курение матери (есть, нет)

8 – физическая активность ребенка

На рис. 1. 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 «образованием матери» и «тип питьевой воды»; как и следовало ожидать, в семьях с более высоким образованием матери чаще употребляют чистую или фильтрованную воду. Проанализировав граф, мы выделили два «системообразующих» (имеющих наибольшее количество связей) фактора: образование матери (фактор № 6) и материальная обеспеченность семьи (№ 4).

Понятие эффекта

При изучении одного, двух и многих факторов возникают, соответственно, одно-, двух- и многофакторные эффекты. Под однофакторным эффектом бинарного фактора ФР, принимающего значения 0 и 1, мы понимаем величину

$$\Delta W1(ФР) = W_1 - W_0, \quad (1)$$

где W_1 = W(ФР=1), а W_0 = W(ФР=0). В задаче о здоровье населения W_1 и W_0 – это распространенности некоторой патологии при наличии и отсутствии одного фактора риска соответственно. В доказательной медицине величина ΔW1 (1) называется «Добавочный риск – Attributable Risk» [Флетчер, 1998г.].

При наличии двух бинарных факторов ФР1 и ФР2 появляются 4 значения отклика: W_{00}, W_{01}, W_{10} и W_{11} с очевидным смыслом.

Двухфакторным эффектом называют величину:

$$\Delta W2 = W_{11} - W_{00}, \quad (2)$$

где W_{11} = W(ФР1=1, ФР2=1);

Электронное научное издание
«Международный электронный журнал. Устойчивое развитие: наука и практика»
www.yrazvitie.ru вып. 2 (5), 2010, ст. 3

W_{00} = W(ФР1=0, ФР2=0).

В случае двух бинарных факторов возможны 4 однофакторных эффекта, когда эффект одного фактора определяется на фиксированном уровне второго фактора: $\Delta W1(ФР1,ФР2=0)=W_{10}-W_{00}$, $\Delta W1(ФР1,ФР2=1)=W_{11}-W_{01}$, $\Delta W1(ФР2,ФР1=0)=W_{01}-W_{00}$, $\Delta W1(ФР2, ФР1=1)=W_{11}-W_{10}$. Так, например, разность ($W_{10} - W_{00}$) - это эффект, обусловленный первым фактором ФР1, на фиксированном уровне второго фактора ФР2=0.

Хорошим способом осмысления двухфакторных эффектов является их графическое представление. Пример из области здоровья населения приведен на рис. 2. Рисунок наглядно показывает, что распространенность заболевания увеличивается при действии каждого из факторов риска (газовая плита и недостаточный уровень физической активности ребенка), причем их совместное действие близко к аддитивному (прямые линии практически параллельны).

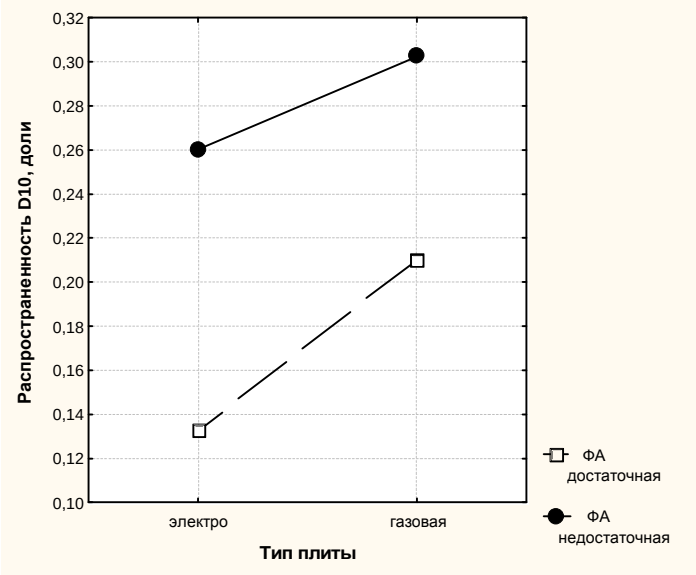

Рис. 2. Распространенность W болезней органов дыхания (класс болезней D10) при совместном действии двух факторов: ФР1-тип плиты в квартире (0-электрическая плита, 1-газовая плита) и ФР2-уровень физической активности ребенка (0-достаточная, 1-недостаточная ФА)

Для науки и практики наиболее интересны так называемые неаддитивные эффекты действия комплекса факторов на отклик W. В случае двух бинарных ФР условие аддитивности выражается соотношением

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

Можно показать, что формула (2) является обобщением соотношения классического дисперсионного анализа, которое (в стандартных обозначениях) имеет вид [Юнкеров, 2005г.]

$$(\overline{y_{ij}} - \overline{y}_{i\bullet} - \overline{y}_{\bullet j} + \overline{y}) = 0 \ . \quad (4)$$

Электронное научное издание
«Международный электронный журнал. Устойчивое развитие: наука и практика»
www.yrazvitie.ru вып. 2 (5), 2010, ст. 3

Формула (2) получается из (3) в предположении, что аналогом среднего значения Y_{ij} в дисперсионном анализе является распространенность патологии W_{ij}.

Согласно представлениям о комбинированном действии факторов, введенным в токсикологии [Кацнельсон, 2002г.], нарушение равенства (2) описывает эффекты антагонизма (когда левая часть (2) меньше правой), либо синергизма (в обратном случае). В первом случае совместный эффект действия двух факторов оказывается меньше суммы однофакторных эффектов, во втором – больше данной суммы. В задаче о здоровье населения это означает, что знание однофакторных эффектов двух ФР при антагонизме позволяет оценить верхний предел угрозы, которую представляют два фактора риска для здоровья населения, а при синергизме однофакторные эффекты не гарантируют такой оценки и возможны сильные неблагоприятные последствия. Естественно, что выполнение (нарушение) равенства (2) должно рассматриваться не в алгебраическом, а в статистическом смысле, т.е. оценка нарушения (2) должна проводиться с помощью статистических критериев на заданном уровне статистической значимости.

Нарушение равенства (2), кроме проявления эффектов антагонизма и синергизма, может также выражаться в предметно-значимых и понятных специалисту эффектах, таких как:

- W_{11} из формулы (2) значительно больше любого однофакторного W_1 из формулы (1); это означает, что одновременное действие двух факторов приводит к *значительному* увеличению W (например, распространенности патологии) по сравнению с действием каждого фактора отдельно;
- W_{00} значительно ниже любого однофакторного W_0; это означает, что одновременное устранение двух факторов приводит к снижению (*резкому* снижению) W по сравнению с устранением одного ФР;
- W_{10} либо W_{01} значительно меньше одного из однофакторных W_1; это означает, что устранение одного ФР способно компенсировать негативное влияние другого ФР.

Влияние коррелированности факторов на отклик

Начнем рассмотрение возможного влияние коррелированности факторов с однофакторного эффекта.

В общем случае проблема формулируется следующим образом. Пусть имеется два фактора ФР1 и ФР2, оказывающих влияние на отклик W. Пусть также между ФР1 и ФР2 имеется статистически значимая взаимосвязь. Для описания этой взаимосвязи используется,

Электронное научное издание
«Международный электронный журнал. Устойчивое развитие: наука и практика»
www.yrazvitie.ru вып. 2 (5), 2010, ст. 3

как отмечалось выше, метод таблиц сопряженности признаков. Пример таблицы сопряженности приведен ниже.

Таблица 1. Таблица сопряженности признаков ФР1 и ФР2 (таблица наблюдаемых частот n_{ij})

Фактор	ФР1=0	ФР1=1	ФР1=0 и 1
ФР2=0	n_{00}	n_{10}	$n_{.0}$
ФР2=1	n_{01}	n_{11}	$n_{.1}$
ФР2=0 и 1	$n_{0.}$	$n_{1.}$	N

В таблице 1 наблюдаемые частоты n_{ij} – это число объектов исследования, которые одновременно принадлежат уровню ФР1= i и ФР2= j. В последней колонке и последней строке таблицы представлены маргинальные частоты, например, $n_{.0}$ равно сумме n_{00} и n_{10}, а $n_{0.}$ равно $n_{00} + n_{01}$. Общее число объектов обозначается N. Если бы факторы ФР1 и ФР2 были независимы, тогда частоты n_{ij} (так называемые ожидаемые частоты n^*_{ij}) были бы равны:

$$n^*_{ij} = \frac{n_{i.} n_{.j}}{N}. \qquad (5)$$

Соотношение (4) выражает тот факт, что вероятность попадания объекта в одну из «клеток» табл. 1, находящуюся на пересечении столбца i и строки j, определяется произведением вероятностей попасть в столбец i и строку j (условие независимости событий).

Представим ситуацию, когда частоты n_{00} и n_{11} оказываются больше соответствующих n^*_{00} и n^*_{11}, а частоты n_{10} и n_{01} – меньше чем n^*_{10} и n^*_{01}. Эта ситуация является аналогом положительной корреляционной связи в том случае, если бы ФР1 и ФР2 были бы количественными переменными. Предположим также, что W(ФР1=1)>W(ФР1=0) и W(ФР2=1)>W(ФР2=0), т.е оба фактора при изолированном действии увеличивают значение отклика W. В этой ситуации однофакторный эффект $\Delta W1(ФР1) = W_1 – W_0$, который должен отражать изменение W при изменении одного фактора ФР1, фактически описывает изменение W при одновременном изменении как ФР1, так и ФР2.

В этой ситуации возникает вопрос: возможно ли определение однофакторного эффекта ФР1 «в чистом виде», без влияния второго фактора ФР2 ? В точной постановке вопроса - ответ отрицательный: невозможно «избавиться» от влияния ФР2 в ситуации, когда ФР2 тесно связан с ФР1. С точки зрения математики отрицательный ответ на поставленный вопрос понятен [Налимов, 1971г.]. Для того, чтобы этот ответ был понятен на уровне «обыденного сознания», проведем аналогию с воздействием на некий отклик W смеси химических реактивов. Итак, пусть имеется смесь химических веществ ФР1 и ФР2,

Электронное научное издание
«Международный электронный журнал. Устойчивое развитие: наука и практика»
www.yrazvitie.ru вып. 2 (5), 2010, ст. 3

вызывающая некоторую реакцию W. Допустим, в проводимых экспериментах делается попытка установить зависимость типа «Доза-ответ», когда реакция W регистрируется при различных дозах химических веществ. Если увеличение дозы организуется путем увеличения объема смеси при неизменном и жестком соотношении количеств первого и второго вещества, тогда и реакция системы наблюдается при одновременном воздействии ФР1 и ФР2. Очевидно, что результаты такого эксперимента не позволяют определить, как действует на W каждое вещество отдельно от другого.

Если же концентрации веществ ФР1 и ФР2 в смеси коррелируют (изменяются синхронно в статистическом смысле), но отсутствует жесткая связь между количеством ФР1 и ФР2, есть возможность сделать оценку порядка величины изменения однофакторного эффекта ΔW1(ФР1) за счет влияния ФР2. Для этого можно использовать так называемую процедуру «линейной интерполяции» для корректировки однофакторных эффектов при наличии коррелированных факторов, описанную в [Вараксин, Константинова, 2009 г.]. Другой способ корректировки однофакторных эффектов основан на использовании данных двухфакторной таблицы. Пусть имеется два бинарных фактора ФР1 и ФР2, оказывающих влияние на отклик W, причем ФР1 считается основным фактором, влияние которого надо изучить, а ФР2 является фактором, «мешающим» (по терминологии доказательной медицины [Флетчер, 1998г.; Привалова, 2003г.]) изучению влияния ФР1. Задача заключается в том, чтобы попытаться выявить влияние ФР1 на W в более «чистом» виде, устранив мешающее влияние ФР2. Для решения поставленной задачи составляется двухфакторная таблица, представленная ниже.

Электронное научное издание
«Международный электронный журнал. Устойчивое развитие: наука и практика»
www.yrazvitie.ru вып. 2 (5), 2010, ст. 3

Таблица 2. Двухфакторная таблица Cell Statistics при одновременном действии на W двух факторов

N	Факторы	Уровни ФР1	Уровни ФР2	Число объектов	W
1	ФР1	0		$n_0.$	W_0 (ФР1)
2	ФР1	1		$n_1.$	W_1 (ФР1)
3	ФР2		0	$n._0$	W_0 (ФР2)
4	ФР2		1	$n._1$	W_1 (ФР2)
5	ФР1 + ФР2	0	0	n_{00}	W_{00}
6	ФР1 + ФР2	0	1	n_{01}	W_{01}
7	ФР1 + ФР2	1	0	n_{10}	W_{10}
8	ФР1 + ФР2	1	1	n_{11}	W_{11}
9	Все объекты вместе			N	$\overline{W}$

Надо отметить, что «устранить» влияние ФР2 можно только путем задания конкретных условий такого устранения. В качестве наиболее естественного условия можно предложить следующее: соотношение распространенностей второго фактора ФР2 на уровне ФР1=0 (это числа n_{00} и n_{01}) и на уровне ФР1=2 (числа n_{10} и n_{11}) должно быть такое же, как для всех объектов вместе на уровнях ФР2 без разделения на уровни ФР1 ($n._0$ и $n._1$). Тогда нескорректированный однофакторный эффект фактора ФР1, введенный ранее соотношением (1), выражается как

$$\Delta W1(\hat{O}\text{Ð}1) = W_1(\hat{O}\text{Ð}1) - W_0(\hat{O}\text{Ð}1) = \frac{W_{10}n_{10} + W_{11}n_{11} - W_{00}n_{00} - W_{01}n_{01}}{N}, \quad (6)$$

а скорректированный (*adjusted*) на влияние ФР2 как:

$$\Delta W1_{adj}(\hat{O}\text{Ð}1) = \frac{W_{10}n._0 + W_{11}n._1 - W_{00}n._0 - W_{01}n._1}{N} \quad (7)$$

Задача выбора факторов и двухфакторных комплексов, оказывающих наибольшее влияние на отклик

Задача выбора факторов (k=1) и двухфакторных (k=2) комплексов выделена в отдельную задачу (отдельную от случая k>2), потому, что она решается точно путем полного

Электронное научное издание
«Международный электронный журнал. Устойчивое развитие: наука и практика»
www.yrazvitie.ru вып. 2 (5), 2010, ст. 3

перебора всех возможных вариантов. В случае $k>2$ такой перебор либо слишком трудоемкий, либо принципиально невозможен.

В случае $k=1$ проводится расчет однофакторных эффектов всех факторов по формуле (1) и отбираются наиболее значимые из них. В техническом плане это простейшая операция, которая с помощью компьютера может быть выполнена для любого исходного числа числа факторов. Например, в задаче о здоровье населения в наших исследованиях первоначальное число факторов равно 150. Число реально значимых факторов, дающих статистически значимые однофакторные эффекты, оказывается гораздо меньше; для разных патологий оно варьируется от 2-3 до 8-10 факторов (гораздо меньше исходных 150).

В случае $k=2$ даже технически задача оказывается гораздо сложнее. В этом случае для каждой пары факторов из первоначально большого числа факторов необходимо построить таблицу типа табл. 1 и, как минимум, сделать оценку величины и статистической значимости двухфакторного эффекта. Если же выполняется программа-максимум, тогда необходимо рассчитать все 4 однофакторных эффекта, оценить их статистическую значимость и определить тип комбинированного действия данной пары факторов.

Даже при большом числе исходных факторов программу-минимум можно выполнить путем полного перебора всех пар факторов. Это возможно двумя путями. Первый – написать собственную компьютерную программу, рассчитывающую двухфакторные эффекты и уровень их значимости. Второй – использовать существующие статистические компьютерные пакеты программ (например, программу Statistica for Windows) для оценки величины и значимости конкретного парного эффекта и собственную подпрограмму, которая работает в среде статистического пакета, с целью организации перебора всех пар факторов. Результаты расчетов заносятся в базу данных, автоматически упорядочиваются по величине эффекта, что позволяет легко отобрать необходимое для дальнейшего анализа число пар факторов, оказывающих максимальное влияние на отклик W.

Именно такой подход реализован нами. Временные затраты оказываются следующими: если путем первичного анализа однофакторных эффектов сократить исходное число факторов до порядка 30, тогда полный перебор всех возможных пар факторов занимает порядка 2-3 часов счета на стандартном персональном компьютере.

Многофакторные эффекты при k>2

Задача поиска многофакторных комплексов ($k>2$), оказывающих на отклик W максимальное влияние, не может быть решена путем полного перебора всех вариантов (всех сочетаний k факторов). Отсюда следует невозможность получения одного абсолютно

Электронное научное издание
«Международный электронный журнал. Устойчивое развитие: наука и практика»
www.yrazvitie.ru вып. 2 (5), 2010, ст. 3

лучшего (с точки зрения заданного критерия) решения. В результате мы попадаем в поле действия широко распространенного принципа множественности моделей [Налимов, 1971г.], согласно которому сложная система может (должна) описываться многими дополняющими друг друга моделями, каждая из которых отражает наиболее полно какую-либо сторону функционирования описываемой системы. Именно такая идеология реализована нами на основе идеи иерархической классификации.

Покажем реализацию данной идеологии на примере задачи о здоровье населения.

Основная задача, которая ставится при изучении многофакторных эффектов – нахождение таких комбинаций факторов и их градаций, которые приводят к резкому повышению (или понижению) распространенности изучаемой патологии W по сравнению с одно- и двухфакторными воздействиями. Какие здесь возникают трудности ?

- большое число многофакторных комбинаций;
- большое число различных градаций для заданной комбинации факторов и соответствующих им значений W_{ij} распространенности патологии;
- для реального (практического) использования полученных многофакторных результатов недостаточно просто перечислить факторы, которые при совместном действии дают максимальный эффект; итогом анализа должна стать формулировка некоего *простого и понятного* специалисту (медику-исследователю и медику-практику) правила, по которому определяются классы детей с низкой и высокой распространенностью патологии. Построение этого правила (решающего правила по терминологии теории классификации) в принципе не может быть алгоритмизировано и является результатом экспертной работы специалиста по анализу данных. Само правило служит для специалиста-медика основой для разработки мероприятий по снижению заболеваемости детей.

Итак, с увеличением числа ФР число различных комбинаций факторов и количество W_{ij} для каждой комбинации факторов увеличивается, в результате чего их экспертный анализ становится непростой задачей; поэтому требуется метод, который позволил бы получать решающее правило минуя стадию «ручного» анализа каждого значения W_{ij}. По нашему мнению, такими свойствами обладают методы классификации.

Нами были проанализированы наиболее известные методы классификации, такие как метод линейной дискриминантной функции, распознавания образов, логистическая регрессия и деревья классификации [Айвазян, 1989г.; Дюк, 2001г.; Загоруйко, 1999г.; Казанцев, 1990г.; Чубукова, 2008г.]. Анализ показал, что ни один метод не может быть

Электронное научное издание
«Международный электронный журнал. Устойчивое развитие: наука и практика»
www.yrazvitie.ru вып. 2 (5), 2010, ст. 3

применен без кардинальных модификаций. Оказалось, что наиболее плодотворная идея содержится в методе деревьев классификации (ДК) – идея последовательного иерархического построения решающего правила.

Оригинальный вариант метода ДК предполагает наличие двух (или более) классов объектов, которые надо разделить на основе анализа набора показателей, характеризующих объекты обоих классов. В нашем случае два класса объектов составляют больные и здоровые дети, а показателями являются факторы риска. По сравнению с оригинальным методом деревьев классификации мы сразу отказались от варианта построения решающего правила, надежно (с высокой точностью) разделяющего здоровых и больных. Такое решение основано на том, что:

- надежное (100-процентное) решающее правило в нашем случае нельзя построить в принципе, поскольку факторы риска лишь увеличивают вероятность заболевания, но не являются его причиной (хорошо известно, что больные и здоровые могут иметь один и тот же набор факторов риска);
- мы считаем, что для нахождения комплекса ФР, повышающего (понижающего) распространенность патологии, достаточно найти два класса детей, в которых распространенность W значительно выше (или ниже), чем в среднем по популяции; при этом максимальная W может быть значительно ниже 100%, а минимальная W – значительно выше нуля.

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 ΔW (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 ФР с максимальными ΔW.

В результате нами предложен алгоритм построения решающего правила, разделяющего детей с низкой и высокой распространенностью патологии [Константинова, Вараксин, 2009г.; Константинова, Вараксин, 2010г.(б)], работу которого можно продемонстрировать примером (рис. 3.).

Электронное научное издание
«Международный электронный журнал. Устойчивое развитие: наука и практика»
www.yrazvitie.ru вып. 2 (5), 2010, ст. 3

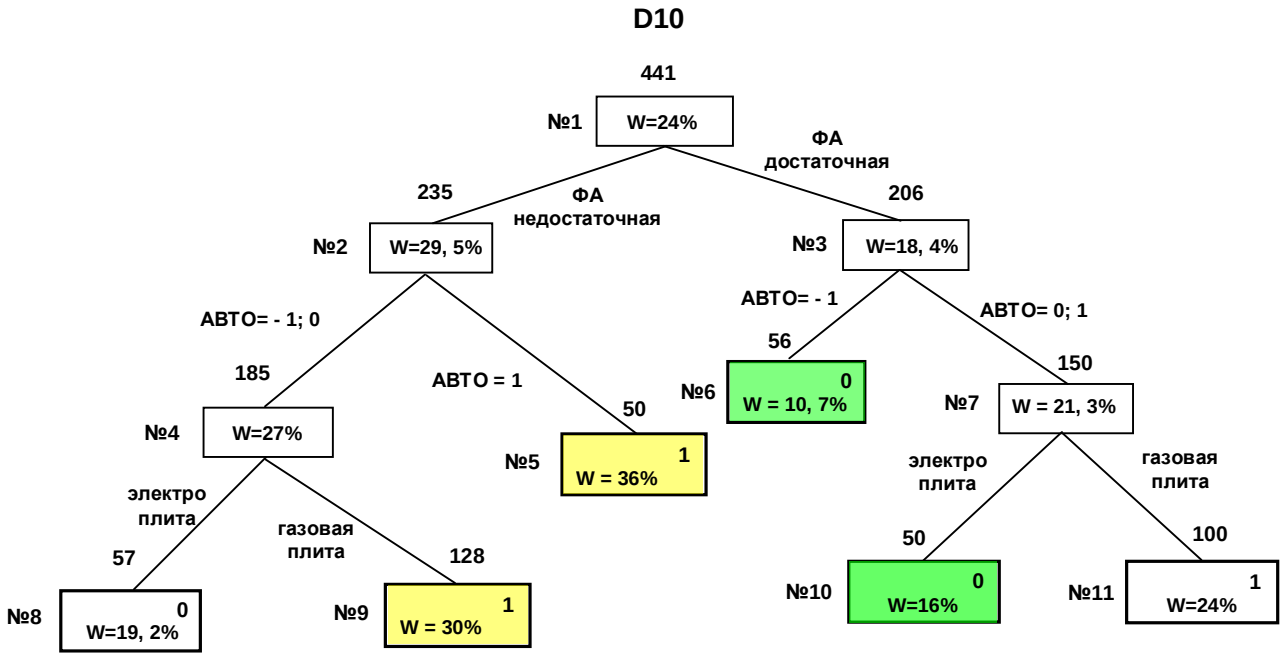

Рис. 3. Дерево классификации для заболеваний верхних дыхательных путей (класс D10)

Алгоритм работает следующим образом.

Первый этап. Построение дерева. В нашем алгоритме для каждого ветвления используется фактор, который дает максимальное различие распространенностей. На рис.6 показано, что разделение по фактору «Физическая активность ребенка – сокр. ФА» дает ΔW=29,5-18,4=11,1% (максимум из всех возможных факторов).

Второй этап. Построение набора ДК. Необходимо построить 12 деревьев, используя для первого ветвления второй и третий по величине ΔW (1) факторы, для второго ветвления используем два варианта (с помощью двух ФР с наибольшими ΔW (1) на этом этапе ветвления), а на третьем и четвертом уровнях используется один ФР с максимальным ΔW. Таким образом, для каждой изучаемой патологии для одного и того же начального списка факторов риска получается 3·2·2=12 различных деревьев классификации (задача вполне решаемая).

Третий этап. Для каждого из 12 деревьев необходимо сформировать два класса детей с низкой и высокой распространенностью исследуемой патологии. На этапе формирования классов мы предлагаем отсечь у деревьев те терминальные вершины, которые имеют W, близкую к средней для всех 441 детей $\overline{W}$.

Электронное научное издание
«Международный электронный журнал. Устойчивое развитие: наука и практика»
www.yrazvitie.ru вып. 2 (5), 2010, ст. 3

Четвертый этап. Для каждого из 12 деревьев надо проанализировать «качество» разделения на классы с высокой и низкой распространенностью заболевания. В нашем подходе «качество» определяется двумя показателями: значением относительного риска RR и числом детей, вошедших в решающее правило. Очевидно, что из одного дерева можно получить несколько правил с различными значениями RR, отсекая разное число терминальных вершин; более высокие значения RR получаются при отсечении большего количества вершин, при этом сокращается число детей в классах и увеличивается неопределенность значения RR (что нежелательно).

Пятый этап. Формулировка предметно-ориентированного решающего правила (РП). Пример такой формулировки РП для данных рис. 2 приведен ниже:

В класс детей с низкой распространенностью D10 попадают дети с достаточным уровнем физической активности в сочетании с проживанием в районе с низким уровнем загрязнения воздуха либо с проживанием в квартире, в которой установлена электрическая плита (электрическая плита компенсирует негативное влияние атмосферного воздуха среднего и высокого уровней загрязнения).

Класс детей с высокой распространенностью D10 характеризуется недостаточным уровнем физической активности в сочетании с проживанием в районе с высоким уровнем загрязнения воздуха либо в квартире с газовой плитой (негативное влияние газовой плиты перевешивает проживание в районе с низким и средним уровнями загрязнения атмосферного воздуха).

Полученные выводы согласуются с данными эпидемиологических исследований о влиянии среды обитания на здоровье детей [Вельтищев, 1995г.; Чеботарев, 2007г.; Экология и здоровье детей, 1998г.]; в отличие от упомянутых работ, в нашем подходе дается количественная оценка действия на здоровье детей комплекса факторов риска.

Очевидно, что решающее правило такого типа понятно специалисту в предметной области (экологическая медицина) и позволяет выработать рекомендации, выполнение которых может снизить заболеваемость детей.

Пример рекомендации: если семья живет в районе с высоким загрязнением воздуха, желательно, чтобы в квартире была установлена электрическая (не газовая) плита; во всех случаях, сильный положительный эффект дает ориентация семьи на активный образ жизни, что позволяет ребенку иметь достаточную физическую активность.

Шестой этап. Анализ промежуточных и отсеченных терминальных вершин.

Промежуточные и отсеченные вершины, не входя в решающее правило, дают важную информацию об изучаемой системе, именно о возможности компенсации действия одних ФР отсутствием других ФР [Константинова, Вараксин, 2010г.(а)].

В качестве примера системного анализа многофакторных эффектов приведем результаты анализа 12 деревьев классификации, построенных для заболеваний верхних дыхательных путей (класс D10). Показано, что для этого класса заболеваний «системообразующими» факторами риска являются загрязнение атмосферного воздуха

Электронное научное издание
«Международный электронный журнал. Устойчивое развитие: наука и практика»
www.yrazvitie.ru вып. 2 (5), 2010, ст. 3

выбросами автотранспорта, Тип плиты в квартире (электрическая/газовая) и Уровень физической активности (ФА) ребенка. Именно эти факторы встречаются ВО ВСЕХ 12 построенных для D10 деревьях классификации (ДК); 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 варианте ДК наличие достаточного уровня физической активности компенсирует негативное действие двух факторов риска – загрязненного воздуха и газовой плиты. В другом варианте ДК электроплита компенсирует наличие в квартире холода, запыленности и недостаточный уровень физической активности ребенка. Если также учесть, что уровень ФА тесно связан с образованием матери и материальной обеспеченностью семьи (см. рис. 1), получаем комплекс социальных и экологических факторов риска, образующих некую «паутину» причинности для заболеваний класса D10.

Эта паутина позволяет ответить на ряд вопросов, на которые невозможно ответить, не имея системной картины процесса. Например, вопрос: что является первичным, когда речь идет о влиянии фактора «Физическая активность ребенка» на заболеваемость детей ? Мы утверждаем, что недостаточная ФА приводит к повышенной распространенности заболевания; возможное утверждение оппонента – повышенная заболеваемость группы детей не позволяет реализоваться достаточной ФА ? Если оставаться на позициях однофакторного анализа, то вопрос останется без ответа; ответ лежит на следующем уровне, где учитываются взаимосвязи между самими факторами риска. Обратившись к ним, обнаруживаем значимую связь уровня физической активности ребенка и уровня образования матери, который можно трактовать так: более образованная мать, понимая важность для организма ребенка физических упражнений, стремится, чтобы ребенок имел достаточный уровень физической активности. В то же время, образование матери едва ли может оказать значимое влияние на распространенность таких заболеваний, которые действительно ограничивают ФА ребенка. Это позволяет высказать обоснованное предположение, что уровень ФА ребенка определяется комплексом социальных факторов семьи, а не распространенностью заболевания. Следовательно, недостаточный уровень ФА ребенка является первичным, а его следствием является повышенный уровень заболеваемости.

Литература

1. Айвазян, С.А. Прикладная статистика: Классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. − М.: Финансы и статистика, 1989. – 607 с.

Электронное научное издание
«Международный электронный журнал. Устойчивое развитие: наука и практика»
www.yrazvitie.ru вып. 2 (5), 2010, ст. 3

2. Вараксин, А.Н. Применение метода корреляционных плеяд в задачах медико-экологического мониторинга / А.Н. Вараксин, А.А. Живодеров, Е.Д. Константинова, И.В. Жовнер. - Экологические системы и приборы. 2009, № 5. – с. 51-54.
3. Вараксин, А.Н. Эффекты взаимной коррелированности факторов риска при изучении связей «Здоровье населения – факторы риска» / А.Н. Вараксин, Е.Д. Константинова. - Экологические системы и приборы. 2009, № 2. – с. 9-13.
4. Вельтищев, Ю.Е. Экопатология детского возраста - Педиатрия. 1995. №4.
5. Дюк, В. Data mining / В.Дюк, А.Самойленко. − СПб.: Питер, 2001. – 368 с.
6. Загоруйко, Н.Г. Прикладные методы анализа данных и знаний / Н.Г. Загоруйко – Новосибирск: Изд-во Института математики, 1999. – 270 с.
7. Казанцев, В.С. Задачи классификации и их программное обеспечение / В.С. Казанцев. − М.: Наука, 1990. – 136 с.
8. Кацнельсон, Б.А. Общая токсикология / под ред. Б.А. Курляндского и В.А. Филова. – М.: Медицина, 2002. – глава 14.
9. Константинова, Е.Д., Вараксин, А.Н., Живодеров, А.А., Жовнер, И.В. Эколого-социальные факторы и зд – с. 48-50.
10. Константинова, Е.Д., Вараксин, А.Н. Метод «Деревья классификации» в задачах оценки комплексного влияния факторов риска на здоровье детей - Экологические системы и приборы. 2009. № 10. – с. 23-28.
11. Константинова, Е.Д., Вараксин, А.Н.(а) Разработка методики нахождения факторов, компенсирующих неблагоприятное действие загрязнения окружающей среды на здоровье человека - Экологические системы и приборы, 2010. №5 – с. 35-38.
12. Константинова, Е.Д., Вараксин, А.Н.(б) Системный подход в изучении влияния комплекса факторов риска на показатели здоровья детей - Информатика и системы управления, 2010. №2(24). – с. 186-189.
13. Миркин, Б.Г. Группировки в социально-экономических исследованиях. Методы построения и анализа / Б.Г. Миркин. - М.: Финансы и статистика, 1985. – 223 с.
14. Налимов, В.В. Теория эксперимента / В.В. Налимов. – М.: Наука, 1971. – 208 с.
15. Привалова, Л.И. Экологическая эпидемиология: принципы, методы, применение / Л.И. Привалова, Б.А. Кацнельсон, С.В. Кузьмин и др. - Екатеринбург, 2003. – 276 с.
16. Сорокин, Ю.А. Устойчивое развитие: многокритериальная оценка и выбор проектов при выведении озоноразрушающих веществ из потребления. - Устойчивое инновационное

развитие: проектирование и управление. Электронное научное издание. 2009. том 3. – 10 с. URL: http: // rypravlenie.ru.

17. Терентьев, П.В. Практикум по биометрии / П.В. Терентьев, Н.С. Ростова. - Л.: ЛГУ, 1977. - 152 с.
18. Флетчер, Р. Клиническая эпидемиология. Основы доказательной медицины / Р. Флетчер, С. Флетчер, Э. Вагнер. - М.: Медиа Сфера, 1998. – 345 с.
19. Чеботарев, П.А. Оценка состояния здоровья детского населения, проживающего в городах с различным загрязнением атмосферного воздуха – Гигиена и санитария. 2007. №6.
20. Чубукова, И.А. Data Mining / И.А. Чубукова. - М.: Изд. дом «Бином», 2008. – 382 с.
21. Экология и здоровье детей / под ред. М.Я. Студеникина, А.А. Ефимовой. – М.: Медицина, 1998. – 384 с.
22. Юнкеров, В.И. Математико-статистическая обработка данных медицинских исследований / В.И. Юнкеров, С.Г. Григорьев. - СПб.: ВМедА, 2005. – 266 с.