УДК 004.435 004.423

ГЕОИНФОРМАЦИОННЫЕ СИСТЕМЫ И МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ ПРОСТРАНСТВЕННОГО АНАЛИЗА ДЛЯ ИССЛЕДОВАНИЯ ЗАБОЛЕВАЕМОСТИ

Д. Р. Струков,
генеральный директор
ООО «Центр пространственных исследований»
В. Л. Горохов,
доктор техн. наук, профессор
Санкт-Петербургский государственный инженерно-экономический университет

Рассматривается модификация средств геостатистического анализа на основе ранговой «нормализации» данных. Предлагается использовать систему динамической визуализации многомерных данных для контроля достоверности и надежности кокригинга. На основе подобной модификации геостатистического анализа разработаны алгоритмы анализа многомерных медицинских данных и показано их применение в практических задачах здравоохранения.

Ключевые слова *— геостатистический анализ, система динамической визуализации многомерных данных, кокригинг, алгоритмы анализа многомерных медицинских данных.*

Введение

Пространственная статистика помогает руководителям здравоохранения в решении следующих задач:

— оптимизации и управления ресурсами здравоохранения;

— логистики;

— анализа тенденций пространственно-временного распространения заболеваемости, прогноза;

— выявления причинно-следственных связей между факторами среды и показателями здоровья.

Геостатистические методы многомерного анализа, реализованные в геоинформационных системах, «помогают» найти взаимосвязи между факторами (их может быть до пяти) и откликами на уровне тенденций. Можно выявлять «источник» того или иного фактора с точки зрения пространственного расположения. И, наконец, современные методы многомерной когнитивной визуализации [1–4] позволяют наглядно строить пространственные модели, учитывая статистическую точность, значения соседей.

При работе использовались данные Санкт-Петербургского государственного учреждения здравоохранения «Медицинский информационно-аналитический центр» СПбМИАП Комитета по здравоохранению Правительства Санкт-Петербурга. В частности, данные ежегодной медицинской статистической отчетности и диспансеризации детей в 2002 г.

Математические методы пространственного анализа

Среди традиционных методов картирования событий, визуализации и моделирования пространственно-временных закономерностей и явлений имеются также более сложные: многомерный анализ территории; интерполяционный анализ, когда необходимо провести пространственный анализ при помощи детерминированных математических методов и спрогнозировать значения между точками пространства (например, моделирование нахождения пыли в атмосфере); геостатистический анализ [5], когда необходимо провести прогноз распространения тенденций в пространстве как в интерполяционном анализе, так и при помощи методов статистики найти взаимосвязи между точками, значения которых отображают зависимость значений одного тематического слоя от значений другого тематического слоя (а может, и нескольких). Геостатистиче-

ские данные представляются в виде пространственно-зависимого показателя Z_i (высота, глубина, концентрация поллютанта, минерала и т. д.). Известны значения этого показателя на конечном наборе опорных точек $p(i)$. Одним из мощнейших методов геостатистического анализа является метод кригинга. В рамках этого метода предполагается, что Z_i является случайным процессом со стационарными приращениями и заданной вариограммой (или ковариационной функцией). Считается, что вариограмма известна или специально оценивается. Требуется построить функцию $Z = p(i)$, чтобы ее значения в опорных точках были приблизительно равны z_i (аппроксимация). Процедура интерполяции задается системой линейных уравнений кригинга. Неизвестные коэффициенты определяются из условий несмещенности оценки и минимизации ее дисперсии. Для оценки вариограммы используется набор модельных функций вариограмм.

Кригинг предполагает модель тренда и случайной ошибки и приводит к уравнению вида $Z(s) = \mu(s) + \varepsilon(s)$, где s — местоположение предсказываемой локации; $Z(s)$ — предсказываемое значение; $\mu(s)$ — детерминированный тренд; $\varepsilon(s)$ — пространственно-коррелированная случайная ошибка. В случае набора электронных карт-слоев возникает многомерный эквивалент кригинга— кокригинг.

Однако подобные геостатистические методы многомерного корреляционного и дисперсионного анализа требуют дополнительных мер для нормализации геостатистических данных в целях увеличения устойчивости (робастности) методов. В данной работе предлагается при использовании методов кригинга осуществлять локальное нормирование данных на основе ранговых и квантильных статистик. Это приводит к асимптотической стабилизации применяемых оценок.

Кроме того, для изучения качества вычисляемых мер статистической близости показателей в многомерном варианте используются методы когнитивной визуализации многомерных данных [6], представленных в электронных картах.

Одним из самых перспективных в настоящее время методов, способных решать разнообразные задачи визуализации многомерных данных, является метод когнитивных динамических проекций [7, 8], развиваемый авторами для оценки пространственного риска действия факторов (экологических, медико-демографических, социальных и пр.).

Суть этого метода [8] заключается в том, что многомерные данные проецируются на выбранную картинную плоскость и при этом осуществляется динамическое вращение «облака» данных в многомерном пространстве. Для этого вычисляются направляющие косинусы вектора **b** направления вращения. Теперь, зная вектор **b**, можно произвести поворот вектора нормали **N** вокруг начала координат на произвольный угол γ. Новое значение нормали будет вычисляться по формуле $\mathbf{N}_{нов} = \mathbf{N}\cos\gamma + \mathbf{b}\sin\gamma$.

Получив новое значение нормали, задают новую плоскость проекции, повернутую на угол γ относительно исходной. Остается только спроецировать на эту плоскость облако точек, чтобы получить новый взгляд на множество наблюдаемых объектов. Таким образом, задав угол поворота γ, процесс поворота можно зациклить, организовав тем самым циклический обзор данных. При этом, в случае небольшого значения γ, у исследователя будет создаваться ощущение объема и структуры вращения псевдотрехмерного образа, адекватно представляющего многомерный набор данных.

Алгоритм анализа многомерных медицинских данных

Авторы попытались объединить в виде алгоритма анализа многомерных медицинских данных различные методы пространственной визуализации (кокригинг и динамическую визуализацию) многомерных территориально-распределенных медицинских данных о населении. Благодаря уникальным данным о здоровье популяции появилась возможность учитывать в медико-экологическом исследовании так называемый коэффициент здоровья популяции. Мы представляем методические аспекты алгоритма пространственного анализа характеристик здоровья, отклонений и их тяжести, а также экологических факторов (далее — алгоритм ПАЗФ).

Прежде чем описывать алгоритм ПАЗФ, обратимся к медико-эпидемиологической модели. В работе учитываются только те заболевания, экологическая обусловленность которых доказана многократными клиническими исследованиями, приведенными в литературе [9]. В методе исследуются классы болезней по Международной классификации болезней 10-го пересмотра (МКБ-10), а также отдельные группы заболеваемостей, объединенные в зависимости от того или иного экологического фактора или группы факторов.

Проведем рассмотрение на примере характеристик болезней органов дыхания (БОД) и тяжести последствий этого класса заболеваний. В качестве экологических факторов в работе рассматриваются химические вещества в атмосфере и почве Санкт-Петербурга.

Алгоритм ПАЗФ.

Цель: исследование характеристик здоровья, отклонений от него и тяжести последствий этих отклонений у населения, проживающего на тер-

ритории города. Выявление причин реакций популяций, проживающих на той или иной территории города, в зависимости от действия факторов загрязнения окружающей среды. Сравнение откликов населения мегаполиса (на примере Санкт-Петербурга) с реакциями популяций других городов по специальным коэффициентам, учитывающим здоровье.

Алгоритм (рис. 1) [6]:

1) условия выбора ареалов $А_1...А_K$;

2) условия выбора ареалов $Э_1...Э_N$;

3) выделение ареалов проживания $А_1...А_K$;

4) выделение ареалов экологических факторов $Э_1...Э_N$;

5) подсчет характеристик популяций в ареалах $А_1...А_K$;

6) получение коэффициента отклонения и тяжести $К_{о.т}$ по $А_1...А_K$;

7) получение коэффициента здоровья $К_з$ по $А_1...А_K$;

8) получение характеристик экологических факторов по ареалам $Э_1...Э_N$;

9) получение зависимости $К_{о.т} = f(К_з)$ для $Э_1...Э_N$;

10) сравнение графиков $К_{о.т} = f(К_з)$ внутри города для $Э_1...Э_N$;

11) получение коэффициента «силы эффекта» (Power), объединяющего характеристики отклонения и тяжести и здоровья внутри ареалов $Э_1...Э_N$;

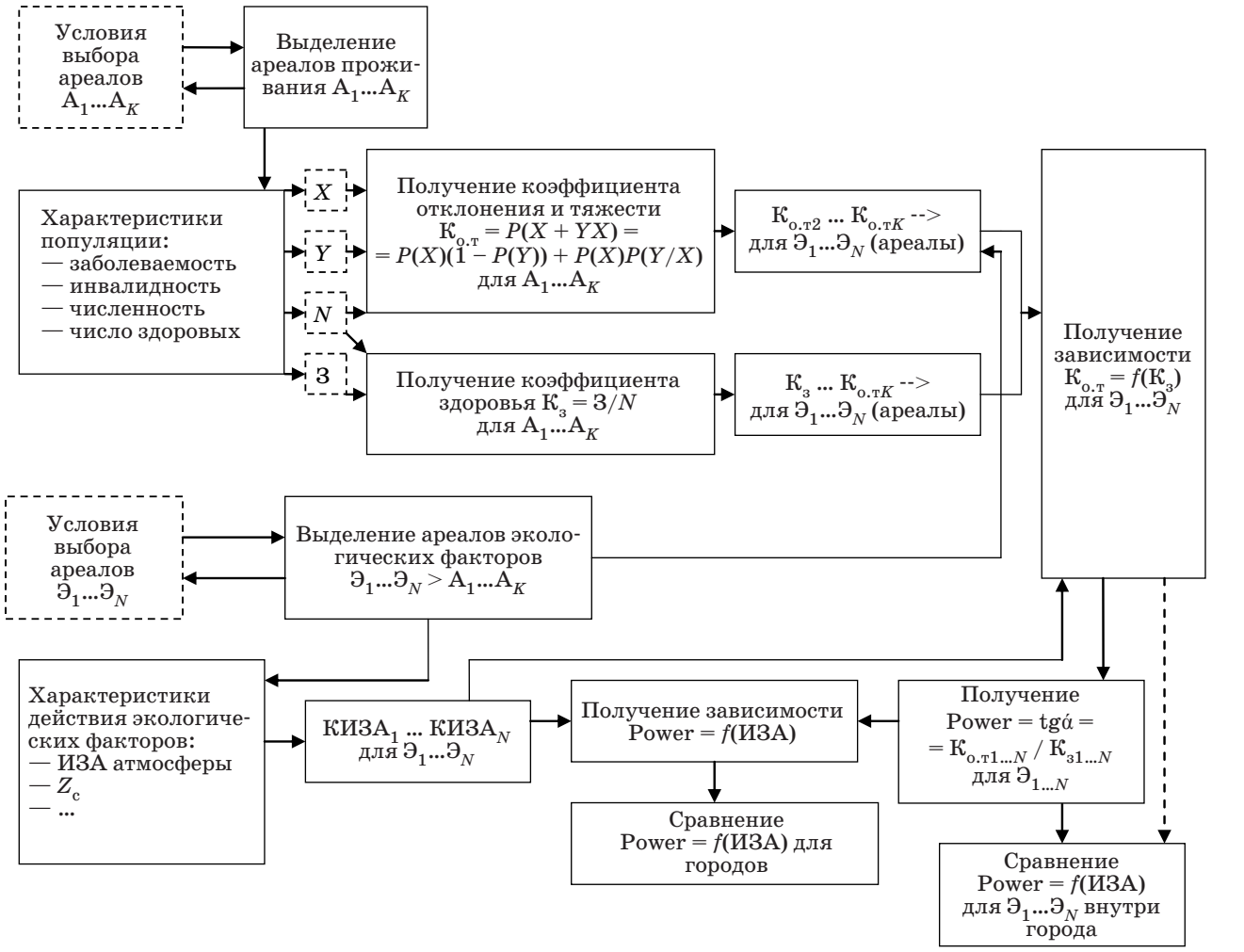
12) сравнение графиков $К_{о.т} = f(К_з)$ внутри города для $А_1...А_K$;

13) получение зависимости Power = f(ИЗА);

14) сравнение Power = f(ИЗА) для Санкт-Петербурга с данными по другим городам (ИЗА — индекс загрязнения атмосферы).

Условия выбора ареалов $А_1...А_K$ для характеристик здоровья населения.

Ареалы $А_1...А_K$ — территориально сгруппированные площадные единицы, главным свойством которых является проживание внутри них населения. Главное условие группировки $А_1...А_K$ — статистически значимое накопление данных внутри каждого ареала. Таким образом, на уровне Санкт-Петербурга такими ареалами могут быть:



■ ***Рис. 1.*** *Алгоритм ПАЗФ*

■ ***Рис. 2.*** *Выбор ареалов первым способом*

1) сумма жилых домов — кварталы (избирательные участки) (рис. 2);

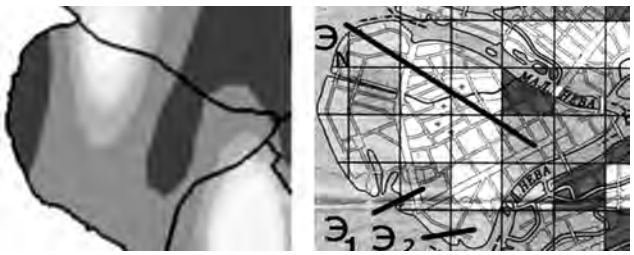
2) сумма кварталов — топонимы (рис. 3), которые объединены по двум критериям: либо постоянная численность N населения (а площади S различны), либо постоянная площадь (а численность населения различна).

Второй способ представляется более интересным с точки зрения пространственного анализа, однако для примера мы будем в качестве ареалов $A_1 ... A_K$ использовать суммы домов по кварталам. Кроме того, этот способ позволяет анализировать характеристики по данным, показывающим, что среднее количество детей в квартале колеблется около 200 человек.

Условия выбора ареалов $Э_1 ... Э_N$ *для характеристик экологических факторов.*

В результате анализа литературы [1–6] мы обозначим три способа выделения ареалов, по которым произведем расчет характеристик факторов.

1. Метод сетки (рис. 4) предполагает разделение пространства на равные квадраты — получение сетки. Сетка «накладывается» на топологию города и на распределение характеристик фактора в пределах города. Метод получения сетки представлен в литературе [8] (по исследованиям, приведенным там, размеры каждой ячейки сетки для Санкт-Петербурга 3 × 3 км, однако в рамках каждого конкретного исследования размер ячейки может быть пересмотрен). Недостаток метода в том, что в некоторых зонах значения фактора могут быть недостоверны (например, модели загрязнения воздуха по пунктам мониторинга).



■ ***Рис. 4.*** *Метод сетки*

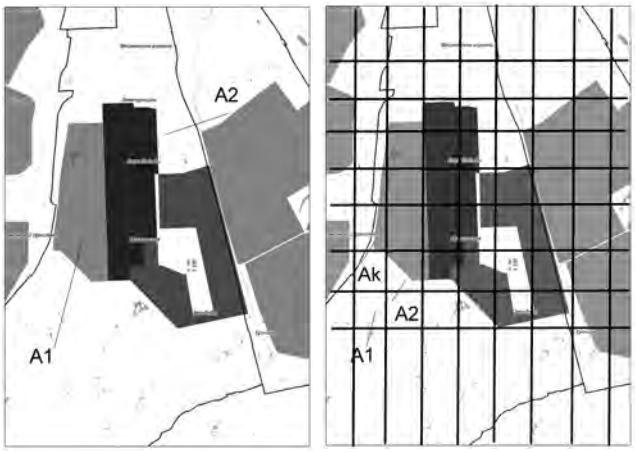
2. Метод соседства с пунктами мониторинга (рис. 5) предполагает выделение определенного пространства (круг, квадрат, полигон в пределах кварталов) вокруг существующих 50 пунктов мониторинга атмосферного воздуха Территориальным управлением «Роспотребнадзор». Принимается [4], что пункты распределены в пространстве достаточно репрезентативным образом и учитывают структуру городской местности (селитебные зоны, промышленные зоны, зеленые насаждения, перекрестки дорог и пр.). Именно в пределах этих зон значение фактора будет приближено к истинному. Однако существуют целые территории в Санкт-Петербурге, где в районе жилых массивов нет пунктов мониторинга или их меньше, чем нужно для анализа, — в этом недостаток метода.

3. Метод, учитывающий структуру города (рис. 6), где основной упор делается на:

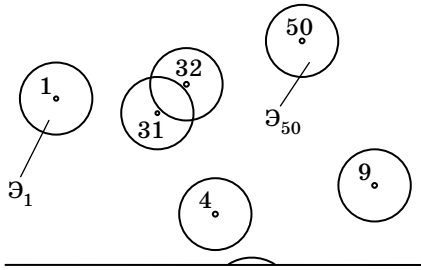
— выделении статистически значимых зон для анализа;

— учете разнообразной структуры внутри города.

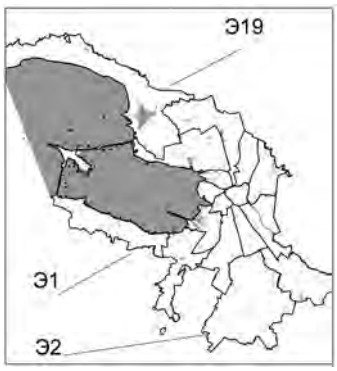
Под структурой здесь понимается как сама инфраструктура города в целом, так и параме-



■ ***Рис. 3.*** *Выбор ареалов вторым способом*



■ ***Рис. 5.*** *Метод соседства с пунктами мониторинга*



■ ***Рис. 6.*** *Метод, учитывающий структуру города*

тры «разброса» населения внутри зон, а также другие демографические и социальные показатели и типы.

Самым простым и грубым примером выделения таких зон в качестве ареалов $Э_1...Э_N$ может быть разбиение территории по административным районам. Недостаток такого разбиения — слишком большая разность площадей районов и неравномерность соотношения структуры города между собой. Для более углубленного пространственного анализа в пределах городской территории необходимо комбинировать этот и другие методы разбиения пространства по ареалам $Э_1...Э_N$ [8].

Однако для нашего исследования должно быть одно обязательное условие — ареалы с экологическими факторами должны «покрывать» ареалы с характеристиками населения, т. е. $A_1...A_K < Э_1...Э_N$ (рис. 7).

Характеристики популяций по ареалам $A_1...A_K$.

Перечислим основные характеристики популяций, которые будем оценивать в рамках нашей модели медико-экологического исследования:

■ ***Рис. 7.*** *Выбранные ареалы*

X — число больных определенным классом болезни (или группой заболеваний);

N — численность детей;

Y — число инвалидов, имеющих заболевание и получивших инвалидность от данного заболевания;

З — число здоровых детей.

Так, например, для БОД берется соответствующее число больных и инвалидов от БОД для каждого квартала $A_1...A_K$.

Получение коэффициента отклонения и тяжести.

Если рассматривать полную группу событий, происшедших с детьми на дату осмотра, то с точки зрения характеристик отклонения и последствий мы можем выделить следующие:

— ребенок заболел без последствий;

— ребенок заболел и получил инвалидность по истечении болезни;

— ребенок заболел и умер.

Последний случай в проекте «Диспансеризация детей 2002» не рассматривается, поэтому во введенной характеристике отклонения и тяжести мы будем учитывать первые две группы событий. Под *отклонением* подразумевается характеристика заболеваемости населения, а под *тяжестью* — инвалидность, возникшая в результате появления той или иной заболеваемости.

Итак, коэффициент отклонения и тяжести $К_{о.т}$ равен сумме вероятности того, что при заболевании инвалидность не возникнет, и вероятности возникновения инвалидности как следствия заболевания, т. е.

$$К_{о.т} = P(X + YX) = P(X)(1 - P(Y)) + P(X)P(Y/X).$$

Коэффициент $К_{о.т}$ учитывает не только само отклонение, но и тяжесть последствий на территории ареалов $A_1...A_K$. Так, основной вклад в значение $К_{о.т}$ для БОД будет иметь слагаемое, характеризующее заболеваемость, а, к примеру, для врожденных деформаций и хромосомных нарушений значение $К_{о.т}$ будет сильно зависеть от показателя инвалидности для данного вида патологии.

Коэффициент здоровья $К_з$.

Под «здоровыми» в результате анкетирования «Диспансеризации детей 2002» мы будем понимать тех детей, которые на момент обследования:

— не имеют диагноза;

— не имеют инвалидности;

— относятся к категории «I группа здоровья»:

$$К_з = З/N.$$

Коэффициент $К_з$ характеризует долю здоровых на территории $A_1...A_K$.

Получение характеристик экологических факторов по ареалам $Э_1...Э_N$.

В зависимости от выбранного метода выделения области $Э_1...Э_N$ (для нашего случая — это административные районы) и с учетом модели медико-экологического исследования мы суммируем по ареалам $Э_1...Э_N$ следующие характеристики:

— ИЗА для одного вещества (компонент — атмосфера);

— комплексный индекс загрязнения атмосферы (КИЗА) для определенных групп химических веществ (компонент — атмосфера);

— показатель суммарного загрязнения почв Z_c (компонент — почвы);

— активность объемную (АО) (для ионизирующих излучений);

— прочие.

В работе в соответствии с моделью исследования учтены экологические факторы атмосферы и почвы по ареалам $Э_1...Э_N$ и берутся показатели ИЗА, КИЗА, усредненные по ареалам $Э_1...Э_N$ (по ареалам административных районов).

Например, к экологическим факторам, которые могут вызвать БОД, можно отнести, прежде всего, химические вещества, находящиеся в атмосфере города: пыль, окислы азота, углерода, серы, аммиак, углеводороды, фенол, сернистый ангидрид.

Для каждого из ареалов $Э_1...Э_N$ строятся таблицы характеристик выбранных ареалов (табл. 1). Таким образом получается N таблиц (по числу ареалов $Э_1...Э_N$) и по каждой таблице строится N графиков зависимостей коэффициента отклонения и тяжести от коэффициента здоровья ($К_{о.т} = f(К_з)$) для $А_1...А_K$.

Сравнение графиков $К_{о.т} = f(К_з)$ *для различных ареалов.*

Кривые демонстрируют зависимость $y = -kx$ и показывают физическое соотношение здоровых и больных на различных территориях $А_1...А_K$ (для каждого $Э_1...Э_N$). Причем характер кривой (ее наклон) должен изменяться в зависимости от величины характеристик фактора (например, КИЗА). На рис. 8 показаны два случая зависимости $К_{о.т} = f(К_з)$ для $Э_1$ и $Э_2 < Э_1$.

Таким образом, сравнение графиков для каждого ареала $Э_1...Э_N$ дает характеристики откликов популяции (отклонений и тяжести) в зависимости от действия факторов внутри того или иного ареала. Чем меньше значение фактора, тем зависимость $К_{о.т} = f(К_з)$ имеет меньший наклон.

Получение коэффициента силы эффекта (*Power*).

Критерием сравнения двух и более графиков $К_{о.т} = f(К_з)$ для $Э_1...Э_N$ является коэффициент силы эффекта, характеризующий зависимость между отклонением, тяжестью и здоровьем популяции внутри ареалов и действия факторов $Э_1...Э_N$ (рис. 9):

Power = (*отклонение* + *тяжесть*) / *здоровье.*

Найти его можно по тангенсу угла наклона прямой $К_{о.т} = f(К_з)$. Чем больше фактор $Э_1$, тем больше коэффициент силы эффекта:

$$\text{Power} = \text{tg}\acute{\alpha} = К_{о.т} / К_з = P(X + YX) / З/N.$$

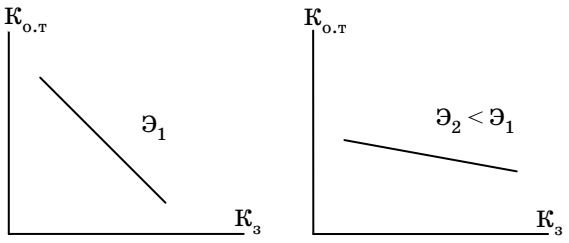
Если зависимости $К_{о.т} = f(К_з)$ нет, то Power $\rightarrow$ $\rightarrow 0$. Это значит, что на той или иной территории $Э_i$ из $Э_1...Э_N$ экологический фактор (или их сумма) не вызывает явных откликов у популяции, проживающей на этой территории.

К примеру, Power для БОД по всем исследуемым районам показал явные зависимости, причем у всех различные. В то время как, например, Power для болезней кожи и подкожной клетчатки практически равен нулю.

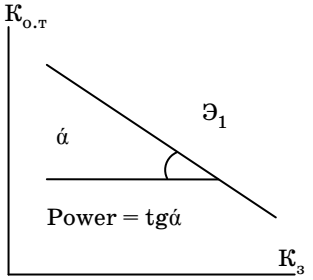
Получение зависимости силы эффекта от характеристик фактора.

■ ***Таблица 1***

Ареал	Число больных x	Численность N	Заболеваемость $P(X) = x/N$	Число инвалидов y	Коэффициент отклонения и тяжести $К_{о.т}$	Число здоровых З	Коэффициент здоровья $К_з$
$А_1$							
$А_2$							
...							
$А_K$							



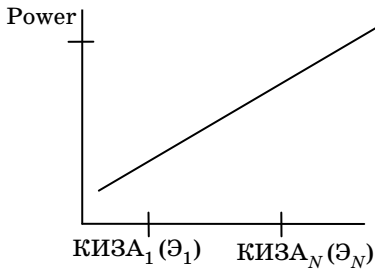
■ ***Рис. 8.*** *Графики зависимости* $К_{о.т} = f(К_з)$ *для различных ареалов*



■ ***Рис. 9.*** *График зависимости* $К_{о.т} = f(К_з)$

■ *Таблица 2*

Ареал	Район	КИЗА	Power
$Э_1$	Невский		
$Э_2$	Приморский		
...			
$Э_N$			



■ ***Рис. 10.*** *График зависимости «силы эффекта» от КИЗА*

Таким образом, после подсчета tgά для каждого графика $К_{о.т} = f(К_э)$ (в соответствии с ареалами $Э_1...Э_N$) мы получаем таблицу зависимости силы эффекта от характеристик фактора (табл. 2).

По данной таблице строится график, отражающий «силу эффекта» действия фактора на территории всех ареалов $Э_1...Э_N$, т. е. на территории всего города (рис. 10). Эта прямая аналогична прямой «доза-ответ» в оценках медико-экологических рисков. Чем больше фактор, тем больше Power [9].

Заключение

1. Рассмотренный алгоритм ПАЗФ может быть применен для достаточно широкого круга медицинских исследований, однако данные должны содержать ровно такую же информацию (по своим свойствам), как и в приведенном примере.

2. Кривая зависимости «силы эффекта» от характеристик экологических факторов, регистрируемых на тех или иных территориях города (например, ИЗА), характеризует распределение характеристик экологических факторов по Санкт-Петербургу в целом.

3. По этой кривой можно сравнивать и части мегаполиса, например, север, юг и центр или пригороды и центр.

4. По данной зависимости можно делать выводы о нахождении экологических факторов в крупных городах и оценивать характерные отклики жителей города на действие этих факторов. Для таких медико-экологических исследований при помощи алгоритма ПАЗФ подойдут крупные российские города, где проводился проект «Диспансеризация детей 2002».

Литература

1. **Бузников А. А. и др.** Обработка экологической информации в системе медико-экологического мониторинга окружающей среды // Дистанционное зондирование земных покровов и атмосферы аэрокосмическими средствами: Материалы 2-й Всерос. конф., 16–18 июня 2004 г. СПб., 2004. С. 9–10.
2. **Красильников И. А., Мерабишвили В. М.** Геоинформационные технологии в онкологии // Новые информационные технологии в онкологической статистике: Материалы Всерос. симпозиума с международным участием / Под ред. В. М. Мерабишвили. СПб., 2001. С. 252–255.
3. **Красильников И. А. и др.** Географические информационные системы в управлении здравоохранением Санкт-Петербурга // ArcReview. Современные геоинформацонные технологии. Спецвыпуск к 300-летию Санкт-Петрбурга. М.: ООО «Дата+», 2003. С. 3–5.
4. **Красильников И. А., Струков Д. Р., Разгуляев К. А.** Внедрение системы медико-экологического мониторинга окружающей среды на базе геоинформационных технологий // Воздух 2004: Материалы Междунар. конф., 9–11 июня 2004 г. / Под ред. Н. З. Битколова и Ю. И. Мусийчука. СПб., 2004. С. 17–18.
5. **Петров Е. И., Струков Д. Р., Красильников И. А.** Геоинформационные технологии в здравоохранении // Региональная информатика 2002: Материалы Юбилейной 8-й Санкт-Петербургской междунар. конф., 26–28 ноября 2002 г. СПб., 2002. С. 23–24.
6. **Попов Г. А., Кононенко Д. В., Струков Д. Р.** Пространственно-временной анализ распределения вредных веществ в атмосфере Санкт-Петербурга // Воздух 2004: Материалы Междунар. конф., 9–11 июня 2004 г. / Под ред. Н. З. Битколова и Ю. И. Мусийчука. СПб., 2004. С. 56–58.
7. **Струков Д. Р.** Пространственно-временной анализ распределения вредных веществ в атмосфере Санкт-Петербурга. Сезонная пространственная изменчивость их распространения // Воздух 2004: Материалы Междунар. конф., 9–11 июня 2004 г. / Под ред. Н. З. Битколова и Ю. И. Мусийчука. СПб., 2004. С. 58–59.
8. **Горохов В. Л., Муравьев И. П.** Когнитивная машинная графика. Методы динамических проекций и робастная сегментация многомерных данных. Методология, методики и интерфейсы / Под ред. проф. А. И. Михайлушкина. СПб.: Изд-во СПбГИЭУ, 2007. 172 с.
9. **Митчелл Э.** Руководство ESRI по ГИС-анализу: Пер с англ. Т. 1. Географические закономерности и взаимодействия. М.: ООО «Дата+», 2001. 190 с.