ПРИМЕНЕНИЕ МЕТОДОВ И СРЕДСТВ МОДЕЛИРОВАНИЯ

УДК 519.6 : 504.064

В.А. Артемчук, И.П. Каменева, кандидаты техн. наук
Ин-т проблем моделирования в энергетике им. Г.Е. Пухова НАН Украины
(Украина, 03164, Киев, ул. Генерала Наумова, 15,
тел. (044) 4249168, e-mail: ak24avo@gmail.com),
А.В. Яцишин, д-р техн. наук
Ин-т геохимии окружающей среды НАН Украины
(Украина, 03680, Киев, пр-т Палладина, 34а)

Модели представления и преобразования данных в задачах экологического мониторинга урбанизированных территорий

Рассмотрены многомерные модели предметной области, формализующие постановку задачи экологического мониторинга урбанизированных территорий. Исследованы различные подходы к проблеме снижения размерности пространства исходных признаков, основанные на критериях информативности многомерных данных. Обоснован критерий информационной полезности проведения наблюдений, обеспечивающий эффективную организацию сети мониторинга состояния атмосферного воздуха (на примере города Киева).

Розглянуто багатовимірні моделі предметної області, що формалізують постановку задачі екологічного моніторингу урбанізованих територій. Досліджено різні підходи до проблеми зниження розмірності простору вихідних ознак, які базуються на критеріях інформативності багатовимірних даних. Обґрунтовано критерій інформаційної корисності проведення спостережень, який забезпечує ефективну організацію мережі моніторингу стану атмосферного повітря (на прикладі міста Києва).

К л ю ч е в ы е с л о в а: модели данных, мера информативности, методы снижения размерности, информативные признаки, критерий информационной полезности.

При мониторинге состояния урбанизированных территорий одной из приоритетных задач является получение достоверной и качественной информации, необходимой для принятия эффективных решений. Проблема информационного обеспечения задач экологического мониторинга на современном этапе рассматривается в рамках концепции устойчивого (сбалансированного) развития, так как именно мониторинг состояния среды обеспечивает информационную базу для определения индексов устойчивого развития на территориальном уровне [1, 2].

К сожалению, в Украине до сих пор не решена проблема согласованности первичных данных для различных областей и населенных пунктов,

отсутствуют эффективные средства информационной поддержки задач мониторинга и оценки экологической безопасности на территориальном уровне.

Для качественного решения задач мониторинга необходимо обеспечить выполение ряда основных требований к проведению измерений, а именно:

достаточную точность измерения основных показателей;

полноту и согласованность данных по временным интервалам (регулярность измерений);

пространственную репрезентативность и охват исследуемых территорий;

согласованность и непротиворечивость данных, полученных на разных постах;

соответствие международным документам.

Рассмотрим два направления исследований, связанных с развитием и оптимизацией структуры экологического мониторинга на региональном и локальном уровнях. Первое ориентировано на повышение качества исходных данных и выбор информативных показателей, второе — это решение задачи эффективного размещения пунктов наблюдения.

В задачах мониторинга и контроля состояния природной среды таблица данных обычно составляется по численным значениям параметров, измеряемых в точках наблюдения за уровнем загрязнения по тем или иным показателям. Поэтому качество данных, полученных в результате мониторинга, существенно зависит от выбора подходящих пунктов наблюдения и основных параметров, определяющих состояние среды (уровень загрязнения) в каждом из выбранных пунктов.

Выбор наиболее подходящих параметров для наблюдения относится к актуальным задачам экологического мониторинга, так как опасность может существенно возрастать в случае одновременного воздействия нескольких факторов риска. В ряде случаев зону опасности можно выявить при исследовании соотношений между отдельными параметрами [3, 4]. Удачный выбор информативных параметров существенно упрощает оценивание и прогнозирование изменений экологического состояния исследуемых территорий.

Таблицы первичных данных наблюдений в редких случаях содержат достаточно полное описание исследуемых объектов или ситуаций. Поэтому для исследования и прогнозирования поведения объектов мониторинга желательно от множества исходных признаков (наблюдений) перейти к более существенным и полезным характеристикам исследуемых объектов. Иными словами, возникает проблема перехода к более эффективному

представлению множества исходных данных, которое может быть реализовано с учетом критериев, определяющих меру информативности этих данных [5]. Как известно, критерии информативности активно используются в задачах анализа данных и распознавания образов, в которых переход к новому набору признаков обеспечивает более качественное разбиение на классы [5, 6].

Базовые модели данных мониторинга. Таблица данных мониторинга воспринимается как непосредственное описание состояния исследуемой системы. Чтобы извлечь пользу из этого описания, необходимо указать метод представления его в более удобной для осмысления и оперирования форме, а именно метод моделирования, или создание информационной модели данных мониторинга. Модель реализует определенный способ интерпретации исследуемых данных, позволяя сопоставить объект исследования с некоторым абстрактным его описанием.

Из множества функций, выполняемых моделями, выделим две наиболее существенные. Во-первых, модель можно рассматривать как «аккумулятор знаний» об исследуемом фрагменте действительности, который позволяет прогнозировать поведение исследуемого объекта без реальных экспериментов. Во-вторых, модели играют определенную роль в образовании смыслов: подводят к выявлению новых знаний, понятий и терминов, способствуют упорядочиванию накопленной информации.

На начальных этапах создания модели рассматривается гипотеза о происхождении данных. В задаче анализа экологических и медико-экологических данных используется две известные гипотезы.

1. Модель основана на гипотезе о статистическом происхождении данных. Эта гипотеза состоит в том, что набор данных является выборкой из бесконечной генеральной совокупности объектов, распределение которой подчинено определенному вероятностному закону. Принятие этой гипотезы обеспечивает возможность применения к исследуемым данным теоретико-вероятностных подходов. Однако корректность этой гипотезы требует строгого доказательства, что не всегда выполнимо.

2. Модель основана на гипотезе о динамическом законе. Можно предположить, что данные получены как результат детерминированного воздействия определенных факторов, но с наложением флуктуаций (шумов), которые могут быть описаны статистическими законами. При этом выбор закона может быть обусловлен априорными соображениями или интуицией исследователя.

Большинство моделей прикладной статистики построено на основе гипотез о происхождении данных. При этом данные привязаны к известным модельным схемам, что не всегда корректно обосновано. Более уни-

версальный подход к анализу данных мониторинга направлен на исследование информационной модели данных без анализа внутренних механизмов, которые ее порождают (подобно принципу «черного ящика» в кибернетике). Исследователь должен описать систему так, как она проявляет себя для внешнего наблюдателя. В то же время, рассматриваемые модели данных могут быть полезны при решении достаточно широкого класса актуальных практических задач.

Информационные модели данных мониторинга состояния окружающей среды могут быть использованы для решения следующих задач:

моделирование реакции окружающей среды на внешнее воздействие;
классификация состояний среды (событий, ситуаций);
прогноз динамических изменений состояния среды;
анализ информативности используемых параметров;
оптимизация системы мониторинга;
управление поведением окружающей среды.

Для решения перечисленных задач целесообразным может быть создание визуальных образов исследуемых данных (визуальных моделей данных), с помощью которых удобно анализировать их структуру и свойства.

Многомерная постановка задачи. Совокупность внешних проявлений взаимодействия между средой и объектом будем рассматривать как экологическое состояние объекта, который может быть представлен в виде определенной последовательности количественных показателей, или параметров [3, 5]. Формально экологическое состояние представлено как многомерный объект $X = \{x_1, x_2, ..., x_m\}$, где m — число рассматриваемых параметров.

При статистической постановке задачи экологического мониторинга предполагается наблюдение некоторого множества экологических состояний (ситуаций): как различных состояний одного объекта, так и состояний различных объектов, сопоставимых между собой в количественном отношении. В качестве параметров могут быть использованы результаты натурных наблюдений или экспериментов и экспертные оценки по отдельным показателям. Обычно рассматривается три группы параметров: физические характеристики экологических объектов, данные об их химических свойствах и определенные биологические показатели, характеризующие отдельный организм или популяцию в целом. В общем случае исследуется n объектов $X_1, X_2, ..., X_n$, т.е. вся информация может быть представлена в виде матрицы размерности $m \times n$.

Многомерный подход основан на предположении, что существует возможность лаконичного объяснения природы рассматриваемой многомерной информации. Можно выявить небольшое число наиболее важных

факторов, с помощью которых могут быть достаточно точно описаны как наблюдаемые характеристики исследуемых состояний, так и характер связей между ними. Иногда такие факторы обнаруживаются среди статистических характеристик, но чаще они являются латентными, т.е. могут быть восстановлены по совокупности исходных данных.

Эта идея составляет основу известных методов многомерного статистического анализа: метода главных компонент, факторного анализа, методов построения многомерных шкал (интегральных индексов).

Следует заметить, что многомерный анализ данных (в частности, факторный анализ) является одной из базовых составляющих методологии оценивания процессов устойчивого развития, утвержденной в ряде международных документов. Такой подход используется для определения индексов устойчивого развития и сравнительного анализа этих индексов на глобальном и региональном уровнях [7].

Для более строгой постановки задачи экологического мониторинга вводится вероятностная мера в пространстве случайных событий [8]. Пусть Ω — непустое множество, B — борелевское поле (или σ-поле) подмножеств множества Ω. Это означает, что B представляет собой набор подмножеств Ω, содержащий пустое множество $\varnothing$ и замкнутый относительно дополнения и объединения его членов.

Пусть P — неотрицательная функция, определенная на B, такая, что

$$P(\Omega) = 1,\ P(\bigcup A_n) = \sum P(A_n),$$

где $A_n \in B$ и $A_n \bigcap A_m = \varnothing$ для любых $n \neq m$. Тогда P определяет вероятностную меру, а тройка (Ω, B, P) представляет вероятностное пространство с заданной мерой.

Наиболее часты случаи, когда Ω — вещественная прямая, а B — поле всех борелевских множеств, т.е. наименьшее σ-поле, содержащее все открытые множества. Для заданной меры P на (Ω, B) можно определить функцию распределения F для любых значений x:

$$F(x) = P(\{t \in R^1 : -\infty < t \le x\}).$$

Это и есть функция распределения меры P, которая указывает вероятность появления различных значений случайной величины в процессе измерения. Легко заметить, что это неубывающая функция, все значения которой находятся в интервале от нуля до единицы.

Вероятностное пространство (Ω, B, P) будем рассматривать как достаточно общее формальное описание предметной области в задаче анализа данных экологического мониторинга [8, 9]. Если для множества Ω исходных данных (наблюдений, измерений) определена вероятностная мера, отображающая это множество в некоторое множество событий B, то каж-

дый элемент множества исходных данных получает определенную интерпретацию, связанную с его соотнесением к одному из подмножеств множества B, которое рассматривается здесь как множество возможных интерпретаций для исходных наблюдений, а мера P — как способ интерпретации, определяющий разбиение исходного множества на классы или категории.

Вероятностные модели предметной области совмещают строгое вероятностное представление многомерных данных мониторинга и достаточно простую визуальную интерпретацию [9]. Взаимодействие с графическими образами способствует активизации творческого процесса, возникновению новых гипотез и принятию решений.

Методы моделирования. Под моделированием семантического пространства знаний понимают переход к формализованному описанию этих знаний. Семантическое пространство знаний может содержать как экспертные оценки, так и информацию, полученную на основе обработки результатов измерений и экспериментальных данных из разных источников. Полем знаний называют неформальное описание понятий и взаимосвязей исследуемой предметной области, которое составляется при работе с экспертом и описывается в виде графических образов, диаграмм или таблиц.

Для моделирования семантических пространств и выявления знаний удобно использовать методы многомерной статистики. В большинстве случаев используют методы факторного анализа, многомерного шкалирования или кластерного анализа. Эти методы позволяют группировать отдельные признаки описания объекта в более емкие категории — факторы. На языке поля знаний факторы являются концептами более высокого уровня абстракции. При геометрической интерпретации такого пространства значение отдельного признака отображается как точка или вектор с заданными координатами в n-мерном пространстве, координатами которого являются выделенные факторы.

Таким образом, при построении семантического пространства предполагается переход к описанию предметной области на более высоком уровне абстракции, т.е. переход от языка с большим числом признаков к более емкому языку концептуализации, который в данном случае является метаязыком по отношению к первому.

В зависимости от личного опыта и профессиональной компетентности специалиста размерность его семантического пространства и расположение в нем тех или иных понятий может существенно меняться. Эти свойства семантических пространств можно использовать для контроля процесса обучения, при тестировании экспертов или пользователей [10, 11]. При анализе индивидуального семантического пространств выяв-

ляются те вопросы, которые не были усвоены и систематизированы. Для контроля проводится сравнение семантического пространства испытуемого со структурами знаний опытных специалистов. По степени согласованности этих пространств, их размерности и конфигурации понятий можно определить уровень знаний и подготовки испытуемого.

В результате использования различных математических методов преобразования сложных структур данных в более простую форму получаем различные структуры данных. Так, кластерный анализ порождает древовидные структуры, факторный анализ и методы многомерного шкалирования — пространственное распределение множества точек.

Методы шкалирования позволяют выявлять структуры знаний косвенным путем при получении ответов на довольно простые вопросы о близости или различии между понятиями X и Y. Многие эксперименты подтверждают такую закономерность: при повышении профессионального уровня специалиста размерность его семантического пространства уменьшается. Этот вывод вполне согласуется с известным в когнитивной психологии положением о том, что процесс познания ведет к обобщению.

Алгоритм моделирования семантического пространства знаний на основе методов многомерной статистики включает следующие этапы [8, 9]:

1. Выбор подходящего метода оценки расстояния между признаками. Этот шаг включает эксперимент, в процессе которого специалисту предлагается оценить совокупность предъявляемых признаков с использованием некоторой шкалы.

2. Построение семантического пространства на основе статистического анализа полученной матрицы расстояний. При этом происходит уменьшение числа исследуемых понятий (параметров) в результате перехода к более общим координатам (факторам или индексам).

3. Поиск смысловых (содержательных) эквивалентов для выделенных структур: идентификация или интерпретация полученных факторов (координатных осей, кластеров или подмножеств).

4. Визуальное отображение отдельных проекций семантического пространства в виде наглядных образов (3D-карт или семантических шкал).

Представляя форму обобщения информации, значение также может выступать как оператор классификации, упорядочивающий объекты или события исследуемой области. Результатом такого анализа будет выявление семантических связей между значениями, которые можно представить в свернутом виде. Обычно формой фиксации семантических связей могут быть семантические поля или онтологии.

Рассмотрим более детально два направления исследований, которые связаны с использованием критериев информативности в задачах экологического мониторинга урбанизированных территорий.

Первое направление ориентировано на выбор наиболее информативных параметров для мониторинга состояния тех или иных объектов (в рассматриваемом случае — территорий). Если измеряемых показателей достаточно много, то следует обратиться к критериям информативности, обеспечивающим переход к более эффективному набору признаков (минимизирующим пространство признаков).

Второе направление ориентировано на решение задач, связанных с модернизацией и оптимизацией структуры сети мониторинга.

Рассмотрим задачу выбора множества пунктов для проведения измерений показателей (задача размещения пунктов наблюдения).

Выбор информативного набора параметров. В работе [3] на теоретическом уровне проанализирована задача перехода к новому набору признаков. В ней каждый возможный вариант такого преобразования оценивается с помощью соответствующего критерия информативности. Допустим, $z = z(x)$ — некоторая k-мерная вектор-функция исходных переменных $x_1, x_2, ..., x_m$, где $k < m$; $I_k(z(x))$ — специально заданная мера информативности k-мерной системы $z(x) = (z_1(x), z_2(x), ..., z_k(x))$.

Возможны два критерия выбора функционала $I_k(z(x))$:

1) критерий автоинформативности, обеспечивающий максимально возможное сохранение информации об исходных признаках, которая содержится в исходном массиве $\{x_i\}$;

2) критерий внешней информативности, обеспечивающий максимально возможное выделение из массива $\{x_i\}$ информации о некоторых внешних показателях.

Задача заключается в определении такого набора признаков z в классе F допустимых преобразований исходных показателей $x_1, x_2, ..., x_m$, для которого $I_k(z(x)) = \max\limits_{F}\{I_k(z(x))\}$.

Выбор класса допустимых преобразований и меры информативности определяют метод уменьшения размерности. Существует три подхода к решению этой задачи.

1. Выбор наиболее информативных признаков из исходной совокупности.

2. Формирование вторичных признаков как линейных или нелинейных комбинаций исходных признаков.

3. Многомерное шкалирование, при котором выбор нового пространства признаков определяется из условия наименьшего искажения для попарных расстояний между объектами.

В прикладной статистике наиболее популярны методы ортогонального проектирования, основанные на построении корреляционных матриц. Их можно рассматривать в рамках второго и третьего подхода.

Если в качестве допустимых преобразований рассмотреть все возможные линейные ортогональные комбинации исходных показателей, т.е.

$$z_i(x) = \sum_{i=1}^{m} a_{ji} x_i,\ j = 1, 2, \ldots, m,$$

а в качестве меры информативности выбрать критерий

$$I_k(z(x)) = \frac{Dz_1 + \ldots + Dz_k}{Dx_1 + \ldots + Dx_m},$$

где Dx_i — дисперсия случайной величины x_i, то получим метод главных компонент.

Геометрически это преобразование сводится к переходу в новую ортогональную систему координат. Если представить n объектов в виде точек m-мерного пространства, каждая ось которого соответствует одному из параметров, то вся совокупность точек для достаточно больших значений n будет иметь форму, сходную с m-мерным эллипсоидом. Главные оси этого эллипсоида образуют новую систему координат — главные компоненты. В этой системе первая ось направлена в сторону наибольшего изменения в совокупности исследуемых параметров, вторая ось координат ортогональна к первой и направлена в сторону наибольшего изменения всех оставшихся параметров, третья ось ортогональна к двум первым и так далее, пока не построим k новых осей, где $k \le m$. Согласно условиям построения новой системы координат, каждая последующая компонента вносит меньший вклад в суммарную дисперсию, чем предыдущая, т.е. $Dz_1 \ge Dz_2 \ge \ldots \ge Dz_k$.

В качестве модели факторного анализа представим исходные переменные в виде линейной комбинации факторов F :

$$X_1\ X_2 \ldots X_m \Rightarrow F_1\ F_2 \ldots F_p,\ X_j = \sum_{k=1}^{p} a_{jk} F_k + U_j,\ p < m,$$

где F_k $(k = \overline{1, p})$ — общие факторы; U_j $(j = \overline{1, m})$ — характерные факторы; a_{jk} — факторные нагрузки.

Главным объектом преобразования в факторном анализе является корреляционная матрица, составленная из коэффициентов корреляций Пирсона. К такому типу преобразований относится R-техника, согласно которой коэффициенты корреляции рассчитываются между переменными X_j, а исходная матрица X сжимается по столбцам, т.е. число переменных сокращается до p.

Один из наиболее распространенных приемов поиска факторов — метод главных компонент. Его основное отличие от факторного анализа состоит в том, что главные компоненты F_k связаны с переменными X_j линейными преобразованиями

$$X_j = \sum_{k=1}^{p} a_{jk} F_k, \quad F_k = \sum_{j=1}^{m} a_{jk} X_j.$$

Основное соотношение метода главных компонент записывается в матричном виде: $Z = AF$, где Z — матрица размерности $m \times n$ стандартизированных исходных данных; A — матрица $m \times p$ факторных нагрузок (факторное отображение); F — матрица $p \times n$ значений факторов; m — число переменных; n — число объектов исходной матрицы; p — число выделенных факторов.

Принятие решения о прекращении процедуры выделения компонент зависит, главным образом, от того, что считается малой частью дисперсии. Это решение достаточно произвольное, однако существует два критерия: критерий Кайзера и критерий «каменистой осыпи» Кеттелла, позволяющие в большинстве случаев рационально выбрать число компонент. Анализ составляющих с малыми величинами собственных значений нецелесообразен, так как они могут быть статистически недостоверными вследствие ошибок различного происхождения.

Для интерпретации главных факторов необходимо придать каждому из них определенное содержание, связанное с предметной областью. Для этого следует определить корреляции полученных факторных нагрузок с исходными переменными. Если при анализе методом главных компонент достигается объяснение максимальной части дисперсии наблюдений, то факторный анализ направлен на объяснение корреляционных связей. В этом случае в качестве критерия информативности выбирается критерий $I_k(z(x)) = 1 - \| R_x - R_z \|^2$, где R_x — корреляционная матрица показателей $x_1, x_2, ..., x_m$; R_z — корреляционная матрица показателей $z_j = \sum_{i=1}^{k} q_{ji} F_i$; $|| R ||$ — евклидова норма матрицы R.

В отличие от компонентного анализа, в котором новая система координат является единственной, факторный анализ позволяет выбрать наиболее удобную для исследования систему координат в пространстве заданной размерности. Обычно такая система координат определяется в процессе ортогонального вращения. Для упрощения интерпретации полученных результатов факторные нагрузки приближают в процессе вращения к нулевым или единичным значениям. При этом матрица нагрузок, полученная методом главных компонент, может быть использована в качестве начального приближения для факторного анализа.

Критерий информационной полезности сети наблюдений. Рассмотрим задачу построения информативной сети наблюдений на примере мониторинга состояния атмосферного воздуха на территории г. Киева [12]. В прикладных задачах экологического мониторинга урбанизированных территорий в настоящее время преобладает экспертный подход к выбору информативных параметров, т.е. выбор информативных параметров из исходного множества осуществляется с помощью экспертов на основе результатов предварительного анализа данных мониторинга отдельных территориальных систем. В частности, при оценивании состояния атмосферного воздуха в промышленных городах Украины эксперты используют индекс загрязнения атмосферы (ИЗА) [13].

Задачу оптимального размещения пунктов наблюдений сети мониторинга состояния атмосферного воздуха (МСАВ) можно сформулировать так.

На заданной территории, разбитой на квадраты фиксированного размера, необходимо разместить не более, чем заданное число постов наблюдения (ПНЗ) с заданным радиусом «представительства», учитывая следующие факторы:

приоритетность и значения следующих величин: уровня загрязнения атмосферы, показателя социально-экономической ценности участка территории и пространственного охвата территории;

ограничение на минимальное расстояние между ПНЗ;

наличие или отсутствие сети мониторинга и возможности ее переноса;

возможность одновременного проведения наблюдений за несколькими загрязняющими веществами на одном ПНЗ;

использование различных типов ПНЗ (стационарных или маршрутных).

При этом все точки квадратов равноценны по размещению ПНЗ, а центры квадратов используются в качестве расчетных точек (рис. 1, *а*). Пункты наблюдений одной сети МСАВ считаются идентичными.

В соответствии с требованиями современного государственного и международного законодательства целесообразность k проведения мониторинга для заданной территории определяется уровнем загрязнения z, что зависит от характера выбросов вредных веществ, метеопараметров, особенностей процессов распространения загрязняющих веществ и значением социально-экономической ценности территории с учетом ее заселенности e. При этом сеть должна охватывать максимально возможную территорию. Поэтому целесообразность проведения МСАВ определяется также расстоянием d от центра заданной территории до ближайшего ПНЗ:

$$k = f\,(z, e, d). \qquad (1)$$

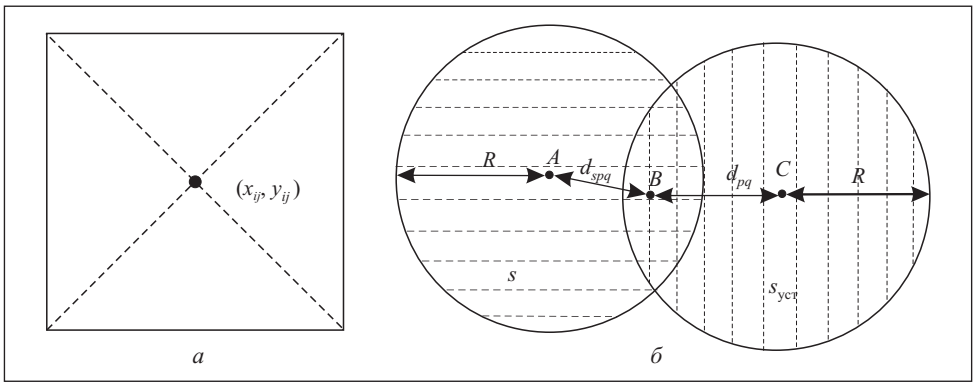

Рис. 1. Квадрат (i, j) (*а*) и схема определения расстояний для модели информационной полезности проведения наблюдений на данной территории (*б*): *A* — центр данной территории; *B* — центр квадрата (p, q); *C* — установленный ПНЗ

Исходя из размещения пунктов наблюдений сети МСАВ целесообразность проведения наблюдений на заданной территории определяем по формуле

$$k = z\left(\alpha + \beta\frac{e}{e_{\max}} + \gamma\frac{d}{d_{\max}}\right), \tag{2}$$

где α, β, γ — приоритетность соответственно учета уровня загрязнения, показателя социально-экономической ценности участка территории и расстояния до ближайшего ПНЗ; $e_{\max}$ — максимальный показатель социально-экономической ценности для исследуемой территории (например, города); $d_{\max}$ — максимальное расстояние между точкой в пределах исследуемого участка территории и ближайшим ПНЗ.

Тогда информационную полезность g проведения наблюдений на данной территории с учетом ее собственных значений k проведения МСАВ, значений взвешенной целесообразности k проведения МСАВ для прилегающих территорий и расположение уже установленных постов определяем по формуле

$$g = \sum_{(p,q)\in(s-s_{\text{уст}})} k_{pq}\frac{R - d_{s\,pq}}{R} + \sum_{(p,q)\in s_{\text{уст}}} k_{pq}\frac{R - d_{s\,pq}}{R}\ \frac{d_{pq}}{R}, \tag{3}$$

где s — площадь, представляемая постом, расположенным в центре данной территории; R — радиус представительства поста (рис. 1, *б*); $s_{\text{уст}}$ — площадь, определяемая пересечением площадей установленных ПНЗ и

предполагаемого поста; $d_{s\,pq}$ — расстояние от центра данной территории до точки (p, q).

Полученный критерий информационной полезности проведения наблюдений на данной территории (1)—(3) предлагается использовать для формальной постановки задачи размещения пунктов наблюдений сети МСАВ.

Задача оптимизации структуры сети МСАВ. Математическая постановка задачи оптимального размещения пунктов наблюдений сети МСАВ для некоторой территории (например, для города) может быть сформулирована так.

Функция цели F, обеспечивающая максимум информационной полезности сети МСАВ, имеет вид

$$F = F(A) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} b_{ij} g_{ij}(A) \to \max. \tag{4}$$

Здесь $m \times n$ — размерность матриц после разбиения исследуемой территории (города) на $m \times n$ квадратов; A — матрица управляемых переменных,

$$a_{ij} = \begin{cases} 1, \text{ если в квадрат } (i, j) \text{ установлен ПНЗ,} \\ 0 \text{ в других случаях;} \end{cases} \tag{5}$$

B — матрица описания исследуемой территории;

$$b_{ij} = \begin{cases} 1, \text{ если квадрат } (i, j) \text{ принадлежит территории города,} \\ 0 \text{ в других случаях;} \end{cases} \tag{6}$$

$g_{ij}(A)$ — информационная полезность проведения наблюдений на данной территории (i, j), которая согласно (3) определяется по формуле

$$g_{ij}(A) = \sum_{(p,q) \in (s_{ij} - s_{ij\ \text{уст}})} k_{pq}(A) \frac{R - d_{s\,pq}}{R} + \sum_{(p,q) \in s_{ij\ \text{уст}}} k_{pq}(A) \frac{R - d_{s\,pq}}{R} \frac{d_{pq}(A)}{R}, \tag{7}$$

где s_{ij} — площадь территории, представляемой постом (i, j); $d_{ij}(A)$ — расстояние между данной точкой (i, j) и ближайшим ПНЗ,

$$d_{ij}(A) = \min\left(\sqrt{(x_k - x_{ij})^2 + (y_k - y_{ij})^2} \right), \; k = \overline{1, N''}; \tag{8}$$

N'' — число зафиксированных ПНЗ на данный момент; C — вектор координат (x_k, y_k) существующей (состоящей из N' ПНЗ) и зафиксированной (состоящей из N'' ПНЗ) сети; (x_{ij}, y_{ij}) — координаты центра квадрата (i, j);

$k_{ij}(A)$ — целесообразность проведения МСАВ для территории квадрата (i, j), определяемой по формуле

$$k_{ij}(A) = z_{ij}\left(\alpha + \beta\frac{e_{ij}}{e_{\max}} + \gamma\frac{d_{ij}(A)}{d_{\max}(A)}\right), \tag{9}$$

где по умолчанию $\alpha = \beta = \gamma = 1$; E_{ij} — показатель социально-экономической ценности участка территории, учитывающий плотность населения; $e_{\max}$ — максимальный показатель социально-экономической ценности участка территории,

$$e_{\max} = \max_{b_{ij}=1}(e_{ij}),\ i = \overline{1, m},\ j = \overline{1, n}; \tag{10}$$

$d_{\max}(A)$ — максимальное расстояние между точкой в пределах исследуемой территории и ближайшим ПНЗ,

$$d_{\max}(A) = \max_{b_{ij}=1}(d_{ij}(A)),\ i = \overline{1, m},\ j = \overline{1, n}; \tag{11}$$

z_{ij} — коэффициент загрязнения ИЗА,

$$z_{ij} = \sum_{p=1}^{t}\left(\frac{q_{ijp}}{\text{ПДК}_{\text{сс}\,p}}\right)^{C_p}; \tag{12}$$

q_{ijp} — концентрация p-го загрязнения воздуха в квадрате (i, j); $\text{ПДК}_{\text{сс}\,p}$ — среднесуточная предельно допустимая концентрация p-го загрязнения воздуха; C_p — безразмерный коэффициент, приводящий уровень p-го загрязнения воздуха к уровню загрязнения вещества третьего класса опасности. Для веществ первого класса опасности C_p = 1,7, второго класса — 1,3, третьего класса — 1,0, четвертого — 0,9.

Накладываются следующие ограничения на число постов и минимальное расстояние между ними:

$$\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij} \le N, \tag{13}$$

$$d_{ij}(A) \ge d_{\min}\ \forall (i, j)\ a_{ij} = 1, \tag{14}$$

где N — число новых постов наблюдения, которые могут быть установлены; $d_{\min}$ — минимальное расстояние между ПНЗ. Кроме того, на значение $d_{ij}(A)$ влияет параметр V:

$$V = \begin{cases} 0, & \text{если в городе нет сети МСАВ или ею можно пренебречь,} \\ 1 & \text{в других случаях.} \end{cases} \tag{15}$$

Если $V = 0$, то для определения координат первого ПНЗ получаем

$$d_{ij}(A) \to \infty \Rightarrow \frac{d_{ij}(A)}{d_{\max}(A)} = 1. \qquad (16)$$

Таким образом, математическая постановка задачи (4)—(16) оптимизации размещения пунктов наблюдений сети МСАВ формализуется следующим образом:

цель исследования операций — нахождение такой матрицы A (и как следствие, набора координат для размещения не более чем N постов), при которой общая информационная полезность сети F является наибольшей;

управляемые переменные — матрица A;

входные данные — вектор C и параметр V, определяющие первоначальный вид матрицы $D(A) = \{d_{ij}(A)\}$, матрицы B, $E = \{e_{ij}\}$ и $Z = \{z_{ij}\}$, значения N и N', параметры R, α, β, γ, $d_{\min}$;

ограничения — на число постов и минимальное расстояние между ними (13), (14);

критерий оптимальности — максимальная общая информационная полезность рассматриваемой сети мониторинга.

Для изучения особенностей исследуемой задачи построены визуальные модели в виде 3D-карт, которые представляют собой коэффициенты пяти матриц, используемых при описании параметров задачи. На рис. 2—6 (см. вклейку) в горизонтальной плоскости расположена карта исследуемой территории (в данном случае г. Киева), разбитая на квадраты. По вертикальной оси отложены значения коэффициентов для тех постов, которые можно расположить в соответствующих квадратах. В качестве исходных данных использованы индексы, определяющие ценность данной территории и уровень ее загрязнения. Остальные матрицы рассчитываются в процессе решения задачи.

Специфика сформулированной задачи оптимального размещения пунктов наблюдений сети мониторинга определяется следующими условиями:

элементы матрицы $D(A) = \{d_{ij}(A)\}$ (см. рис. 2), а следовательно, и коэффициенты $g_{ij}(A)$ функции цели F (см. рис. 6), зависят от матрицы управляемых переменных A и поэтому изменяются в процессе решения задачи. Таким образом, при решении задачи неизменными остаются только матрицы E (см. рис. 3) и Z (см. рис. 4), а матрицы $D(A) = \{d_{ij}(A)\}$ (см. рис. 2), $K(A) = \{k_{ij}(A)\}$ (см. рис. 5) и $G(A) = \{g_{ij}(A)\}$ (см. рис. 6) постоянно изменяются, так как зависят от матрицы управляемых переменных A;

коэффициенты b_{ij} (6) функции цели F, описывающие исследуемую территорию, т.е. область, в которой осуществляется размещение постов, могут принимать такие значения, при которых область размещения может

оказаться невыпуклой (например, для территории Киева) или даже несвязной (если оптимизируется сеть нескольких городов-соседей одновременно), а следовательно, задача (4)—(16) не является задачей выпуклого программирования;

в соответствии с формулой (8) ограничение (14) — нелинейное, а следовательно, задача (4)—(16) является задачей нелинейного программирования;

поскольку в задаче (4)—(16) есть ограничения (13), (14), эта задача относится к задачам условной оптимизации;

множество решений задачи (4)—(16) является дискретным (целочисленным и даже бинарным), т.е. задача (4)—(16) является задачей комбинаторной оптимизации.

Следовательно, сформулированная задача (4)—(16) — детерминированная бинарная нелинейная задача условной динамической оптимизации на невыпуклой или бессвязной области. Для решения этой задачи разработаны специальные алгоритмы, основанные на различных способах соединения жадного алгоритма с методом полного перебора. Эти алгоритмы, а также описание экспериментов, связанных с их применением, приведены в работе [12].

Следует заметить, что 3D-карты могут быть использованы как весьма информативные визуальные образы исходных данных и результатов анализа. Например, 3D-карта загрязнения (см. рис. 4) представляет распределение уровней загрязнения атмосферного воздуха для территории города, построенное согласно индексу ИЗА. 3D-карта, представленная на рис. 6, отображает возможные распределения постов наблюдения в соответствии с их целесообразностью.

Выводы

Адаптированный к задачам экологического мониторинга алгоритм моделирования семантического пространства знаний обеспечивает преобразование многомерных данных в информативное множество параметров (индексы экологического состояния) с учетом выбранных для конкретной задачи критериев информативности. Предложенный и обоснованный критерий информационной полезности проведения наблюдений за состоянием окружающей среды позволил сформулировать и решить практическую задачу эффективного размещения пунктов наблюдения для сети МСАВ на примере г. Киева. Предложенные результаты рекомендуется использовать для усовершенствования уже существующих и проектирования новых сетей мониторинга состояния окружающей среды на региональном и локальном уровнях.

СПИСОК ЛИТЕРАТУРЫ

1. *Боголюбов В.М., Клименко М.О., Мокін В.Б. та ін.* Моніторинг довкілля: Підручник під ред. В.М. Боголюбова. 2-е вид. — Вінниця: ВНТУ, 2010. — 232 с.
2. *Основи* стійкого розвитку. Навчальний посібник / За ред. проф. Л.Г. Мельника. — Суми: ВТД «Університетська книга», 2005. — 654 с.
3. *Сердюцкая Л.Ф., Каменева И.П.* Системный анализ и математическое моделирование медико-экологических последствий аварии на ЧАЭС и других техногенных воздействий. — Киев: «Медэкол», 2000. — 173 с.
4. *Управление* риском: Риск. Устойчивое развитие. Синергетика. — М. : Наука, 2000. — 431 с.
5. *Айвазян С.А., Буштабер В.М., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика. Классификация и снижение размерностей. — М. : Финансы и статистика, 1989. — 607 с.
6. *Бусыгин Б.С., Мирошниченко Л.В.* Распознавание образов при геолого-геофизическом прогнозировании. — Днепропетровск: Изд-во ДГУ, 1991. — 168 с.
7. *Аналіз* сталого розвитку — глобальний і регіональний контексти : У 2-х ч. / Міжнар. рада з науки (ICSU) [та ін.]; наук. кер. М.З. Згуровський. — Київ: НТУУ «КПІ», 2010. — Ч. 1. Глобальний аналіз якості та безпеки життя людей. — 252 с.
8. *Каменєва І.П.* Просторово-семантичні моделі репрезентації знань в геоекологічних дослідженнях // Геоінформатика. — 2005. — № 4. — С. 64 — 69.
9. *Каменева И.П.* Вероятностные модели репрезентации знаний в интеллектуальных системах принятия решений // Искусственный интеллект. — 2005. — № 3. — С. 399 — 409.
10. *Гаврилова Т.А., Хорошевский В.Ф.* Базы знаний интеллектуальных систем. — СПб.: Питер, 2001. — 384 с.
11. *Петренко В.Ф.* Психосемантика сознания. — М.: Изд-во МГУ, 1988. — 230 с.
12. *Артемчук В.О.* Математичні та комп'ютерні засоби для вирішення задач розміщення пунктів спостережень мережі моніторингу стану атмосферного повітря: Автореф. дис... канд. техн. наук. Ин-т проблем моделирования в энергетике им. Г.Е. Пухова НАН Украины.— Київ, 2011. — 20 с.
13. *Щомісячний* бюлетень забруднення атмосферного повітря в Києві та містах Київської області. — Київ: Центральна геофізична обсерваторія, 2005—2012 р.

V.O. Artemchuk, I.P. Kameneva, A.V. Iatsyshyn

MODELS OF REPRESENTATION AND DATA TRANSFORMATION IN THE PROBLEMS OF ENVIRONMENTAL MONITORING IN URBANIZED AREAS

Multidimensional domain models formalizing formulation of the problem of environmental monitoring of urbanized areas have been considered. Different approaches to the problem of reducing the dimension of the space of original signs, based on the criteria of information content of multidimensional data have been studied. The criterion of usefulness of the information observation, providing more efficient organization of the air monitoring network (by the example of the city of Kiev) has been substantiated.

K e y w o r d s: data models, information content measure, dimension reduction techniques, informative signs, the criterion of information usefulness.

REFERENCES

1. Bogolyubov, V.M., Klimenko, M.O., Mokin, V.B. and et al. (2010), *Monitoring dovkillya: Pidruchnik* [Environmental monitoring: A textbook], Ed. by Bogolyubov, V.M., VNTU, Vinnytsya, Ukraine.
2. *Osnovy stiykogo rozvytku. Navchalny posibnik* [Basics of sustainable development. Tutorial] (2005), Ed. by prof. Melnik, L.G., VTD «Universitetska kniga», Sumi, Ukraine.

3. Serdyutskaya, L.F. and Kameneva, I.P. (2000), *Sistemnyi analiz i matematicheskoe modelirovanie mediko-ekologicheskikh posledstviy avarii na ChAES i drugih tehnogennykh vozdeystviy* [System analysis and mathematical modeling of medical and environmental consequences of the Chernobyl accident and other anthropogenic impacts], Medekol, Kyiv, Ukraine.
4. *Upravlenie riskom: Risk. Ustoychivoe razvitie. Sinergetika* [Risk management: Risk. Sustainable development. Synergetics], 2000, Nauka, Moscow, Russia.
5. Ayvazyan, S.A., Bushtaber, V.M., Enyukov, I.S. and Meshalkin, L.D. (1989), *Prikladnaya statistika. Klassifikatsiya i snizhenie razmernostey* [Applied statistics. Classification and reduced dimensions], Finansyi statistika, Moscow, Russia.
6. Busygin, B.S. and Miroshnichenko, L.V. (1991), *Raspoznavanie obrazov pri geologo-geofizicheskom prognozirovanii* [Pattern recognition under geological and geophysical forecasting], DGU, Dnepropetrovsk, Ukraine.
7. *Analiz stalogo rozvytku – globalny i regionalny konteksty.* Ch. 1. Globalny analiz yakosti ta bezpeky zhittya lyudey. [Analysis of sustainable development – global and regional contexts. Ch.1. Global analysis of people life quality and security], (2010), Sc. superv. Zgurovskiy, M.Z., NTUU «KPI», Kyiv, Ukraine.
8. Kameneva, I.P. (2005), "Spatial-semantic representations of knowledge in geoecological investigations", *Geoinformatika*, Vol. 4, pp.64-69.
9. Kameneva, I.P. (2005), "Probability models of knowledge representation in intellectual systems of decision making", *Iskusstvennyi intellekt: IPII NAN Ukrainy*, Vol. 3, pp. 399-409.
10. Gavrilova, T.A. and Horoshevskiy, V.F. (2001), *Bazy znaniy intellektualnykh sistem* [Knowledge bases of Intelligent Systems], Piter, Saint Petersburg, Russia.
11. Petrenko, V.F. (1988), *Psikhosemantika soznaniya* [Psychosemantics of consciousness], MGU, Moscow, Russia.
12. Artemchuk, V.O. (2011), "Mathematical and computer tools for solving the problems of placement of observation points of air monitoring network", Abstract of Cand. Sci. (Tech.) dissertation, 01.05.02, Pukhov Institute for Modeling in Energy Engineering, NAS of Ukraine, Kyiv, Ukraine.
13. *Schomisyachnyi byuleten zabrudnennya atmosfernogo povitrya v Kyevi ta mistakh Kyivskoyi oblasti* [Monthly bulletin of air pollution in the cities of Kyiv and Kyiv region], (2005-2012), Tsentralna geofizichna observatoriya, Kyiv, Ukraine.

Поступила 04.01.16

АРТЕМЧУК Владимир Александрович, канд. техн. наук, ст. науч. сотр. Ин-та проблем моделирования в энергетике им. Г.Е. Пухова НАН Украины. В 2008 г. окончил Житомирский государственный технологический университет. Область научных исследований — математическое моделирование и численные методы, информационные технологии.

КАМЕНЕВА Ирина Петровна, канд. техн. наук, ст. науч. сотр. Ин-та проблем моделирования в энергетике им. Г.Е. Пухова НАН Украины. В 1976 г. окончила Киевский госуниверситет им. Тараса Шевченко. Область научных исследований — анализ данных и математическое моделирование в экологии.

ЯЦИШИН Андрей Васильевич, д-р техн. наук, вед. науч. сотр. Ин-та геохимии окружающей среды НАН Украины. В 2002 г. окончил Киевский национальный университет им. Тараса Шевченко. Область научных исследований — математическое моделирование экологических процессов, экологический мониторинг техногенных объектов, информационные технологии.

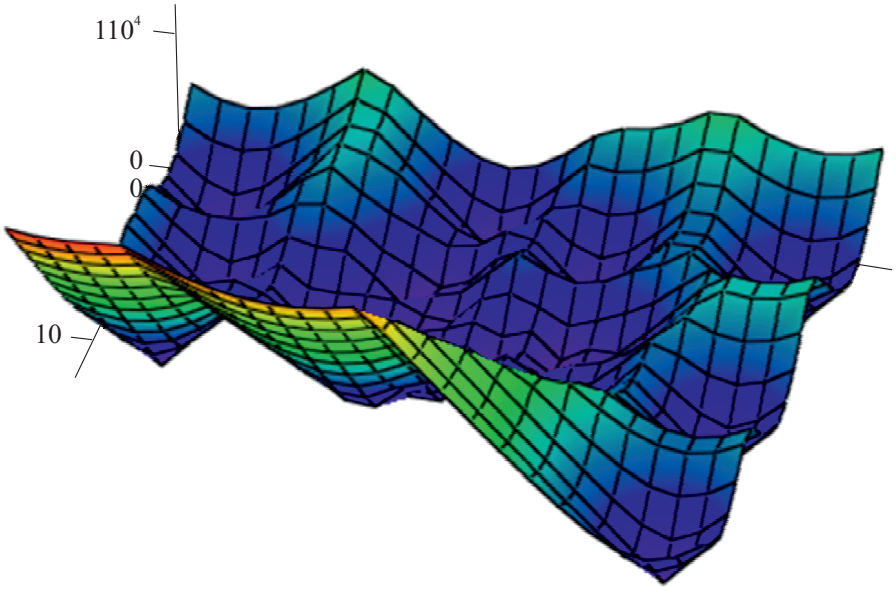

Рис 2. Поверхность, образуемая матрицей $D(A) = \{d_{ij}\}$

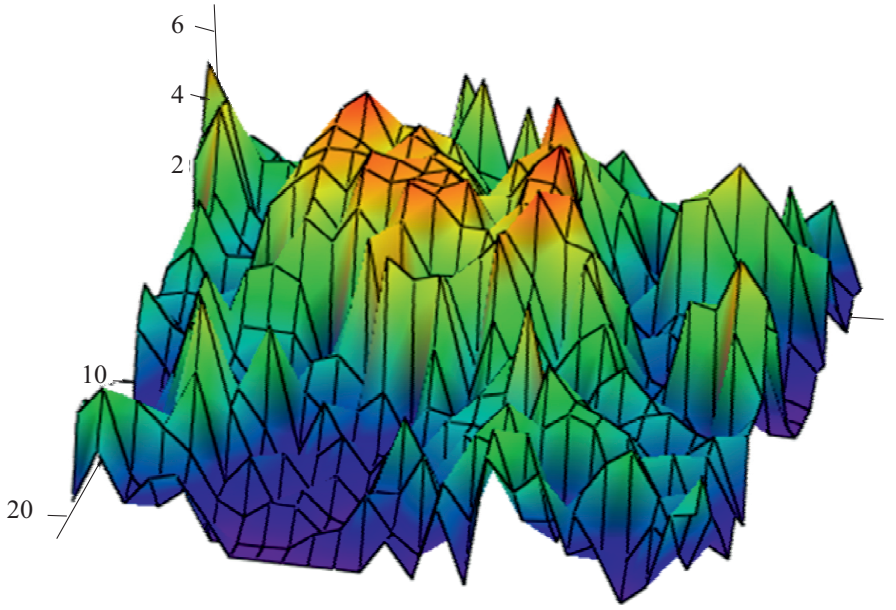

Рис. 3. Поверхность, образуемая матрицей $E = \{e_{ij}\}$

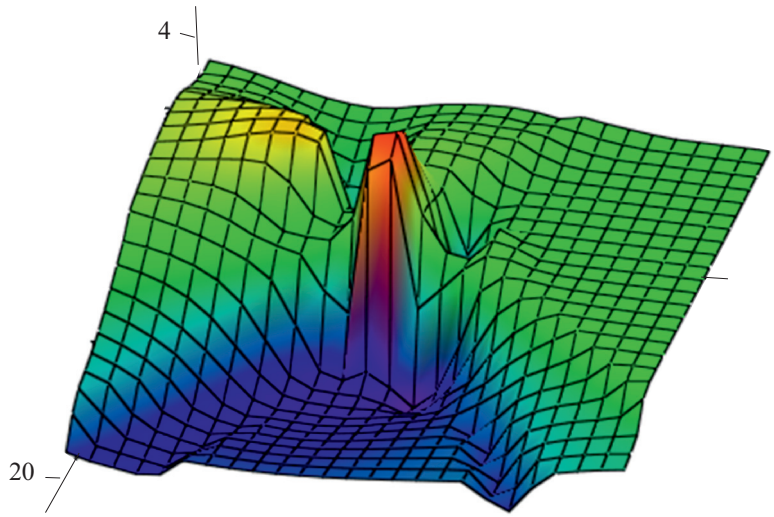

Рис. 4. Поверхность, образуемая матрицей $Z = \{z_{ij}\}$

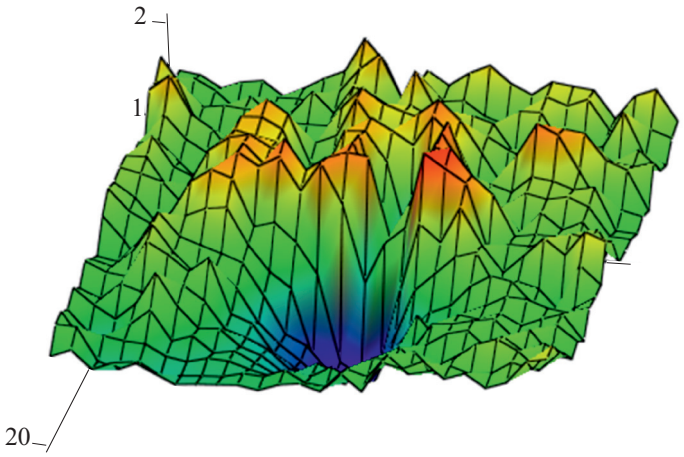

Рис. 5. Поверхность, образуемая матрицей $K\ (A) = \{k_{ij}\}$

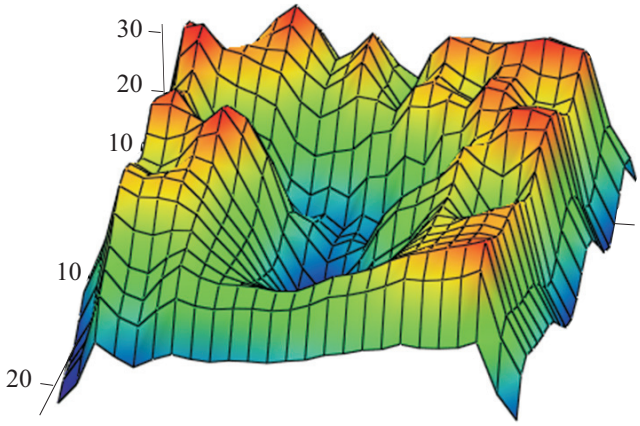

Рис. 6. Поверхность, образуемая матрицей $G\,(A) = \{g_{ij}\}$