

# ОХОРОНА НАВКОЛИШНЬОГО СЕРЕДОВИЩА (ІНЖЕНЕРНА ЕКОЛОГІЯ) І РЕСУРСОЗБЕРЕЖЕННЯ

УДК 504.064.36; 004.8.032.26

С.П. Алёшин, к.т.н.

## НЕЙРОСЕТЕВАЯ МОДЕЛЬ ТЕХНОГЕННОЙ НАГРУЗКИ РЕГИОНА ПО РЕЗУЛЬТАТАМ ЭКОЛОГИЧЕСКОГО МОНИТОРИНГА

Полтавский национальный технический университет имени Юрия Кондратюка, [aleshsp@ukr.net](mailto:aleshsp@ukr.net)

*Предложена методика разделения данных на однородные группы путем кластерного анализа и построения многомерной регрессии для каждого кластера в формате множества весовых коэффициентов обученной нейронной сети. Найден ансамбль моделей с приемлемой для практики производительностью при обоснованной надежности и точности полученных результатов. Показана реализуемость предложенных алгоритмов в среде стандартного пакета технического анализа, что позволяет избежать дополнительных материальных и временных затрат при создании систем поддержки принятия решений.*

**Ключевые слова:** Техногенная нагрузка, экологический мониторинг, искусственная нейронная сеть, кластерный анализ, многомерная регрессия, распознавание, обучение.

### Введение - постановка проблемы

Для эффективного анализа влияния техногенной нагрузки на окружающую среду в регионе необходимо знать внутреннюю структуру экологических факторов, их влияние на показатели качества и динамику их изменений. Для этого целесообразно найти аналитическую зависимость входных экологических факторов с индикаторами экологии в регионе. Принятие решений по коррекции экологии зависит от значения детерминирующей силы каждого фактора в отдельности и его вклада при совместном многофакторном воздействии всего массива факторов на целевую функцию. Если построить информационно-аналитическую модель анализа данных экологического мониторинга с оценкой этих показателей, то при прочих равных условиях, полученные знания могут существенно повлиять на доминантные индикаторы техногенной нагрузки в регионе [1,2].

Техногенное давление на окружающую среду и человека можно характеризовать рядом показателей экологического мониторинга, а индикатором этого воздействия рассматривать, например, статистику заболеваний среди жителей этого региона. На примере Черкасской области собран значительный массив данных (несколько тысяч) по годам и по месту замера, который при определенных условиях можно рассматривать как репрезентативную выборку примеров для извлечения предметных знаний [3]. Анализ имеющихся данных позволяет сделать вывод о многообразии и различной природе факторов, высокой степени неопределенности динамики индикаторов состояния экологии, сложности аналитического описания физических процессов и прогнозирования их развития. Поэтому построение аналитических моделей техногенной нагрузки в регионе для поддержки принятия решений является актуальной задачей как с научной, так и с практической точки зрения.

### Основной материал исследования и полученные научные результаты

Публикации в этой области исследований, несмотря на многофакторность, разную силу воздействия на среду и случайный характер факторов в функции техногенной нагрузки, свидетельствуют о принципиальной возможности решения подобной задачи. Наиболее перспективным отмечают направление исследований на основе интеллектуальных технологий извлечения знаний из массива данных по результатам мониторинга экологической нагрузки [3]. Основу этого направления составляет применение интеллектуальных технологий, построенных на принципах обучения и реализованных в формате пакетов технического анализа [3,4,5,6]. Анализ публикаций позволяет сделать вывод о целесообразности постановки и решения задачи моделирования процесса оценки и управления состоянием экологии в регионе с целью минимизации техногенной нагрузки. Если найти адекватную математическую модель, связывающую данные экологического мониторинга почвы, воды, воздуха с индикаторами состояния здоровья жителей региона, то можно построить оптимальную систему принятия

решений. Это особенно важно при ограниченных ресурсах и необходимости их продуктивного точечного использования с максимальной эффективностью. Однако в условиях большой размерности вектора входных данных (экологические показатели) и сложности взаимосвязей с индикаторами техногенной нагрузки (статистика заболеваемости) построить многомерную регрессию для всего массива данных с хорошей производительностью, как правило, не удается [3,5,8]. Поэтому логично предположить, что нахождению многофакторной зависимости, как решению задачи многомерной регрессии, должна предшествовать процедура структурирования исходных данных с целью формирования однородных подмножеств. Тогда значительно легче добиться заданных границ производительности при построении модели регрессионного анализа для каждого из полученных подмножеств. Решению этой задачи посвящена данная статья.

Для построения программной модели анализа структуры входных данных предлагается использовать среду стандартного пакет технического анализа Statsoft ( версия Statistika 6.1) и обеспечить адаптацию эмуляторов базовых функций к задаче регрессионного анализа техногенной нагрузки региона в рамках установленных ограничений.

### Постановка задачи

По результатам экологического мониторинга в Черкасской области [3] сформирован массив показателей (табл. 2). Необходимо построить математическую модель внутренней структуры факторов техногенной нагрузки на основе исходных данных экологического мониторинга, которая по заданному входному вектору наблюдений выдавала бы отклик с заданной точностью:

$$F: X \rightarrow Y_{opt}, X \subset \mathbb{R}^m, Y_{opt} \subset \mathbb{R}, \quad (1)$$

где  $X$  — множество выборок признаков описания экологического состояния;

$Y_{opt}$  — верифицированное выходное значение функции регрессии.

В нейросетевом формате эта задача может быть представлена в виде [9]:

$$y^k(x^m) = \sum_{j=1}^H v_j (w_{j1}x_1 + w_{j2}x_2 + \dots + w_{jm}x_m + u_j), \quad (2)$$

где  $y^k$  - вектор состояний;  $k$ -номер индикатора состояний (дыхание, кровь, % обращений к врачу и т.д.);  $(x^m)$ - вектор факторов;  $m$  - размерность входного вектора данных – (NO, SO<sub>2</sub>, CO и т.д.);  $H$  – мощность обучающей выборки;  $v_j$  - параметры нейросети;  $w_{j1}, w_{j2}, \dots, w_{jm}$  – весовые (синаптические) коэффициенты нейросетевой модели.

Нахождение однородных подмножеств во входном массиве данных рассмотрим как задачу кластерного анализа. Имеется конечный массив обучающих выборок  $X^n = \{x_1, x_2, \dots, x_n\} \subset X$ . Известен вид функции расстояния между объектами  $\rho(x, x^*)$  (евклидово расстояние). Необходимо разделить массив данных на непересекающиеся подмножества, так, чтобы каждое состояло из объектов, близких по метрике  $\rho(x, x^*)$ , а объекты разных подмножеств существенно отличались между собой, то есть:

$$\sum_i^n d(X_i, \bar{X}_j) \Rightarrow \min; \quad (3)$$

$$\sum_j^K d(\bar{X}_h, \bar{X}_j) \max, \bar{X} = (x_1, \dots, x_n) \in X; \quad \bar{K} = (k_1, \dots, k_j) \in K, \quad (4)$$

где  $X$  – множество выборок признаков описания инвестиционных состояний;  $K$  – множество номеров кластеров;  $\bar{X} = (x_1, \dots, x_m) \in X$ ;  $\bar{K} = (k_1, \dots, k_j) \in K$  - множества значений входных факторов и номеров классов;  $h, j = 1, 2, \dots, K$  – средние значения в кластерах;  $d(X_i, \bar{X}_j)$  – расстояние между объектом и центром кластера;  $d(\bar{X}_h, \bar{X}_j)$  – расстояние между центрами кластеров.

Решив задачу кластерного анализа и добившись однородности данных выборок в каждом подмножестве, зависимость «вход—выход», будет представлять собой непрерывную функцию [8]. Построим для каждого из кластеров отдельную нейронную сеть для многомерной

регрессии, получим физическую модель техногенной нагрузки как реализацию функции (1). Если производительность сети и ошибки на обучающем, контрольном и тестовом множествах в допустимых пределах, то модель становится инструментом поддержки принятия решений в оценке техногенной нагрузки региона. Прогноз динамики заболеваний и выбор оптимальных управляющих факторов для достижения целевого состояния имеют детерминированную связь через массив синаптических весов обученной нейронной сети.

Решение задачи исследования. Решение задачи основано на извлечении знаний из массива примеров путем обучения нейросетевой модели анализа данных. В качестве программного инструментария используем пакет технического анализа Statistika Neural Network. Учитывая, что число рассматриваемых факторов анализа техногенной нагрузки достигает нескольких десятков [3], эти факторы, как правило, оказывают влияние друг на друга, часть из них зашумлена или отсутствует, решения приходится принимать в условиях высокой априорной неопределенности. В этих условиях наиболее эффективны методы и алгоритмы нейросетевого распознавания образов [4,5,7]. В нашей задаче в качестве образа следует рассматривать каждую строку информативных признаков, описывающих состояние экологии в регионе. Так как исходная классификация техногенной нагрузки в формате информативных признаков (Таблица1) отсутствует, то на первом этапе задача распознавания классов сводится к распознаванию компактных, однородных группировок в данных, что позволяет корректно применить теорему Колмогорова [9,10] для регрессионного анализа в каждом кластере.

Внутри кластеров данные однородны, а, значит, поведение физического процесса (техногенной нагрузки) в рамках одного кластера более предсказуемо, чем динамика этого процесса во всем пространстве наблюдений имеющихся примеров.

Таким образом, если будут найдены группировки в данных мониторинга как значения факторов компактных в многомерном пространстве на основе некоторой метрики близости, то новые наблюдения можно классифицировать по принадлежности к тому или иному кластеру. Если на вход обученной модели подать выборку значений факторов заданной размерности, модель отнесет ее к кластеру с минимальным значением меры близости. Это эквивалентно присвоению номера класса техногенного состояния по исследуемой выборке данных.

Манипуляции размерностью вектора и информативностью его элементов с многократным повторением данной процедуры кластеризации, позволяют эмпирическим путем найти совокупность факторов соответствующих нужному классу экологической нагрузки региона.

### Модели анализа данных и его результаты

Воспользуемся обобщенными показателями экологического мониторинга на примере Черкасского региона [3], составим исходную таблицу данных в пакете технического анализа Statsoft (Statistika 6.1) и проведем кластеризацию данных в пространстве наблюдений (строк) по известной методике [8] для стандартизованных данных. Результатом стандартизации является приведение всех переменных к единой шкале: данные изменяются в пределах нуля в диапазоне  $\pm 3$ , причем большая часть всех значений будет принадлежать интервалу  $(-1, 1)$ . Процедура стандартизации не изменяет структуру взаимодействий между переменными, не влияет на структуру кластеров. Результат кластеризации представлен таблицей 1.

Таким образом, можно утверждать, что все множество данных хорошо разделяется на три кластера. Стоит отметить, что данные таблиц хорошо

Таблица 1

Кластеры в данных экологического мониторинга

	Элементы кластера номер 1 (РЕЗУЛЬТАТ ЭКОЛОГИИ) и расстояния до центра кластера. Кластер содержит 95 набл.					
	3	4	5	6	7	
Расст.	356,3160	902,6745	348,7101	201,4843	329,9767	271
	Элементы кластера номер 2 (РЕЗУЛЬТАТ ЭКОЛОГИИ) и расстояния до центра кластера. Кластер содержит 232 набл.					
	Var1	2	101	102	103	1
Расст.	710,9604	710,9604	39,85736	80,80621	343,3665	194
	Элементы кластера номер 3 (РЕЗУЛЬТАТ ЭКОЛОГИИ) и расстояния до центра кластера. Кластер содержит 15 набл.					
	26	51	76	118	119	1
Расст.	1259,087	886,3760	948,9191	435,3337	661,3876	281

иллюстрируют их устойчивую группировку внутри кластеров и надежное разделение наблюдений на однородные группы между кластерами. Алгоритмы, реализующие процедуры группировки данных, обеспечивают однородность данных внутри кластеров и максимальную отдаленность самих кластеров [7]. Следовательно, условия (3) и (4) при выполнении процедуры формирования однородных групп в исходных данных соблюдены.

На следующем этапе реализуем процедуру (1) как задачу многомерной регрессии внутри каждого кластера. Для этого наблюдения соответствующего кластера используем в задаче многомерного регрессионного анализа. В отличие от первого этапа исследования, где реализован алгоритм анализа данных с самообучением, второй этап реализует процедуру построения многомерной регрессии в режиме «обучения с учителем» [3]. При этом процедура (2) в постановке задачи выполняется итерационно в процессе обучения сети по алгоритму обратного распространения ошибки [7] применительно к данным формата экологической

нагрузки региона  $\frac{1}{mn} \sum_{j=1}^n \sum_{i=1}^m (y_{ij} - d_{ij})^2 \Rightarrow \min$ , где  $y_{ij}$  – вектор индикаторов выходных

состояний из табл. 1;  $d_{ij}$  – результат обучения сети на  $j$  - выходе, при  $i$  – м примере обучающей выборки,  $j = 1, n$  – номер выхода сети;  $i = 1, m$  – номер примера;  $m, n$  – размерность массива примеров и числа выходных элементов сети.

После выполнения процедуры обучения получим ансамбль нейросетевых моделей (Рис 2).

Анализ профилей результатов моделирования позволяет выделить сети с приемлемыми для практики производительностью и ошибками на обучающих, контрольных и тестовых множествах. Это свидетельствует об адекватности математической модели физическому содержанию исследуемого процесса, который формализован выражением (2) в постановке задачи.

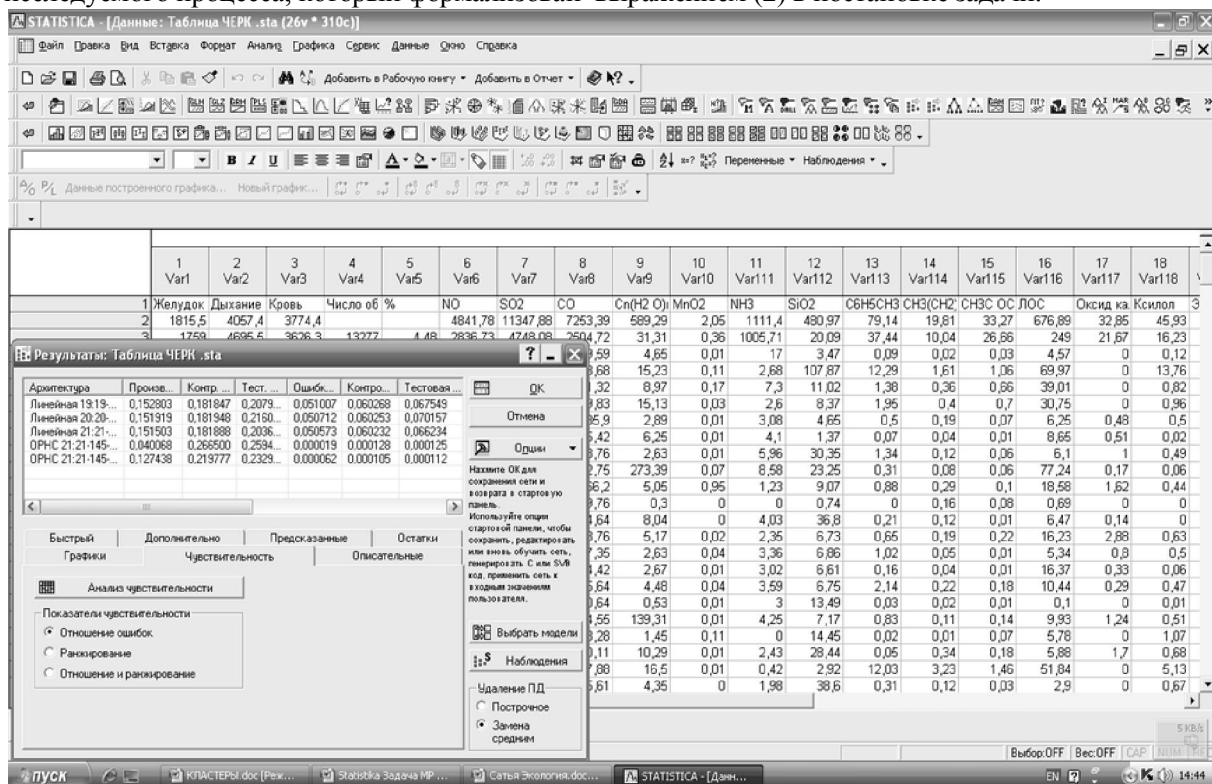


Рис 2. Профили результатов моделирования и фрагмент обучающей выборки.

На примере кластера №2 приведен графический результат построения многомерной регрессионной зависимости (2) для пяти моделей (Рис. 2).

Найденная регрессионная зависимость устанавливает связь входных факторов и выходных индикаторов системы через массивы синаптических коэффициентов пяти обученных нейросетевых моделей. Производительность моделей представлена таблицей 2.

Сеть успешно распознала структуру обучающего множества и пригодна для использования в прогнозировании значений зависимой переменной. Таким образом, математическая модель техногенной нагрузки региона построена. Вариации входными факторами позволяют получать различные значения функции, что обеспечивает системного менеджера данными для принятия решений.

### Выводы

Анализ экологической нагрузки региона как решение задачи распознавания кластеров и построение регрессии в каждом из них реализуется в среде стандартных эмуляторов пакета технического анализа StatSoft в пределах заданных показателей адекватности.

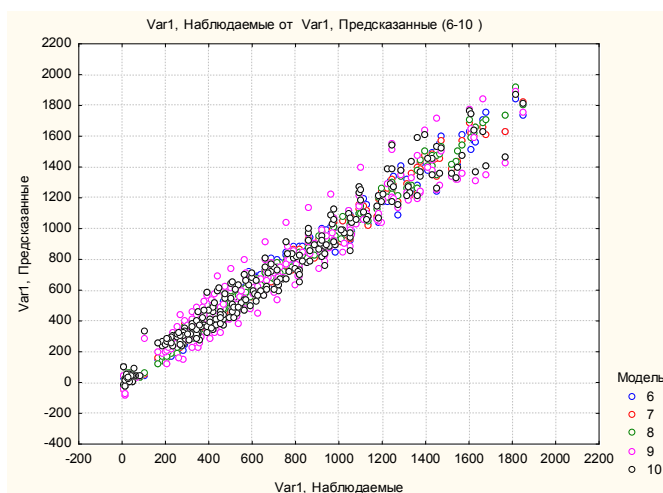


Рис 2. Характеристика качества многомерной регрессии

Таблица 2.

Показатели производительности моделей					
Модели (6-10) многомерной регрессии (ЭКОПРОЕКТ ЧЕРКАССКАЯ ОБЛ .SNN. sta)					
Показатели/модели	Var5.6	Var5.7	Var5.8	Var5.9	Var5.10
Ст. отклонение данных	2042,348	2042,348	2042,348	2042,348	2042,348
Ст. отклонение ошибки	355,446	359,548	351,992	381,685	373,371
Отношение ст. отклонений	0,174	0,176	0,172	0,187	0,183
Корреляция	0,985	0,984	0,985	0,982	0,983

Научная новизна статьи состоит в разработке методики извлечения знаний для принятия решений из массива данных экологического мониторинга региона путем использования стандартного нейроэмулятора Statistika Neural Network.

Практическая значимость результатов исследования состоит в снижении рисков и повышении объективности при принятии решений по оптимизации техногенной нагрузки в регионе.

### Список литературных источников

1. Бусленко Н.П. Моделирование сложных систем. 2 – изд. – М.: Наука, 1978. – 400с.
2. Морозов А.А., Теслер Г.С. Ситуационное управление и системы поддержки принятия решений //Збірник доповідей науково – практичної конференції. ІПММС НАН України / Системи підтримки прийняття рішень. Теорія і практика. – Київ – 2005. – С.5-9.
3. Голуб С.В. Методологія створення автоматизованих систем багаторівневого соціоекологічного моніторингу: автореф. дис. на здобуття наукового ступеня д-ра техн. наук: 05.13.06 / С.В. Голуб. – Київ, 2008. – 35 с.
4. Хайкин С. Нейронные сети: полный курс, 2-е издание. Пер. с англ. - М.: Издательский дом «Вильямс», 2006 г. - 1104с.
5. Морозов А.А. Состояние и перспективы нейросетевого моделирования СППР в сложных социотехнических системах / В.П. Клименко, А.Л. Ляхов, С.П. Алёшин // Математичні машини і системи. – 2010. - № 1.- С. 127 – 149.
6. Ляхов А.Л. Интеллектуальный анализ данных в прикладных экономических задачах / А.Л. Ляхов, С.П. Алёшин // Науковий вісник Полтавського національного технічного університету. Економіка і регіон – 2009. - № 4(23). – С. 140 – 147.
7. Барабаш Ю. Л. Коллективные статистические решения при распознавании.- М.: Радио и связь , 1983.- 224с.
8. Боровиков В.П. STATISTICA NN - Техническое описание. М.: Мир,1999. – 239 с.
9. Колмогоров А.Н. О представлении непрерывных функций нескольких переменных в виде суперпозиций непрерывных функций одного переменного и сложения / Колмогоров А.Н. // Доклады АН СССР. – 1957. – Т. 114. – С. 953 – 956.
10. Ляхов А.Л. Сложная социотехническая система как объект управления искусственной нейрон-ной сетью/А.Л. Ляхов,С.П. Алёшин//Вісник інженерної академії України.–2010.–№ 1.–С.93–97.