УДК 504.052+004.62

М.З. ЗГУРОВСКИЙ, А.А. БОЛДАК, К.В. ЕФРЕМОВ

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ И СИСТЕМНОЕ СОГЛАСОВАНИЕ НАУЧНЫХ ДАННЫХ В МЕЖДИСЦИПЛИНАРНЫХ ИССЛЕДОВАНИЯХ

Ключевые слова: *информационное пространство, согласованность данных, Мировой центр данных, интеллектуальный анализ данных, устойчивое развитие.*

ВВЕДЕНИЕ

Современное исследование сложных систем предполагает их одновременное изучение с позиций многих научных дисциплин, что является необходимым условием формирования более полной научной картины мира на основе создания междисциплинарных моделей, полученных в результате системного согласования эмпирических данных, моделей, методов, используемых в различных научных областях. Такие исследования зачастую основаны на взаимодействии многих участников на уровне средств обмена и преобразования данных, средств их обработки и анализа, включая инструменты системного согласования междисциплинарных данных, их систематизации, интеллектуальной обработки, оценки адекватности, анализа качества, корректности и т.д. Задачи долгосрочного хранения и управления научными данными, организация всеобщего и равноправного доступа к ним, содействие соблюдению стандартов данных возложены на Мировую систему данных (МСД, World Data System — WDS), работающую под эгидой Международного совета по науке (МСН, International Council for Science — ICSU) [1, 2]. В настоящей статье рассмотрен комплекс задач интеллектуального анализа и системного согласования научных данных различной природы. Представлены математические и программно-технические инструменты решения указанного класса задач. Рассмотрен пример системного согласования данных экономической, экологической и социальной природы в задаче глобального моделирования процессов устойчивого развития, ежегодно выполняемого в рамках деятельности Мирового центра данных «Геоинформатика и устойчивое развитие» (МЦД–Украина) [3].

ОРГАНИЗАЦИЯ МЕЖДИСЦИПЛИНАРНЫХ ИССЛЕДОВАНИЙ НА БАЗЕ МЦД–УКРАИНА

Исходя из понимания важности углубления сотрудничества ученых в области сбора, обмена и использования данных различной природы для научных исследований, Президиум НАН Украины в 2006 г. инициировал создание в Украине Междисциплинарного центра данных, который, пройдя необходимые этапы международной сертификации, в 2011 г. вошел в состав Мировой системы данных Международного совета по науке и получил статус Мирового центра данных по геоинформатике и устойчивому развитию в Украине (МЦД-Украина). Украинский Центр является 53-м МЦД, созданным в Мировой системе данных, которая в настоящее время охватила 13 стран. Центр работает на базе Национального технического университета Украины «Киевский политехнический институт» и Института прикладного системного анализа МОНМС Украины и НАН Украины.

На МЦД-Украина возложены задачи сбора, обработки и анализа национальных и мировых данных, необходимых для исследований в области устойчивого развития, а также оказания содействия национальным научным организациям

в сборе и предоставлении конечным пользователям наборов данных по широкому спектру дисциплин [4]. Отличительной особенностью МЦД-Украина является уникальная для Мировой системы данных сетевая модель построения Центра типа «Network of networks». Согласно этой модели каждая группа научных учреждений Национальной академии наук Украины, которая координирует деятельность одного или нескольких научных направлений, организовала сотрудничество в рамках МЦД-Украина. По каждому научному направлению собирают и предоставляют научные данные в МЦД-Украина следующие учреждения:

— Институт прикладного системного анализа НАН Украины и МОНМС Украины (системное согласование междисциплинарных данных, исследование устойчивого развития);

— Институт геофизики НАН Украины им. С.И. Субботина (исследование в области сейсмологии, гравиметрии, тепловых потоков, археo- и палеомагнетизма, магнитных измерений);

— Научный центр аэрокосмических исследований Земли Института геологических наук НАН Украины (аэрокосмическая съемка для использования в геологии, экологии, сельском, лесном и водном хозяйстве);

— Главная астрономическая обсерватория НАН Украины (исследование в области космической геодезии и геодинамики; космические лучи);

— Морской гидрофизический институт НАН Украины (сбор океанологических и гидрометеорологических данных);

— Институт географии НАН Украины (сбор картографических данных);

— Чернобыльский центр по проблемам ядерной безопасности, радиоактивных отходов и радиоэкологии (сбор данных о радиационных, биологических и медицинских последствиях Чернобыльской катастрофы, о безопасности «саркофага»).

Анализ процессов устойчивого развития стран мира и территорий (областей) Украины и оценивание влияния совокупности глобальных угроз на устойчивое развитие этих объектов является классическим примером сложной междисциплинарной задачи, требующей извлечения и обработки больших массивов геофизических и социально-экономических данных из разнотипных источников, существенно отличающихся один от другого по форматам представления и хранения данных [3].

К моменту создания Мирового центра данных в Украине (2006 г.) существовали все предпосылки для успешного создания его сетевой инфраструктуры. В качестве коммуникационной платформы внутри страны была использована Украинская научно-образовательная телекоммуникационная сеть (Ukrainian Research & Academic Network — URAN), физически объединившая вышеуказанную группу учреждений НАН Украины [5, 6]. Учитывая, что с 2007 г. сеть URAN вошла в состав общеевропейской академической сети GEANT2, появилась возможность информационно связать МЦД-Украина с 52-мя другими МЦД Мировой системы данных, а также с партнерскими научными учреждениями всего мира.

При решении задач МЦД-Украина, требующих больших вычислительных мощностей и значительных объемов памяти, используется вычислительный кластер НТУУ «КПИ» производительностью 7 ТФлопс, входящий в состав единой национальной GRID-инфраструктуры Украины, которая насчитывает в настоящее время 31 кластер.

ИНСТРУМЕНТАРИЙ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА И СИСТЕМНОГО СОГЛАСОВАНИЯ ДАННЫХ

Информационная среда МЦД-Украина представляет собой распределенную информационно-аналитическую систему, предназначенную для поддержки междисциплинарных научных исследований.

Нижний уровень системы обеспечивает стандартизацию процессов обслуживания данных. Эти процессы образуют непрерывный жизненный цикл со следующими фазами: создание, обработка, анализ данных, их сохранение, публикация и повторное использование [7–9]. В рамках требований, предъявляемых Мировой системой данных к отдельным центрам, условие обеспечения полного жизненного цикла данных является обязательным для сертификации.

Исходя из концепции объединения источников данных и сервисов — Global System of Data Systems, принятой Мировой системой данных, информационные ресурсы всей совокупности интегрируемых источников и сервисов должны быть представлены как новый единый источник [1, 2, 10]. Решение этой задачи осложняется тем, что в информационных системах центров Мировой системы данных используются различные модели данных, которые построены на основе различных (в лучшем случае клиент-серверных или сервис-ориентированных) архитектур, что часто требует разработки специализированных средств для организации взаимодействия с унаследованным программным обеспечением.

Для решения этой задачи МЦД-Украина разработал прототип распределенной информационно-аналитической системы [11], в которой интеграция данных и сервисов осуществляется на семантическом уровне с использованием языка описания онтологий OWL, обеспечивающего поддержку единого представления данных с учетом их семантических свойств в контексте единой онтологии предметной области [12]. Каталогизация источников данных в этой системе осуществляется с использованием онтологии Global Change Master Directory (GCMD) Science Keywords [13].

Для реализации такой системы в работе [12] предложен агентно-ориентированный подход, используя который в рамках сервис-ориентированной архитектуры можно осуществлять интеграцию не только данных, но и сервисов, а также организовывать их взаимодействие за счет применения агентов. Эти агенты представляют собой специализированные мигрирующие программные компоненты, которые имеют уникальные для их предметной области OWL-онтологии и файлы XML, определяющие параметры их подключения к источнику, правила проецирования словаря на источник данных и параметры функционирования [14].

Взаимодействие пользователей с разработанной системой организовано через портал МЦД-Украина: http://wdc.org.ua. На рис. 1 представлен внешний вид пользовательского WEB-интерфейса, который обеспечивает организацию виртуального пространства источников данных и сервисов: регистрация источников данных и сервисов с привязкой к единой GCMD-онтологии, мониторинг состояния источников данных и сервисов, получение справочной метаинформации о них; поиск и интеграцию данных: формирование запросов путем задания параметров фильтров или непосредственного введения SPARQL-запроса, ограничение зоны поиска по источникам данных, приведение результатов поиска к виду, определяемому пользователем (поддерживаются форматы XML, JPEG, CSV, HTML, различные виды графиков и таблиц). Система включает также сервисы для аналитической обработки данных (факторный анализ, кластерный анализ, корреляционный анализ и т.п.).

Используя разработанную систему, МЦД-Украина обеспечивает взаимодействие организаций-владельцев данных (ведущих научных учреждений НАН Украины) в рамках модели «Network of Networks». Этот принцип впервые был предложен МЦД-Украина и принят Мировой системой данных как образец для других междисциплинарных центров [4]. Взаимодействие научных учреждений, в том числе в рамках Российско-Украинского сегмента Мировой системы данных [11, 15], при поддержке НАН Украины, РАН, Российского и Украинского фондов фундаментальных исследований позволяет МЦД-Украина и его партнерам успешно проводить ряд междисциплинарных исследований в области науки о Земле, в частности в области исследования сложных эколого-социально-экономических систем в контексте их устойчивого развития (выполнено 10 совместных украинско-российских проектов).

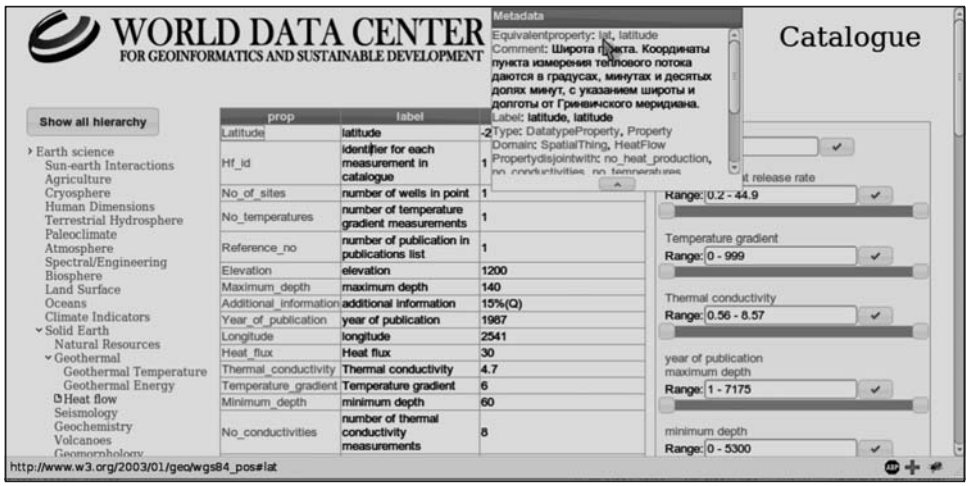

Рис. 1. Внешний вид пользовательского WEB-интерфейса для поиска и интеграции данных

СИСТЕМНОЕ СОГЛАСОВАНИЕ ДАННЫХ РАЗЛИЧНОЙ ПРИРОДЫ

Данные, используемые в междисциплинарных исследованиях, как правило, имеют различную природу, определяемую их объективным содержанием, целевым назначением и способом получения. В этом случае построение агрегированных междисциплинарных моделей [3] требует решения целого комплекса задач, связанных с приведением этих данных к единой семантике, единому диапазону значений и единым единицам измерения при условии минимизации информационных потерь, неизбежно возникающих в результате согласования. Такое согласование данных обусловливает необходимость решения нескольких типов задач. Первый тип задач связан с оценкой информационных потерь, возникающих в результате согласования данных. Второй тип задач связан с разработкой методологии количественной оценки согласованности данных различной природы, третий — с разработкой алгоритмов и методов такого согласования.

В общем случае исследование определенного явления связано с анализом и обработкой сведений о нем. Эти сведения представляют собой количественные и (или) качественные оценки свойств некоторой совокупности объектов $O = \{o_i\}$, $i = \overline{1, n}$, где o_i — значения идентификаторов номинальной шкалы объектов из представленной совокупности. Можно считать, что такие данные являются результатом отображений вида

$$O \xrightarrow{I_j} X^j, \quad j = \overline{1, m}, \tag{1}$$

где I_j, $j = \overline{1, m}$, — отображение, определенное на множестве объектов ($D(I_j) = O$), с областью значений, соответствующей области определения показателя X^j, т.е. $E(I_j) = D(X^j)$.

Семантика отображений I_j, $j = \overline{1, m}$, и отождествляемых с ними показателей X^j в соотношении (1) формулируется, исходя из целей исследования, и определяет объективное содержание (свойство оценивания), целевое назначение (предназначение этих оценок) и способ получения (данные могут быть результатом измерений, моделирования, экспертного оценивания).

Информационные потери при преобразовании измерительных шкал. Возможности совместного использования данных различной природы во многом зависят от типов шкал, в которых они измерены [16].

Цель измерительных экспериментов состоит в определении состояния эмпирической системы $E = (O, R)$ (где $R = \{r_k\}, k = \overline{1, m}$, — множество отношений между объектами множества O) с помощью поставленной ей в соответствие измерительной системы $M = (O', R')$, в которой $O' = \{o'_k\}, k = \overline{1, n'}$, — множество символов, а $R' = \{r'_l\}, l = \overline{1, m'}$, — множество допустимых отношений на O'. Соответствие между эмпирической и измерительной системой задано с помощью сюръекции $g : E \to M$.

Шкала измерений представляет собой тройку $S = (E, M, g)$, которая полностью определяет процесс измерения [17].

Оценка информационных потерь при измерении с помощью шкалы $S = (E, M, g)$ зависит от оценки неопределенности обратного отображения $g^{-1} : M \to E$. Согласно [18] информация, полученная в результате непосредственного выяснения состояния эмпирической системы, определяется как $I_E = H(E)$, где $H(E)$ — ее собственная энтропия. Но, как правило, состояние эмпирической системы можно определить лишь опосредовано с помощью шкалы $S = (E, M, g)$. В этом случае эмпирическая и измерительная системы рассматриваются как зависимые и имеет место соотношение

$$I_{M \to E} = H(E) - H(E \mid M), \tag{2}$$

где $H(E \mid M)$ — условная энтропия, характеризующая неопределенность состояния эмпирической системы в том случае, когда состояние измерительной системы полностью определено.

Таким образом, информационные потери можно оценить с помощью величины

$$\Delta I(M, E) = I_E - I_{M \to E} = H(E \mid I). \tag{3}$$

Полная информация систем E и M, определенная по формуле (2), является симметричной:

$$H(E) - H(E \mid M) = I_{M \to E} = I_{E \to M} = H(M) - H(M \mid E), \tag{4}$$

где $H(M|E)$ — условная энтропия измерительной системы.

В случае отсутствия побочного влияния на измерительную систему она считается подчиненной эмпирической системе ($H(M \mid E) = 0$), а выражение (3) с учетом соотношения (4) принимает вид

$$\Delta I(M, E) = H(E) - H(M). \tag{5}$$

Процесс преобразования данных из одной шкалы S_1 в другую S_2 можно представить как процесс измерения, в котором S_1 выступает в роли эмпирической системы, а S_2 — измерительной, т.е. можно определить новую шкалу $T_{S_1 \to S_2} = (S_1, S_2, \varphi)$, где $\varphi : S_1 \to S_2$.

Шкалы S_1 и S_2 будем считать эквивалентными, если существует $\varphi \in \Phi : S_1 \to S_2$, при котором $\Delta I(S_1, S_2) = 0$. Очевидно, что относительно классов $\Phi = \{\varphi\}$ множество шкал $T_{S_1 \to S_2} = (S_1, S_2, \varphi)$ разбивается на классы эквивалентности (типы) S_Φ и преобразование данных в пределах шкал одного типа не приводит к информационным потерям. Если определить $O = X^i \subseteq R$, $O' = X^j \subseteq R$, то шкалы можно классифицировать следующим образом [17]:

— количественные шкалы, для которых $\Phi = \{\varphi : x_i = ax_j + b\}$, $S_i, S_j \in S_\Phi$, $x_i \in X^i, x_j \in X^j$, a, b — параметры масштаба и сдвига;

— порядковые шкалы, для которых $\Phi = \{\varphi : \forall x_i \le x_j \Rightarrow \varphi(x_i) \le \varphi(x_j)\}$, $S_i, S_j \in S_\Phi$, $x_i, x_j \in X^i$, $\varphi(x_i), \varphi(x_j) \in X^j$;

— номинальные шкалы, для которых $\Phi = \{\varphi : \forall x_i \neq x_j \Rightarrow \varphi(x_i) \neq \varphi(x_j)\}$, $S_i, S_j \in S_\Phi,\ x_i, x_j \in X^i, \varphi(x_i), \varphi(x_j) \in X^j$.

Перечисленные типы шкал являются наиболее распространенными и позволяют давать как количественную, так и качественную оценку свойств исследуемых объектов.

Таким образом, решение задач согласования данных может быть сведено к построению процедуры преобразования $\varphi : S_1 \to S_2$, при этом шкалы S_1 и S_2 могут относиться как к одному, так и к разным типам. В том случае, когда S_1 и S_2 относятся к разным типам шкал, имеет место соотношение $\Delta I(S_1, S_2) = H(S_1) - H(S_2) > 0$. В этом случае считают, что шкала S_1 «сильнее», чем шкала S_2, и при переходе $S_1 \to S_2$ возникают информационные потери. При обратном переходе, когда $\Delta I(S_2, S_1) < 0$, в модели возникает неопределенность, численно равная $-\Delta I(S_2, S_1)$.

Среди известных шкал наиболее «слабыми» являются номинальные шкалы, наиболее «сильными» — количественные шкалы, порядковые шкалы в этом смысле занимают промежуточное положение.

Численная оценка информационных потерь, возникающих при переходе от количественной шкалы к порядковой (или номинальной) шкале, формулируется в терминах количества отсчетов исходной (количественной) шкалы n и количества отсчетов целевой (порядковой или номинальной) шкалы m. Это выражение для порядковой шкалы определяется как

$$\Delta I(n, m) = \sum_{l=1}^{m} \frac{R(n, l) C_m^l \log (C_{\max(n,l)}^{\min(l,n)})}{m^m}, \tag{6}$$

а для номинальной шкалы приобретает вид

$$\Delta I(n, m) = \sum_{l=1}^{m} \frac{S(n, l) A_m^l \log (A_{\max(n,l)}^{\min(l,n)})}{m^m}. \tag{7}$$

В формулах (6) и (7) C_m^l, $C_{\max(n,l)}^{\min(l,n)}$ обозначают число сочетаний, A_m^l и $A_{\max(n,l)}^{\min(l,n)}$ — число размещений, $R(n, l) = \sum_{j=0}^{l} (-1)^j C_l^j (l - j)^n$ — число Моргана, а $S(n, l) = \frac{R(n, l)}{l!}$ — число Стирлинга второго рода.

Таким образом, при преобразовании данных с целью их согласования следует учитывать возможные либо информационные потери, либо рост неопределенности модели, которые возникают при переходе от одних типов шкал к другим.

Неопределенность модели, возникающая при переходе от более «слабых» шкал к более «сильным» шкалам, может быть частично уменьшена за счет использования дополнительной информации об исследуемых объектах. Например, если при кластеризации объектов по какому-либо показателю X^j сохранить информацию о средних значениях для каждого кластера C^j, то использование этой дополнительной информации при обратном переходе от номинальных значений кластеров C^j к количественным оценкам X^j позволяет снизить неопределенность модели.

Полезность использования дополнительной информации при преобразовании данных можно оценить с помощью выражения

$$U(I, S_1, S_2) = \Delta I(S_1, S_2) - \Delta I((S_1, S_2) \mid I), \tag{8}$$

где $U(I, S_1, S_2)$ — полезность дополнительной информации I при переходе от шкалы S_1 к шкале S_2, $\Delta I(S_1, S_2)$ и $\Delta I((S_1, S_2) \mid I)$ — потери информации при преобразовании данных соответственно без использования и с использованием дополнительной информации I.

На рис. 2 представлена экспериментально полученная зависимость между длиной протокола измерений k (количество анализируемых объектов) и нормированными на $\max\limits_{k} \Delta I(S_1, S_2)$ значениями $\Delta I(S_1, S_2)$ (кривая *1*), $\Delta I((S_1, S_2) \mid I)$ (кривая *2*) и $U(I, S_1, S_2)$ (кривая *3*) для приведенного выше примера. Как видно, именно для критических длин протоколов, когда информационные потери достигают максимума, полезность дополнительной информации оказывается наибольшей и составляет в относительных единицах около 60 % максимального значения информационных потерь.

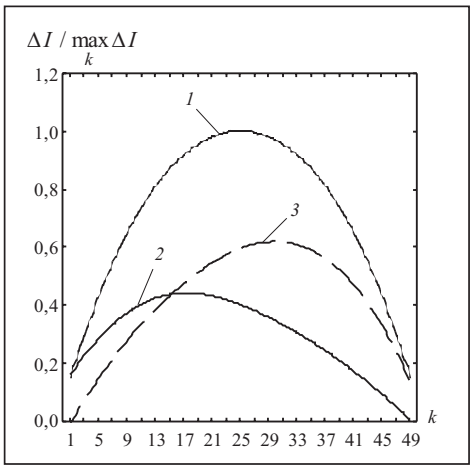

Рис. 2. Зависимость информационных потерь $\Delta I / \max\limits_{k} \Delta I$ от длины протокола измерений k для перехода от номинальной шкалы к порядковой шкале

Построение метрик для оценки согласованности данных. Все имеющиеся в распоряжении исследователя данные вида (1) могут быть представлены матрицей «объект–свойство» [19]:

$$X = (x_{i,j})_{i=1, j=1}^{n,m}, \tag{9}$$

в которой строка X_i соответствует набору значений, характеризующему свойства объекта o_i, а столбец X^j задает значения j-го показателя для всей выборки объектов.

Определенные в терминах данных различные «представления» одного и того же явления, такие как выборки объектов, временные ряды, проекции свойств и т.п., соответствуют различным подмножествам $V \subseteq X$, которые могут быть заданы как композиции на множестве элементарных выборок $V_i \equiv X_i$ — горизонтальных сечений матрицы (9), и элементарных проекций $V^j \equiv X^j$ — вертикальных сечений той же матрицы. Если обозначить $\widetilde{V}$ произвольную выборку или проекцию, а $\widetilde{X}$ — элементарную выборку или проекцию, то справедливо разложение $\widetilde{V}$ по множеству $\widetilde{X}$:

$$\widetilde{V} \subseteq \bigcup_{i=1}^{|\widetilde{X}|} \widetilde{X}_i. \tag{10}$$

Подход к построению метрических пространств для количественной оценки согласованности данных, подробно изложенный в работе [20], основан на использовании разложения (10).

В общем случае на множестве $\widetilde{X} = \{\widetilde{X}_i\}$ можно задать сигма-алгебру $\mathbb{S}$ [21] и определить метрическое пространство $(\widetilde{X}, \mathbb{S}, \mu)$, изоморфное вероятностному пространству. Иными словами, независимо от типов шкал, в которых представлены данные, происходит отображение их значений в номинальные события, определенные на множестве $\widetilde{X}$.

В случае, когда данные представлены в количественных шкалах, $\widetilde{X}_i \in \widetilde{X}$ можно рассматривать как элементы линейного пространства $E(X^1) \times \ldots \times E(X^m)$ или использовать другие модели, например кватернионы [22], на которых введена метрика.

Если удается определить пространство с нормой

$$|| \widetilde{X}_i || = \left(\sum_{j=1}^{|\widetilde{X}|} w_j (x_{i,j})^p \right)^{1/p}, \tag{11}$$

где w_j — весовые коэффициенты объектов или показателей, появляется возможность интегральной оценки представлений с помощью нормы

$$||\widetilde{V}_i|| = \left(\sum_{\widetilde{X}_k \in \widetilde{V}_i} ||\widetilde{X}_k||^p \right)^{1/p},$$

а также их близости

$$||\widetilde{V}_i|| - ||\widetilde{V}_j|| \le d(\widetilde{V}_i, \widetilde{V}_j) = ||\widetilde{V}_i - \widetilde{V}_j|| \le ||\widetilde{V}_i|| + ||\widetilde{V}_j||. \qquad (12)$$

При $p = 2$ имеем евклидово пространство [23], что дает возможность определить в нем понятие ортогональности (независимости) представлений.

Следует заметить, что определение расстояния в соотношении (12) может варьироваться в зависимости от целей исследования. Например, вместо $||\widetilde{V}_i - \widetilde{V}_j||$ часто используют $||\overline{V}_{i,j}||$, где $\overline{V}_{i,j}$ — некоторое усреднение (среднее или медиана) [17], найденное как решение задачи $\arg\min\limits_{\overline{V} \in \mathbb{S}} (d(\overline{V}_{i,j}, \widetilde{V}_i) + d(\overline{V}_{i,j}, \widetilde{V}_j))$.

Могут быть использованы и другие нормы, наиболее точно отражающие смысл того, какие представления считать близкими.

Так, для оценки «близости» статистических распределений в работе [20] предложена мера

$$L(P, Q) = 2H\left(\frac{P \oplus Q}{2}\right) - H(P) - H(Q),$$

где P и Q — оцениваемые распределения, $\frac{P \oplus Q}{2} = \left\langle \frac{p_1 + q_1}{2}, \frac{p_2 + q_2}{2}, \ldots, \frac{p_n + q_n}{2} \right\rangle$ — усредненное распределение, $H\left(\frac{P \oplus Q}{2}\right)$, $H(P)$ и $H(Q)$ — энтропия Шеннона [24].

На основании этой меры можно разработать ряд оценок, имеющих практическую ценность при решении разнообразных задач.

Предложенный в работе [20] подход к построению таких критериев состоит в том, чтобы найти распределения, выступающие в роли наилучшего $P_{\sup}$ и наихудшего $P_{\inf}$ с точки зрения решения конкретной прикладной задачи. В этом случае критерий может быть определен следующим образом:

$$S(P) = \frac{L(P, P_{\inf}) - L(P_{\inf}, P_{\inf})}{L(P_{\sup}, P_{\inf}) - L(P_{\inf}, P_{\inf})}. \qquad (13)$$

Если учесть, что $L(P_{\inf}, P_{\inf}) = 0$, то соотношение (13) примет окончательный вид

$$S(P) = \frac{L(P, P_{\inf})}{L(P_{\sup}, P_{\inf})}.$$

Очевидно, что $S(P) \in \{0, 1\}$. Значение этого критерия можно расценивать как количественную оценку распределения P, выраженную в относительной шкале.

Пусть имеется множество распределений $\pi = \{P_j,\ j = \overline{1, m}\}$, а также определены граничные распределения $P_{\sup}$ и $P_{\inf}$. Тогда на множестве π может быть определен нестрогий порядок $\alpha \subseteq \pi \times \pi : \langle p_i, p_j \rangle \in \alpha \Leftrightarrow S(p_i) \ge S(p_j)$, т.е. могут быть решены оптимизационные задачи вида $\max\limits_{P \in \pi} (S(P))$, $\arg\max\limits_{P \in \pi} (S(P))$.

Методы преобразования данных. При разработке и использовании методов преобразования данных необходимо учитывать ограничения, накладываемые типами шкал, в которых представлены данные. Так, преобразования данных в пределах одного типа шкал должны обеспечивать отсутствие информационных потерь. Если типы исходной и целевой шкал не совпадают, необходимо такие потери минимизировать.

В случае, когда данные представлены в количественных шкалах ($X^j \in R$, $j = \overline{1, m}$, где R — поле действительных чисел [21]) и есть необходимость приведения данных к единой семантике и диапазону значений, то такое преобразование данных (значений показателей) $C_j : X^j \xrightarrow{C_j} R$ должно быть без информационных потерь, что накладывает на эти преобразования условие монотонности:

$$x_{k,j} \prec x_{l,j} \Leftrightarrow C_j(x_{k,j}) \le C_j(x_{l,j});\ k, l \in [1, n],\ k \ne l.$$

Этим требованиям соответствуют линейные и нелинейные нормировки — монотонные параметрические преобразования.

Линейная нормировка определяется в соответствии с выражением

$$C_{ln}(x_{i,j}) = \frac{x_{i,j} - a}{b}, \tag{14}$$

где $x_{i,j}$ — значение из матрицы (9), a — параметр, задающий смещение; b — параметр, определяющий масштаб нормировки.

Примером нелинейной нормировки может служить преобразование

$$C_{nn}(x_{i,j}) = \left(1 - \exp\left(\frac{a - x_{i,j}}{b}\right)\right)^{-1}, \tag{15}$$

задающее логистическую кривую [25], в которой параметры a и b имеют тот же смысл, что и в формуле (14).

Параметры преобразований (14) и (15) выбирают таким образом, чтобы привести исходные данные к заданному диапазону значений (наиболее часто к интервалу [0, 1]). При этом параметры таких нормировок должны иметь те же единицы измерения, что и исходные показатели, что обеспечит безразмерность нормированных значений показателей. Эти значения следует трактовать как положение анализируемого объекта относительно некоторого множества эталонов. Способ выбора таких эталонов будет определять окончательно семантику нормированных значений.

Часто для вычисления значений параметров a и b используют выражения

$$a = \min_{i=\overline{1,n}}(x_{i,j}),\ b = \max_{i=\overline{1,n}}(x_{i,j}) - \min_{i=\overline{1,n}}(x_{i,j}) \tag{16}$$

или

$$a = \overline{X^j} = \frac{1}{n}\sum_{i=1}^{n} x_{i,j},\ b = \sigma(X^j) = \sqrt{\frac{\sum_{i=1}^{n}(x_{i,j} - \overline{X^j})^2}{n}}, \tag{17}$$

где $\overline{X^j}$ — среднее значение, $\sigma(X^j)$ — стандартное отклонение показателя X^j, определяемое на выборке объектов O. В этом случае нормировка проводится относительно среднего по выборке O.

Если в качестве эталонов использовать некоторые объекты, характеризующиеся предельными состояниями, то отображения (14) и (15) могут интерпретироваться как функции принадлежности для лингвистических переменных [20].

В случае, когда показатели заданы в номинальных шкалах, спектр эквивалентных преобразований данных ограничивается биекциями. На рис. 3 представлена схема биективного согласования данных X и Y, представленных в номинальных шкалах M и N, в рамках которой задача согласования данных рассматривается как задача оптимизации.

В качестве критерия оптимизации в указанной схеме используется вероятностная мера — альфа Клиппендорфа [26]:

$$\alpha_k = \frac{(n-1)\sum_{i=1}^{n} o_{i,i} - \sum_{i=1}^{n} s_i(s_i - 1)}{n(n-1) - \sum_{i=1}^{n} s_i(s_i - 1)}.$$

Значения α_k в диапазоне [0,75–1,0] соответствуют высокой степени, в диапазоне [0,5–0,75] — средней степени, а значения в диапазоне меньше 0,5 — низкой степени согласованности данных, представленных в номинальных шкалах.

В работе [27] рассмотрен метод преобразования в порядковую шкалу данных, представленных в количественных шкалах, при условии минимизации возникающих при этом информационных потерь. С этой целью для показателя X, выраженного в количественной шкале, необходимо построить вариационный ряд $\widetilde{X} = \langle \widetilde{x} \rangle$: $\forall \widetilde{x} \in \widetilde{X},\ \widetilde{x} \in X,\ \forall \widetilde{x}_i, \widetilde{x}_j \in \widetilde{X},\quad i < j,\ \widetilde{x}_i \le \widetilde{x}_j$, и использовать кусочно-линейную аппроксимацию его кумулянты, которая определяется следующим образом:

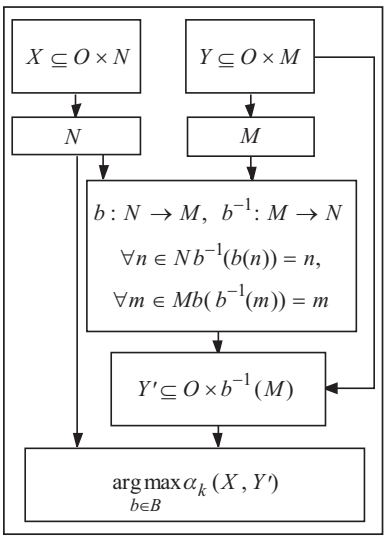

Рис. 3. Схема биективного согласования данных, представленных в номинальных шкалах

$$C_k = \langle c_i \rangle : c_i = \sum_{l=1}^{i} \widetilde{x}_l,\ i = \overline{1, n},$$

где c_i — i-е значение кумулянты, $\widetilde{x}_l \in \widetilde{X}$ — l-й член вариационного ряда, n — длина этого ряда.

Определение значений показателя X в порядковой шкале связано с построением такого разбиения вариационного ряда $\widetilde{X}$ на сегменты:

$$\pi(\widetilde{X}) = \left\langle \langle \widetilde{x}_1, \widetilde{x}_2 \rangle, \langle \widetilde{x}_2, \widetilde{x}_3 \rangle, \ldots, \langle \widetilde{x}_{r-1}, \widetilde{x}_r \rangle \right\rangle,$$

$$\widetilde{x}_l \in \widetilde{X},\ l = \overline{1, r}\ \ \forall l_1, l_2 = \overline{1, r},\ l_1 < l_2 : \widetilde{x}_{l_1} \le \widetilde{x}_{l_2},$$

для которого выполняется условие $\sum\limits_{\substack{\widetilde{x} \in \langle \widetilde{x}_i, \widetilde{x}_{i+1} \rangle, \\ i = \overline{1, r-1}}} (\widetilde{x} - M(\langle \widetilde{x}_i, \widetilde{x}_{i+1} \rangle))^2 \to \min$, где $M(\langle \widetilde{x}_i, \widetilde{x}_{i+1} \rangle)$ — математическое ожидание значений показателя X в сегменте $\langle \widetilde{x}_i, \widetilde{x}_{i+1} \rangle$ вариационного ряда. Таким образом, каждому сегменту $\langle \widetilde{x}_i, \widetilde{x}_{i+1} \rangle \in \pi(\widetilde{X})$ ставится в соответствие величина $M(\langle \widetilde{x}_i, \widetilde{x}_{i+1} \rangle)$, а каждой из них, в свою очередь, — значение ординальной шкалы:

$$X \to \widetilde{X} \to \pi(\widetilde{X}) \to \left\langle M(\langle \widetilde{x}_i, \widetilde{x}_{i+1} \rangle) \right\rangle,\ i = \overline{1, r-1} \to \overline{1, r-2}.$$

Если ряд $\left\langle M(\langle \widetilde{x}_i, \widetilde{x}_{i+1} \rangle) \right\rangle$ использовать для обратного преобразования, то неопределенность модели будет описываться зависимостью, изображенной на рис. 2.

ПРИМЕР СИСТЕМНОГО СОГЛАСОВАНИЯ ДАННЫХ В ЗАДАЧАХ МОДЕЛИРОВАНИЯ ПРОЦЕССОВ УСТОЙЧИВОГО РАЗВИТИЯ

Характерным примером использования описанных выше принципов системного согласования данных различной природы является задача моделирования влияния совокупности угроз на процессы устойчивого развития определенной территории [27].

Согласно [3] для определения компоненты безопасности жизни людей воспользуемся моделью пространства угроз, когда каждой территории j поставлен в соответствие вектор

$$\vec{Tr}_j = (t_i^j),\ i = \overline{1, n}, \qquad (18)$$

с координатами t_i^j, которые характеризуют степень проявления соответствующих угроз.

Суммарное влияние совокупности угроз на отдельные территории (страны или регионы страны) будем оценивать с помощью компоненты безопасности жизни людей C_{sl}, определяемой как норма Минковского для вектора угроз:

$$C_{sl} = ||\vec{T}r_j|| = \left(\sum_{i=1}^{n} (t_i^j)^p \right)^{1/p}. \qquad (19)$$

Поскольку с увеличением параметра p увеличивается отклик (чувствительность) модели на изменение каждой составляющей вектора $\vec{T}r_j$, и наоборот, его снижение сглаживает (огрубляет) эту чувствительность, в модели принято считать $p = 3$.

Рассмотрим представленную задачу на примере расчета влияния совокупности угроз на территориально-административные единицы (области и территории) Украины.

В табл. 1 рассмотрены виды угроз, которые были определены экспертным путем с учетом специфики устойчивого развития регионов Украины. Показатели угроз отличаются содержанием, процедурой и единицами измерений, т.е. имеют различную природу, поэтому построение агрегированных (интегральных) оценок безопасности было выполнено системным согласованием значений этих показателей.

Таблица 1

Вид угрозы	Показатель угрозы	Единицы измерения показателя
Снижение продолжительности жизни (S1)	Средняя ожидаемая продолжительность жизни	Годы
Преступность (S2)	Коэффициент преступности	Количество зарегистрированных преступлений на 100 тыс. человек
Коррупция (S3)	Индекс восприятия коррупции	%
Социальное неравенство (S4)	Индекс Джини (коэффициент неравенства в получении доходов)	усл. ед.
Рост безработицы (EC1)	Уровень зарегистрированной безработицы	на конец года, %
Износ технологической инфраструктуры (EC2)	Степень износа основных фондов	%
Снижение благосостояния населения (EC3)	Доход населения в расчете на одного человека	грн
Загрязнение городского воздуха (E1)	Индекс загрязнения атмосферы	усл. ед.
Ухудшение качества питьевой воды (E2)	Пробы воды, которые не отвечают требованиям Госстандарта	%
Загрязнение окружающей среды (E3)	Плотность выбросов загрязняющих веществ в атмосферу и воду	т / км^2
Рост выбросов парниковых газов (E4)	Плотность выбросов парниковых газов	т CO_2 / км^2

С использованием метода экспоненциального нормирования значения всех показателей угроз приведены к безразмерным величинам, которые изменяются в диапазоне [0–1]. При этом значение 0,5 отвечает среднему по стране влиянию угрозы, а значение, близкое к 1,0, — наибольшему влиянию. Определив критический порог (в данном случае 0,7), можно выделить критические значения показателей угроз для каждого региона.

На основе применения соотношений (18), (19) была рассчитана компонента безопасности жизни людей C_{sl} (табл. 2), которая агрегирует все 11 угроз, представленных в табл. 1.

Таблица 2

Регион	Рейтинг	C_{sl}	Угрозы социального характера				Угрозы экономического характера			Угрозы экологического характера			
			S1	S2	S3	S4	EC1	EC2	EC3	E1	E2	E3	E4
Высокий уровень ($C_{sl} > 1{,}1$)													
Ивано-Франковская	1	1,31	72,52	419,00	40,70	19,91	2,00	47,60	14720,30	3,40	99,48	19,87	565,96
Черновицкая	2	1,29	72,64	737,00	38,30	21,18	1,90	37,70	**13181,50**	4,80	98,20	7,39	95,34
Львовская	3	1,22	72,57	680,00	43,60	23,48	1,70	70,10	16561,30	5,60	97,75	20,35	182,56
Тернопольская	4	1,20	72,82	495,00	28,80	24,99	2,60	46,80	**13572,90**	3,90	92,35	5,78	103,96
Киевская	5	1,13	**69,45**	964,00	**46,70**	22,31	1,60	38,70	19327,80	3,26	90,20	9,57	383,23
Черниговская	6	1,12	**69,21**	867,00	26,80	24,28	**2,90**	56,30	16427,50	3,10	96,10	3,70	77,01
Хмельницкая	7	1,11	71,35	818,00	40,60	26,08	2,60	64,10	15480,30	5,20	94,11	4,35	145,81
Уровень выше среднего ($1{,}1 > C_{sl} > 0{,}99$)													
Харьковская	8	1,09	71,20	1025,00	38,80	21,39	1,90	**88,70**	18402,30	3,60	90,75	18,31	380,90
Черкасская	9	1,05	70,91	753,00	32,30	26,40	**3,30**	66,90	15445,20	5,90	95,30	8,98	201,35
Волынская	10	1,05	70,55	805,00	**53,80**	22,30	2,20	49,10	**13913,60**	8,60	95,05	3,88	65,60
г. Севастополь	11	1,03	70,65	**1452,00**	**52,00**	20,71	0,60	48,30	16763,00	3,80	99,65	**91,38**	621,58
Винницкая	12	1,02	71,49	781,00	25,50	24,89	**3,00**	**97,10**	15857,00	4,50	96,50	7,44	226,05
Херсонская	13	1,01	**69,28**	1129,00	34,60	26,31	1,70	67,30	14586,50	6,30	91,50	4,03	45,80
Полтавская	14	1,01	70,38	1095,00	25,20	23,52	**3,80**	**73,50**	17958,80	4,43	91,57	7,66	131,51
Сумская	15	1,00	70,36	915,00	**55,20**	24,53	**2,90**	63,80	16619,20	5,40	95,40	4,88	96,06
Уровень ниже среднего ($0{,}99 > C_{sl} > 0{,}87$)													
Закарпатская	16	0,98	70,23	544,00	29,80	20,99	1,80	**74,30**	**12226,90**	**14,40**	92,45	8,58	89,36
Ровненская	17	0,98	70,76	649,00	33,40	23,31	**2,90**	50,90	14352,10	**14,20**	89,35	3,70	79,74
Запорожская	18	0,95	70,53	**1543,00**	26,80	23,54	2,30	72,60	19856,60	**12,90**	94,25	14,08	549,28
АР Крым	19	0,94	70,45	**1705,00**	**52,00**	24,27	1,60	69,40	15232,10	5,24	96,85	11,51	103,16
Житомирская	20	0,94	**69,24**	797,00	24,60	**31,63**	**3,00**	57,40	15571,60	4,20	93,52	3,57	53,53
Николаевская	21	0,92	**68,71**	1067,00	31,00	24,04	**2,70**	**74,30**	16600,60	9,20	**88,49**	4,67	105,35
Низкий уровень ($0{,}87 > C_{sl}$)													
Луганская	22	0,86	69,58	**1396,00**	35,70	25,06	1,40	55,90	17836,10	**10,13**	**76,57**	41,87	443,26
Одесская	23	0,82	**68,95**	1039,00	**46,30**	**30,09**	1,40	52,70	15996,10	**13,56**	90,30	12,84	165,12
Днепропетровская	24	0,78	**69,16**	**1482,00**	35,10	**26,69**	1,60	**78,70**	20687,40	**11,42**	**83,00**	48,33	683,99
г. Киев	25	0,73	73,66	**1308,00**	40,20	**30,46**	0,30	53,30	37573,20	6,80	99,69	**449,28**	**11631,22**
Кировоградская	26	0,72	**69,03**	1174,00	**60,30**	**27,07**	**3,20**	**96,70**	15214,50	4,88	93,00	4,02	65,10
Донецкая	27	0,68	**69,07**	**1406,00**	**51,00**	25,45	1,20	64,50	21258,20	**13,49**	**78,20**	**112,73**	**2318,49**
Наименьшее значение		0,68	68,71	419,00	24,60	19,91	0,30	37,70	12226,90	3,10	76,57	3,57	45,80
Среднее значение		1,00	70,55	1001,67	38,86	24,63	2,15	63,58	17082,32	7,12	92,58	34,55	726,31
Наибольшее значение		1,31	73,66	1705,00	60,30	31,63	3,80	97,10	37573,20	14,40	99,69	449,28	11631,22

По значениям компоненты безопасности жизни людей была выполнена кластеризация, позволившая подразделить регионы Украины на четыре кластера: с высоким, выше среднего, ниже среднего и низким уровнями безопасности.

Первый кластер с высоким уровнем безопасности жизни людей ($C_{sl} > 1.1$) включает семь регионов: Ивано-Франковская, Черновицкая, Львовская, Тернопольская, Киевская, Черниговская, Хмельницкая области. Эти области характеризуются умеренным влиянием угроз социального, экономического и экологического характера. Отметим, что для Киевской и Черниговской областей существенной является угроза «Снижение продолжительности жизни». Кроме того, для Киевской области важной угрозой является «Коррупция», а для Черниговской — «Рост безработицы». Угроза «Снижение благосостояния населения» существенной является для Черновицкой и Тернопольской областей.

Ко второму кластеру с уровнем безопасности жизни людей выше среднего ($1.1 > C_{sl} > 0.99$) отнесено восемь областей Украины: Харьковская, Черкасская, Волынская, Винницкая, Херсонская, Полтавская, Сумская области и г. Севастополь. При этом высокий уровень угроз: «Преступность», «Коррупция» и «Загрязнение окружающей среды» характерны для г. Севастополя, для Волынской области существенными угрозами являются «Коррупция» и «Снижение благосостояния населения», для Винницкой и Полтавской областей — «Рост безработицы» и «Износ технологической инфраструктуры», для Сумской области — «Коррупция» и «Рост безработицы».

В третий кластер с уровнем безопасности жизни людей ниже среднего ($0.99 > C_{sl} > 0.87$) вошло шесть регионов Украины: Закарпатская, Ровенская, Запорожская, Житомирская, Николаевская области и АР Крым. Для Закарпатской и Житомирской областей существенными являются три угрозы, для Николаевской — четыре.

Четвертый кластер с низким уровнем безопасности жизни людей ($0.87 > C_{sl}$) включает шесть регионов Украины: Луганскую, Одесскую, Днепропетровскую, Кировоградскую, Донецкую области и г. Киев. На снижение безопасности жизни в регионах этой группы одновременно влияют от трех (Луганская область) до семи угроз (Донецкая область).

ЗАКЛЮЧЕНИЕ

Рассмотренная методология исследования сложных систем, функционирующих на основе использования междисциплинарных моделей, получена в результате системного согласования эмпирических данных, моделей, методов из различных научных областей. Представлены математические и программно-технические инструменты для интеллектуальной обработки, анализа и системного согласования данных различной природы, их систематизации, оценки адекватности, анализа качества, корректности и пр. На примере моделирования процессов устойчивого развития административных регионов Украины (областей, АР Крым и городов Киев и Севастополь), ежегодно выполняемого в рамках деятельности Мирового центра данных «Геоинформатика и устойчивое развитие» (МЦД-Украина), рассмотрен комплекс задач интеллектуального анализа и системного согласования научных данных экономического, экологического и социального характера.

СПИСОК ЛИТЕРАТУРЫ

1. Constitution of the International Council for Science World Data System (ICSU WDS), 2012. — http://icsu-wds.org/images/files/WDS sub Constitution sub 04 sub 04 sub 12.pdf.
2. Minster J.B. The ICSU world data system as a global system of data systems // Abstract Proc. the XXVth IUGG General Assembly “Earth on the Edge — Science for a Sustainable Planet”, 2011, Melbourne (Australia), 2011. — P. 46.
3. Згуровський М.З., Болдак А.О., Єфремов К.В. та ін. Аналіз сталого розвитку — глобальний і регіональний контексти: у 2 ч. — К.: НТУУ «КПІ», 2010. — Ч. 1: Глобальний аналіз якості та безпеки життя людей. — 252 с.
4. Згуровский М.З., Гвишиани А.Д., Ефремов К.В., Пасичный А.М. Интеграция украинской науки в Мировую систему данных // Кибернетика и системный анализ. — 2010. — № 2. — С. 49–58.
5. Згуровский М.З., Патон Б.Е., Якименко Ю.И. Состояние и перспективы развития национальной телекоммуникационной академической сети, 1977. — http://www.uazone.org/inet/uren/uran-dop1w97.html.

6. Yakymenko Yu., Timofeyev V., Galagan V., Dombrougov M. Development and European Integration of Ukrainian Research and Academic Network (URAN) for Provision of High Speed Services to Science and Education // Materials of the 21st Int. CODATA Conf., 2008, Kiev (Ukraine), 2008. — P. 253.
7. Create & manage data. URL. — http://www.data-archive.ac.uk/create-manage/life-cycle/.
8. Shearer C. The CRISP-DM model: the new blueprint for data mining // J. Data Warehousing. — 2000. — N 5. — P. 13–22.
9. Rohanizadeh S.S., Moghadam M.B. A proposed data mining methodology and its application to industrial procedures // J. Industr. Eng. — 2009. — N 4. — P. 37–50.
10. ICSU Annual Report 2008. — http://www.icsu.org/publications/annual-reports/annual-report-2008/annual-report-2008-file.
11. Yefremov K. Integration of heterogeneous data sources of Russian-Ukrainian WDS Segment based on ontology and agent-oriented approach // JpGU Intern. Symposium 2012, Makuhari Messe, Chiba (Japan), 2012. — P. 23.
12. Yefremov K. Agent-oriented approach for integration of WDC-Ukraine partner network resources // Materials of the 22nd Int. CODATA Conf., 2010, Cape Town (South Africa), 2010. — P. 32.
13. Olsen L.M., Major G., Shein K. et al. GCMD's Science keywords and associated directory keywords, 2007. — P. 26.
14. Шаповалова С.И., Ефремов К.В., Глуханик А.И. Организация интегрированного доступа к информационным ресурсам // Сб. тр. XI Междунар. конф. «Интеллектуальный анализ информации». — К.: Просвіта, 2011. — С. 101–108.
15. Shaimardanov M., Gvishiani A., Zgurovsky M. et al. Development of WDS Russian-Ukrainian segment // Proc. 1st ICSU-WDS Conf. "Global Data for Global Science", 2011, Kyoto (Japan), 2011. — P. 19–28.
16. Айвазян С.А., Бухштабер В.М., Енюков И.С. и др. Прикладная статистика. Классификация и снижение размерности. — М.: Финансы и статистика, 1989. — 607 с.
17 Luce R.D., Krantz D.H., Suppes P. et al. Foundation of measurement. — San Diego: Academic Press, 1990. — Vol. 3: Representation, axiomatization and invariance. — 368 p.
18. Вентцель Е.С. Теория вероятностей. — М.: Высш. шк., 1999. — 576 с.
19. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Исследование зависимостей: Справ, изд. / Под ред. С. А. Айвазяна. — М.: Финансы и статистика, 1985. — 487 с.
20. Згуровский М.З., Болдак А.А. Системное согласование данных разной природы в мультидисциплинарных исследованиях // Кибернетика и системный анализ. — 2011. — **46**, № 5. — С. 152–163.
21. Колмогоров А.Н., Фомин С.В. Элементы теории функций и функционального анализа. — 7-е изд. — М.: Физматлит, 2004. — 572 с.
22. Conway J.H., Smith D. On quaternions and octonions. — Abington, Oxfordshire: AKPeters/CRC Press, 2003. — 159 p.
23. Ильин В.А., Позняк Э.Г. Линейная алгебра: учебник для вузов. — 6-е изд. — М.: Физматлит, 2007. — 278 с.
24. Шеннон К. Работы по теории информации и кибернетике. — М.: Изд-во иностр. лит., 2002. — 836 с.
25. Balakrishnan N. Handbook of the logistic distribution. — New York: Marcel Dekker, 1992. — 601 p.
26. Klippendorff K. Content analysis: An introduction to its methodology / Thousand Oaks, CA: Sage, 2004. — P. 219–250.
27. Згуровський М.З., Болдак А.О., Єфремов К.В. та ін. Аналіз сталого розвитку: глобальний і регіональний контексти. 2011–2012. — К.: НТУУ «КПІ», 2012. — Ч. 2: Україна в індикаторах сталого розвитку. — 240 с.

Поступила 23.01.2013