Cybernetics and Systems Analysis, Vol. 49, No. 4, July, 2013

INTELLIGENT ANALYSIS AND THE SYSTEMIC ADJUSTMENT OF SCIENTIFIC DATA IN INTERDISCIPLINARY RESEARCH

M. Z. Zgurovsky[ab†], **A. A. Boldak,**[a‡] **and K. V. Yefremov**[c]

UDC 504.052+004.62

Abstract. *A set of problems in intelligent analysis and systemic adjustment of various scientific data are considered. The mathematical and software and hardware tools are presented to solve this class of problems. An example of the systemic adjustment of economic, ecological, and social data in the annual global modeling of sustainable development processes, which is carried out by the World Data Center for Geoinformatics and Sustainable Development (WDC-Ukraine), is considered.*

Keywords: *information space, data consistency, World Data Center, intelligent data analysis, sustainable development.*

INTRODUCTION

The modern study of complex systems implies their simultaneous analysis in the context of many scientific disciplines, which is necessary for a complete scientific worldview based on interdisciplinary models obtained as a result of the systemic adjustment of empirical data, models, and methods used in various scientific domains. Such research is often based on the interaction of many participants at the level of the tools of data exchange and conversion, their processing and analysis, including tools of the systemic adjustment of interdisciplinary data, their classification, intelligent processing, adequacy assessment, quality and correctness analysis, etc. The tasks of long-term storage and control of scientific data, universal and equivalent access to them, observation of standards of data are imposed on the World Data System (WDS), operating under the aegis of the International Council for Science (ICSU) [1, 2]. In the present paper, we will consider a complex of problems of the intelligent analysis and systemic adjustment of various scientific data and present mathematical and software–hardware tools to solve this class of problems. We will also consider an example of the systemic adjustment of economic, ecological, and social data in the global modeling of processes of sustainable development carried out annually within the framework of the activity of the World Data Center "Geoinformatics and Sustainable Development" (WDC-Ukraine) [3].

ORGANIZATION OF INTERDISCIPLINARY RESEARCH BASED ON WDC-UKRAINE

Understanding the importance of a close cooperation of scientists in the acquisition, exchange, and use of various data for scientific research, Presidium of the National Academy of Sciences (NAS) of Ukraine initiated the creation of the Interdisciplinary Data Center in Ukraine in 2006. It passed all the necessary stages of the international certification and became a part of the World Data System of the International Council for Science and obtained the status of the World Data Center on Geoinformatics and Sustainable Development in Ukraine (WDC-Ukraine) in 2011. The Ukrainian Center is the

[a]National Technical University of Ukraine "Kyiv Polytechnic Institute," Kyiv, Ukraine, [†]*zgurovsm@hotmail.com*, [‡]*boldak@wdc.org.ua*. [b]Institute for Applied Systems Analysis, National Academy of Sciences of Ukraine and Ministry of Education and Science, Youth and Sports of Ukraine, Kyiv, Ukraine. [c]World Data Center "Geoinformatics and Sustainable Development," Kyiv, Ukraine, *k.yefremov@wdc.org.ua*. Translated from Kibernetika i Sistemnyi Analiz, No. 4, July–August, 2013, pp. 62–75. Original article submitted January 23, 2013.

53rd WDC created in the World Data System, which envelopes 13 countries. The center operates on the basis of the National Technical University of Ukraine "Kyiv Polytechnic Institute" (NTUU "KPI") and of the Institute for Applied Systems Analysis of the Ministry of Education and Science, Youth and Sports (MESYS) of Ukraine and NAS of Ukraine.

The WDC-Ukraine is intended to acquire, process, and analyze national and world data necessary for studies in the field of sustainable development, as well as to assist national scientific organizations in gathering and providing data sets on a wide range of disciplines to end users [4]. A distinctive feature of the WDC-Ukraine is the "network of networks" structure model of the Center, which is unique for the World Data System. According to this model, each group of scientific institutions of the National Academy of Sciences of Ukraine, which coordinates the activity of one or several research fields, has organized cooperation within the framework of the WDC-Ukraine. In each research field, the following institutions acquire scientific data and submit them to the WDC-Ukraine:

— Institute for Applied Systems Analysis of the NAS of Ukraine and MESYS of Ukraine (systemic adjustment of interdisciplinary data, analysis of sustainable development);

— S. I. Subbotin Institute of Geophysics of the NAS of Ukraine (studies in seismology, gravimetry, heat flows, archeo- and paleomagnetism, magnetic measurements);

— Scientific Center for Aerospace Research of the Earth of the Institute of Geological Sciences of the NAS of Ukraine (GPS survey to be used in geology, ecology, agriculture, forestry, and water industry);

— Main Astronomical Observatory of the NAS of Ukraine (research in space geodesy and geodynamics; cosmic rays);

— Sea Hydrophysical Institute of the NAS of Ukraine (acquisition of oceanologic and hydrometeorological data);

— Institute of Geography of the NAS of Ukraine (acquisition of cartographic data);

— Chornobyl Center for Nuclear Safety, Radioactive Waste and Radioecology (data acquisition in radioactive, biological, and medical consequences of the Chornobyl disaster and safety of the Shelter).

An analysis of the processes of sustainable development of countries of the world and territories (regions) of Ukraine and assessment of the influence of the set of global threats on the sustainable development of these objects is a classical example of a complex interdisciplinary problem, which requires extracting and processing large arrays of geophysical and socio-economic data from various sources substantially different in data representation and storage formats [3].

By 2006, when the World Data Center was established in Ukraine, there had been all premises for the successful creation of its network infrastructure. As the communication platform inside the country, the Ukrainian Research and Academic Network (URAN) was used, which physically united the above-mentioned group of the institutions of the NAS of Ukraine [5, 6]. Since 2007, the URAN has become a part of the all-European academic network GEANT2, and it became possible to informationally connect the WDC-Ukraine with 52 other WDCs of the World Data System and also with partner scientific institutions of the world.

In solving problems of the WDC-Ukraine that require large computational capabilities and considerable memory, the NTUU "KPI" 7 TFlops computing cluster is used, which is a part of the integrated national GRID infrastructure of Ukraine containing 31 cluster.

TOOLS OF THE INTELLIGENT ANALYSIS AND SYSTEMIC DATA ADJUSTMENT

The WDC-Ukraine information environment is a distributed information-analytical system intended to support interdisciplinary scientific research.

The lower level of the system ensures the standardization of data management processes. These processes form a continuous life cycle with the following phases: creation, processing, analysis of data, their storage, publication, and reuse [7–9]. Within the framework of the requirements put forward by the World Data System to individual centers, ensuring the full data life cycle is an obligatory condition for the certification.

Based on the concept of the union of data sources and services, Global System of Data Systems, approved by the World Data System, information resources of all the integrable sources and services should be presented as a new integrated source [1, 2, 10]. The solution of this problem is complicated by the fact that intelligent systems of the centers of the World Data System employ different data models, which are based on different (at the best case, client–server or service-oriented) architectures, which often needs specialized tools to be developed to organize the interaction with the inherited software.

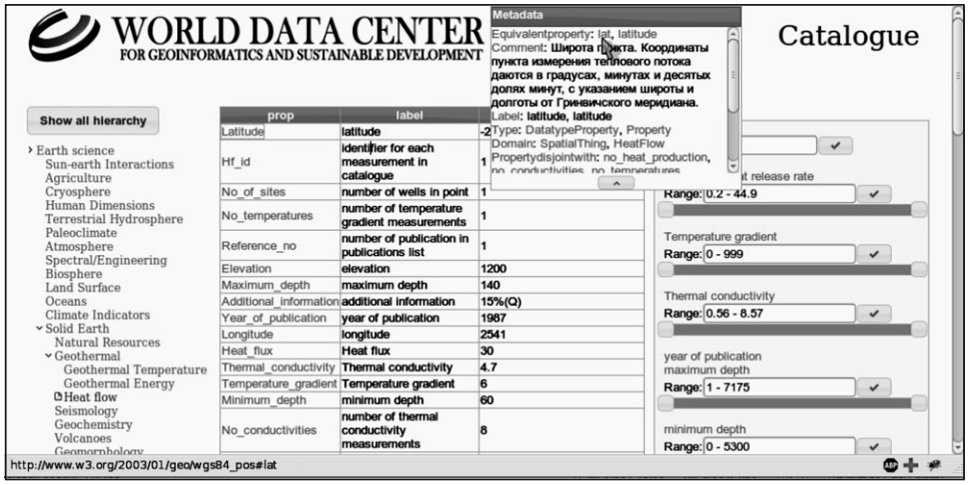

Fig. 1. User WEB-interface for data search and integration.

To solve this problem, the WDC-Ukraine has developed a prototype of a distributed information-analytical system [11], where data and services are integrated at the semantic level with the use of the ontology description language OWL. This language supports the unified data presentation taking into account their semantic properties in the context of the integrated ontology of the subject domain [12]. Data sources are catalogued in this system with the use of the ontology Global Change Master Directory (GCMD) Science Keywords [13].

To implement such a system, the study [12] proposes an agent-oriented approach. Within the framework of service-oriented architecture, it allows integrating not only data but services as well and organizing their interaction due to the application of agents. These agents are specialized migrating software components, which have unique (for their subject domain) OWL-ontologies and XML files determining the parameters of their connection to the source, rules of the projection of the dictionary on the data source, and operation parameters [14].

The users interact with the developed system via the WDC-Ukraine portal http://wdc.org.ua. Figure 1 shows the user WEB-interface, which organizes the virtual space of data sources and services: registration of data sources and services related to the integrated GCMD-ontology, monitoring of the state of data sources and services, acquiring reference metainformation about them; data search and integration: making queries by specifying filter parameters or directly introducing SPARQL query, restricting the search zone based on data sources, reducing the search results to the form defined by the user (XML, JPEG, CSV, and HTML formats are supported, as well as various curves and tables). The system also includes services for analytic data processing (factor analysis, cluster analysis, correlation analysis, etc.).

Using the developed system, the WDC-Ukraine provides the interaction of the organizations being data owners (leading scientific institutions of the NAS of Ukraine) within the framework of the “Network of Networks” model. This principle was proposed by the WDC-Ukraine and was approved by the World Data System as an example for other interdisciplinary centers [4]. The interaction of scientific institutions, including that within the framework of the Russian–Ukrainian segment of the World Data System [11, 15], supported by the NAS of Ukraine, Russian Academy of Sciences, and the Russian and Ukrainian Funds for Fundamental Research, allows the WDC-Ukraine and its partners to successfully conduct interdisciplinary research in geoscience, in particular, analysis of complex ecological, social, and economic systems in the context of their sustainable development (ten joint Ukrainian–Russian projects have been completed).

SYSTEMIC ADJUSTMENT OF DATA OF DIFFERENT NATURE

Interdisciplinary studies usually use data of different nature defined by their objective content, purpose, and the way they are acquired. In this case, developing aggregated interdisciplinary models [3] needs a complex of problems to be solved, which involve reducing these data to common semantics, common range of values, and common units of measure and minimizing the information losses inevitably resulting from the adjustment. Such data adjustment necessitates the solution of

several types of problems. Problems of the first type involve the assessment of the information losses resulting from the data adjustment. Problems of the second type imply the development of a methodology for the quantitative assessment of the consistency of data of different nature. Problems of the third type are related to the development of algorithms and methods for such adjustment.

Generally, studying a certain phenomenon involves analyzing and processing the information about it, which is quantitative and (or) qualitative estimates of the properties of some set of objects $O = \{o_i\}$, $i = \overline{1, n}$, where o_i are the values of identifiers of the nominal scale of objects from the presented set. Such data can be assumed to result from the mappings

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

where I_j, $j = \overline{1, m}$, is a mapping defined on the set of objects ($D(I_j) = O$), with the range of values corresponding to the domain of definition of the index X^j, i.e., $E(I_j) = D(X^j)$.

The semantics of mappings I_j, $j = \overline{1, m}$, and indices X^j identified with them in (1) is formulated based on the objectives of the research and defines the objective content (property of the estimation), the purpose of these estimates, and the way of obtaining (the data can result from measurement, modeling, and expert estimation).

Information Losses in the Transformation of Measurement Scales. The possibilities of the common use of data of different nature depend in many respects on the types of scales they are measured [16].

The purpose of the measurement experiments is to determine the state of empirical system $E = (O, R)$ (where $R = \{r_k\}, k = \overline{1, m}$, is the set of relations among objects from the set O) using the associated measuring system $M = (O', R')$, where $O' = \{o'_k\}, k = \overline{1, n'}$, is the set of symbols and $R' = \{r'_l\}, l = \overline{1, m'}$, is the set of admissible relations on O'. The correspondence between the empirical and measuring system is defined by the surjection $g : E \to M$.

The measurement scale is a triple $S = (E, M, g)$ that completely defines the measurement process [17].

The estimate of information losses during measurements using the scale $S = (E, M, g)$ depends on the estimate of the uncertainty of the inverse mapping $g^{-1} : M \to E$. According to [18], the information obtained from the direct determination of the state of empirical system is defined as $I_E = H(E)$, where $H(E)$ is its own entropy. However, as a rule, the state of an empirical system can only be determined indirectly using the scale $S = (E, M, g)$. In this case, the empirical and measuring systems are considered as dependent and the relation

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

holds, where $H(E \mid M)$ is the conditional entropy characterizing the uncertainty of the state of the empirical system when the state of the measuring system is completely defined.

Thus, the information losses can be estimated by

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

The total information of the systems E and M defined by formula (2) is symmetric:

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

where $H(M|E)$ is the conditional entropy of the measuring system.

If there is no side effect on the measuring system, it is subordinated to the empirical system ($H(M \mid E) = 0$), and expression (3) with (4) taken into account becomes

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

Data conversion from one scale S_1 to the other S_2 can be presented as a measurement process, where S_1 is empirical system and S_2 is measuring one, i.e., it is possible to define a new scale $T_{S_1 \to S_2} = (S_1, S_2, \varphi)$, where $\varphi : S_1 \to S_2$.

The scales S_1 and S_2 are assumed equivalent if there exists $\varphi \in \Phi : S_1 \to S_2$ such that $\Delta I(S_1, S_2) = 0$. It is obvious that with respect to the classes $\Phi = \{\varphi\}$, the set of scales $T_{S_1 \to S_2} = (S_1, S_2, \varphi)$ is divided into equivalence classes (types) S_Φ and data conversion within scales of one type does not cause information losses. If we define $O = X^i \subseteq R$ and $O' = X^j \subseteq R$, then we can classify the scales as follows [17]:

— quantitative scales, for which $\Phi = \{\varphi : x_i = ax_j + b\}$, $S_i, S_j \in S_\Phi$, $x_i \in X^i, x_j \in X^j$, a and b are scale and shift parameters;

— ordinal scales, for which $\Phi = \{\varphi : \forall x_i \le x_j \Rightarrow \varphi(x_i) \le \varphi(x_j)\}$, $S_i, S_j \in S_\Phi$, $x_i, x_j \in X^i$, $\varphi(x_i), \varphi(x_j) \in X^j$;

— nominal scales, for which $\Phi = \{\varphi : \forall x_i \ne x_j \Rightarrow \varphi(x_i) \ne \varphi(x_j)\}$, $S_i, S_j \in S_\Phi$, $x_i, x_j \in X^i$, and $\varphi(x_i), \varphi(x_j) \in X^j$.

These types of scales are the most popular and allow both quantitative and qualitative estimates of the properties of objects under study.

Thus, the solution of data adjustment problems can be reduced to constructing the procedure of transformation $\varphi : S_1 \to S_2$; the scales S_1 and S_2 may be either of the same or of different types. If S_1 and S_2 are of different types of scales, the relation $\Delta I(S_1, S_2) = H(S_1) - H(S_2) > 0$ holds. In this case, the scale S_1 is said to be "stronger" than the scale S_2, and the transition $S_1 \to S_2$ causes information losses. In the inverse transition, where $\Delta I(S_2, S_1) < 0$, an uncertainty numerically equal to $-\Delta I(S_2, S_1)$ appears in the model.

Among the well-known scales, nominal scales are the "weakest," quantitative scales are the "strongest," and ordinal scales are intermediate in this sense.

The numerical estimate of the information losses appearing in the transition from a quantitative to ordinal (or nominal) scale is formulated in terms of the number of references of the original (quantitative) scale *n* and the number of references of the objective (ordinal or nominal) scale *m*. This expression for ordinal scale is defined as

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

and for a nominal scale becomes

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

In formulas (6) and (7) C_m^l and $C_{\max(n,l)}^{\min(l,n)}$ denote the number of combinations, A_m^l and $A_{\max(n,l)}^{\min(l,n)}$ is the number of arrangements, $R(n, l) = \sum_{j=0}^{l} (-1)^j C_l^j (l - j)^n$ is Morgan's number, and $S(n, l) = \frac{R(n, l)}{l!}$ is the Stirling number of the second kind.

Thus, when transforming data to be adjusted, it is necessary to take into account possible either information losses or growth of the uncertainty of the model, which appear when passing from one type of scales to others.

The model uncertainty appearing when passing from "weaker" to "stronger" scales can partially be reduced by using additional information about the objects under study. For example, if when clustering objects according to some index X^j, information about the mean values for each cluster C^j is saved, then using this additional information in the inverse transition from nominal values of clusters C^j to quantitative estimates X^j reduces the uncertainty of the model.

The utility of the additional information in data conversion can be estimated by the expression

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

where $U(I, S_1, S_2)$ is the utility of the additional information I when passing from scale S_1 to scale S_2 and $\Delta I(S_1, S_2)$ and $\Delta I((S_1, S_2) \mid I)$ are information losses in data conversion without and with the use of additional information *I*, respectively.

Figure 2 shows the experimentally obtained dependence between the length of the test log *k* (the number of objects being analyzed) and normalized by $\max_k \Delta I(S_1, S_2)$ values of $\Delta I(S_1, S_2)$ (curve *1*), $\Delta I((S_1, S_2) \mid I)$ (curve *2*), and $U(I, S_1, S_2)$ (curve *3*) for the above example. As is seen, for critical log lengths, where information losses are maximum, the utility of additional information is the greatest and is (in relative units) about 60% of the maximum value of information losses.

Constructing Metrics to Estimate Data Consistency. All the available data (1) can be represented by an "object–property" matrix [19]:

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

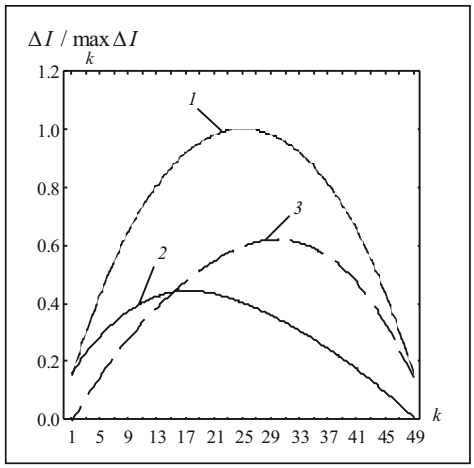

Fig. 2. Dependence of information losses $\Delta I / \max_k \Delta I$ on the test log length k for changing over from nominal to ordinal scale.

where row X_i corresponds to the set of values that characterizes properties of object o_i, and column X^j specifies values of the jth index for the whole sample of objects.

Different "representations" of one phenomenon defined in terms of data, such as samples of objects, time series, projections of properties, etc., correspond to different subsets $V \subseteq X$, which can be specified as compositions on the set of elementary samples $V_i \equiv X_i$ (horizontal sections of matrix (9)) and of elementary projections $V^j \equiv X^j$ (vertical sections of the same matrix). If we denote by $\widetilde{V}$ an arbitrary sample or projection and by $\widetilde{X}$ an elementary sample or projection, then the following expansion of $\widetilde{V}$ in the set $\widetilde{X}$ is true:

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

The approach to constructing metric spaces for the quantitative estimate of data consistency, presented in detail in [20], is based on expansion (10).

In the general case, a sigma-algebra $\mathbb{S}$ can be defined on the set $\widetilde{X} = \{\widetilde{X}_i\}$ [21] and a metric space $(\widetilde{X}, \mathbb{S}, \mu)$ isomorphic to the probability space can be defined. In other words, irrespective of the types of scales in which the data are presented, their values are mapped into nominal events defined on the set $\widetilde{X}$.

If the data are presented in quantitative scales, $\widetilde{X}_i \in \widetilde{X}$ can be considered as elements of the linear space $E(X^1) \times \ldots \times E(X^m)$ or other models can be used, for example, quaternions [22] on which a metric is introduced.

If it is possible to define a space with the norm

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

where w_j are weight coefficients of objects or indices, it becomes possible to make an integral estimate of the representations using the norm

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

and of their proximity

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

For $p=2$ we have a Euclidean space [23], which makes it possible to define in it the concept of orthogonality (independence) of representations.

The definition of distance in (12) may vary depending on the research objectives. For example, $||\overline{V}_{i,j}||$ is often used instead of $||\widetilde{V}_i - \widetilde{V}_j||$, where $\overline{V}_{i,j}$ is some averaging (mean or median value) [17] found as a solution of the problem $\arg\min\limits_{\overline{V}\in\mathbb{S}} (d(\overline{V}_{i,j},\widetilde{V}_i)+d(\overline{V}_{i,j},\widetilde{V}_j))$.

It is possible to use other norms, which most exactly reflect the sense of what representations should be considered close.

For example, to estimate the "proximity" of statistical distributions, the measure

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

is proposed in [20], where P and Q are the estimated distributions, $\frac{P\oplus Q}{2}=\left\langle \frac{p_1+q_1}{2},\frac{p_2+q_2}{2},\ldots,\frac{p_n+q_n}{2}\right\rangle$ is averaged distribution, and $H\left(\frac{P\oplus Q}{2}\right)$, $H(P)$, and $H(Q)$ are the Shannon entropy [24].

This measure can be used to develop a number of estimates being of practical value in solving various problems.

The approach proposed in [20] to the construction of such criteria implies finding the distributions that are the best $P_{\sup}$ and the worst $P_{\inf}$ from the point of view of the solution of a specific applied problem. In this case, the criterion can be defined as follows:

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

With $L(P_{\inf},P_{\inf})=0$, relation (13) finally becomes

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

It is obvious that $S(P)\in\{0,1\}$. The value of this criterion can be considered as a quantitative estimate of the distribution P expressed in a relative scale.

Let there be a set of distributions $\pi=\{P_j, j=\overline{1,m}\}$, and the boundary distributions $P_{\sup}$ and $P_{\inf}$ be defined. Then the total ordering $\alpha\subseteq\pi\times\pi:\langle p_i,p_j\rangle\in\alpha\Leftrightarrow S(p_i)\geq S(p_j)$ can be defined on the set π, i.e., optimization problems $\max\limits_{P\in\pi}(S(P)), \arg\max\limits_{P\in\pi}(S(P))$ can be solved.

Data Conversion Methods. In developing and using data conversion methods, it is necessary to consider the constraints imposed by the types of scales in which the data are presented. For example, data conversion within one type of scales should ensure the absence of information losses. If the types of the original and objective scales do not coincide, such losses should be minimized.

If the data are presented in quantitative scales ($X^j\in R, j=\overline{1,m}$, where R is the field of real numbers [21]) and it is necessary to reduce the data to a unified semantics and range of values, such conversion of data (values of indices) $C_j: X^j\xrightarrow{C_j} R$ should be without information losses, which imposes the monotonicity condition on these transformations:

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

Linear and nonlinear normalizations (monotonic parametric transformations) correspond to these requirements.

Linear normalization is defined according to the expression

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

where $x_{i,j}$ is the value from matrix (9), a is the parameter specifying the displacement; and b is the parameter defining the normalization scale.

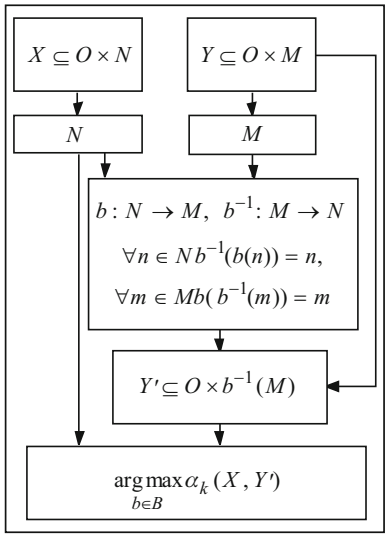

Fig. 3. Schematized bijective adjustment of data presented in nominal scales.

An example of a nonlinear normalization may be the transformation

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

which defines the logistic curve [25], where the parameters a and b have the same meaning as in formula (14).

The parameters of transformations (14) and (15) are selected so that to reduce the initial data to the preset range of values (most often to the interval $[0, 1]$). The parameters of such normalizations should have the same units of measure as the original indices, which ensures the dimensionless of the normalized values of indices. These values should be treated as the position of the object being analyzed with respect to some set of standards. The way of choosing such standards determines the final semantics of the normalized values.

To calculate the parameters a and b, the expressions

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

or

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

are often used, where $\overline{X^j}$ is the mean value, $\sigma(X^j)$ is a standard deviation of index X^j defined on the sample of objects O. In this case, the normalization is carried out with respect to the sample O mean.

If some objects characterized by limiting states are used as standards, then mappings (14) and (15) can be treated as membership functions for linguistic variables [20].

If the indices are specified in nominal scales, the spectrum of equivalent data conversions is limited to bijections. Figure 3 schematizes the bijective adjustment of the data X and Y presented in the nominal scales M and N; within the framework of this scheme, data adjustment is considered as an optimization problem.

As an optimization criterion, this scheme employs probability measure, the Krippendorff alpha [26]:

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

The values of α_k within the range [0.75–1.0] correspond to the high degree, within the range of [0.5–0.75] to average degree, and within the range of less than 0.5 to the low degree of the consistency of data presented in nominal scales.

The study [27] considers the method of data conversion to an ordinal scale of data presented in quantitative scales, provided that the information losses are minimized. To this end, for the index X expressed in a quantitative scale, it is necessary to form a variational series $\widetilde{X} = \langle \widetilde{x} \rangle : \forall \widetilde{x} \in \widetilde{X},\ \widetilde{x} \in X,\ \forall \widetilde{x}_i, \widetilde{x}_j \in \widetilde{X}, i < j, \widetilde{x}_i \le \widetilde{x}_j$, and use the piecewise linear approximation of its cumulant defined as follows:

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

where c_i is the ith value of the cumulant, $\widetilde{x}_l \in \widetilde{X}$ is the lth term of the variational series, and n is the length of this series.

Determining the values of X in ordinal scale involves finding the partition of the variational series $\widetilde{X}$ into segments:

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

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

for which the condition $\sum_{\substack{\widetilde{x} \in \langle \widetilde{x}_i, \widetilde{x}_{i+1} \rangle, \\ i = \overline{1, r-1}}} (\widetilde{x} - M(\langle \widetilde{x}_i, \widetilde{x}_{i+1} \rangle))^2 \to \min$ is satisfied, where $M(\langle \widetilde{x}_i, \widetilde{x}_{i+1} \rangle)$ is the expectation of values of the index X in the segment $\langle \widetilde{x}_i, \widetilde{x}_{i+1} \rangle$ of the variational series. Thus, each segment $\langle \widetilde{x}_i, \widetilde{x}_{i+1} \rangle \in \pi(\widetilde{X})$ is associated with the quantity $M(\langle \widetilde{x}_i, \widetilde{x}_{i+1} \rangle)$, and each of them, in turn, with the value of the ordinal scale:

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

If the series $\left\langle M(\langle \widetilde{x}_i, \widetilde{x}_{i+1} \rangle) \right\rangle$ is used for inverse transformation, then the uncertainty of the model can be described by the dependence represented in Fig. 2.

AN EXAMPLE OF THE SYSTEMIC ADJUSTMENT OF DATA IN MODELING OF SUSTAINABLE DEVELOPMENT PROCESSES

A typical example of applying the above principles of the systemic adjustment of data of different nature is the problem of modeling the influence of a set of threats on processes of sustainable development of a certain territory [27].

According to [3], to determine the human life safety component, we will use the threat space model, where each territory j is associated with the vector

$$\vec{T}r_j = (t_i^j),\ i = \overline{1, n}, \tag{18}$$

with coordinates t_i^j, which characterize the degree of manifestation of the corresponding threats.

To estimate the total influence of the set of threats on individual territories (countries or regions of a country), we will use the component of human life safety C_{sl} defined as the Minkowsky norm for the vector of threats:

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

As the parameter p increases, the response (sensitivity) of the model to variations in each component of the vector $\vec{T}r_j$ increases, and vice versa, its decrease smoothens (coarsens) this sensitivity; therefore, $p = 3$ is assumed in the model.

Let us consider the problem on the example of calculating the influence of a set of threats on territorial divisions (regions and territories) of Ukraine.

Table 1 shows the threats determined by experts with regard for the specific features of the sustainable development of the regions of Ukraine. The threat indices differ in the sense, procedure, and units of measurements, i.e., are of different nature; therefore, the aggregated (integral) estimates of safety were constructed by the systemic adjustment of the values of these indices.

TABLE 1

Threat	Threat index	Units of measurement of the index
Life expectancy drop (S1)	Average life expectancy	Years
Criminality (S2)	Criminality coefficient	Registered crimes per 100,000 population
Corruption (S3)	Corruption perception index	%
Social disparity (S4)	Gini index (income inequality coefficient)	cond. unit
Growth of unemployment (EC1)	Level of the registered unemployment	as of the end of the year, %
Deterioration of the technological infrastructure (EC2)	Deterioration of the fixed capital	%
Drop in the well-being of the population (EC3)	Income per capita	UAH
Contamination of city air (E1)	Atmosphere contamination index	cond. unit
Deterioration of potable water (E2)	Water samples not compliant with the State Standard	%
Environmental pollution (E3)	Density of pollution emission in the atmosphere and water	t / km^2
Growth of greenhouse gas emission (E4)	Density of greenhouse gas emission	$t\ CO_2 / km^2$

With the use of the method of exponential normalizing, the values of all the threat indices were reduced to dimensionless values, which vary within the range [0–1]. The value of 0.5 corresponds to the threat effect average for the country and a value close to 1.0 corresponds to the greatest effect. Determining the critical threshold (0.7 in this case), it is possible to select the critical values for threat indices for each region.

Relations (18) and (19) were used to calculate the component of human life safety C_{sl} (Table 2), which aggregates all the 11 threats presented in Table 1.

According to the values of the human life safety component, regions of Ukraine were divided into four clusters: with high, upper intermediate, low intermediate, and low safety levels.

The first cluster with high level of human life safety ($C_{sl} > 1.1$) includes seven regions: Ivano-Frankivs'ka, Chernovyts'ka, L'vivs'ka, Ternopil's'ka, Kyivs'ka, Chernigivs'ka, and Khmel'nyts'ka regions. These regions are characterized by the moderate influence of social, economic, and ecological threats. Note that the threat "Life expectancy drop" is significant for Kyivs'ka and Chernigivs'ka regions. Moreover, "Corruption" is an important threat for Kyivs'ka region and "Growth of unemployment" for Chernigivs'ka region. The threat "Drop in the well-being of the population" is significant for Chernovyts'ka and Ternopil's'ka regions.

The second cluster with the upper intermediate level of human life safety ($1.1 > C_{sl} > 0.99$) includes eight domains of Ukraine: Kharkivs'ka, Cherkas'ka, Volyns'ka, Vinnyts'ka, Khersons'ka, Poltavs'ka, and Sums'ka regions and the City of Sevastopol'. A high level of threats "Criminality," "Corruption," and "Environmental pollution" are typical of the City of Sevastopol', "Corruption" and "Drop in the well-being of the population" are significant threats for the Volyns'ka region, "Growth of unemployment" and "Deterioration of the technological infrastructure" for Vinnyts'ka and Poltavs'ka regions, and "Corruption" and "Growth of unemployment" for the Sums'ka region.

The third cluster with low intermediate level of human life safety ($0.99 > C_{sl} > 0.87$) includes six regions of Ukraine: Zakarpats'ka, Rivnens'ka, Zaporiz'ka, Zhytomyrs'ka, and Mykolaivs'ka regions and AR of Crimea. Three threats are significant for Zakarpats'ka and Zhytomyrs'ka regions, and four for Mykolaivs'ka region.

The fourth cluster with low level of human life safety ($0.87 > C_{sl}$) includes six regions of Ukraine: Lugans'ka, Odes'ka, Dnipropetrovs'ka, Kirovograds'ka, and Donets'ka regions and City of Kyiv. From three (Lugans'ka region) to seven (Donets'ka region) threats simultaneously influence the low level of human safety in the regions of this group.

TABLE 2

Region	Rating	C_{sl}	Social Threats				Economic Threats			Ecological Threats			
			S1	S2	S3	S4	EC1	EC2	EC3	E1	E2	E3	E4
High level (C_{sl} >1.1)													
Ivano-Frankivs'ka	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
Chernovyts'ka	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
L'vivs'ka	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
Ternopil's'ka	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
Kyivs'ka	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
Chernigivs'ka	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
Khmel'nyts'ka	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
Upper intermediate level (1.1 > C_{sl} > 0.99)													
Kharkivs'ka	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
Cherkas'ka	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
Volyns'ka	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
City of Sevastopol'	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
Vinnyts'ka	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
Khersons'ka	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
Poltavs'ka	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
Sums'ka	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
Lower intermediate level (0.99 > C_{sl} > 0.87)													
Zakarpats'ka	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
Rivens'ka	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
Zaporiz'ka	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
AR of Crimea	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
Zhytomyrs'ka	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
Mykolaivs'ka	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
Low level (0.87 > C_{sl})													
Lugans'ka	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
Odes'ka	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
Dnipropetrovs'ka	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
City of Kyiv	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**
Kirovograds'ka	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
Donets'ka	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**
Least value		0.68	68.71	419.00	24.60	19.91	0.30	37.70	12226.90	3.10	76.57	3.57	45.80
Mean value		1.00	70.55	1001.67	38.86	24.63	2.15	63.58	17082.32	7.12	92.58	34.55	726.31
Greatest value		1.31	73.66	1705.00	60.30	31.63	3.80	97.10	37573.20	14.40	99.69	449.28	11631.22

CONCLUSIONS

The considered methodology of the analysis of complex systems based on interdisciplinary models has been obtained as a result of the systemic adjustment of empirical data, models, and methods from various scientific fields. We have presented mathematical and software–hardware tools for the intelligent processing, analysis, and systemic adjustment of data of different nature, their ordering, assessment of their adequacy, analysis of quality, correctness, etc. As an example, we have used the modeling of processes of sustainable development of administrative regions of Ukraine (regions, AR of Crimea and cities of Kyiv and Sevastopol') annually performed within the framework of the World Data Center "Geoinformatics and Sustainable Development" (WDC-Ukraine) and have considered a set of problems of intelligent analysis and systemic adjustment of economic, ecological, and social scientific data.

REFERENCES

1. Constitution of the International Council for Science World Data System (ICSU WDS), http://icsu-wds.org/images/files/WDS_Constitution_04_04_12.pdf (2012).
2. J. B. Minster, "The ICSU world data system as a global system of data systems," in: Abstr. Proc. 25th IUGG General Assembly "Earth on the Edge: Science for a Sustainable Planet", 2011, Melbourne, Australia (2011), p. 46.
3. M. Z. Zgurovsky, A. O. Boldak, K. V. Yefremov, et al., Analysis of Sustainable Development, Pt. 1, Global Analysis of the Human Life Quality and Safety [in Ukrainian], NTUU KPI, (2010).
4. M. Z. Zgurovsky, A. D. Gvishiani, K. V. Yefremov, and A. M. Pasichny, "Integration of the Ukrainian science into the World Data System," Cybern. Syst. Analysis, **46**, No. 2, 211–219 (2010).
5. M. Z. Zgurovsky, B. E, Paton, and Yu. I. Yakimenko, "State of the art and prospects of the development of the national telecommunication academic network," http://www.uazone.org/inet/uren/uran-dop1w97.html (1977).
6. Yu. Yakymenko, V. Timofeyev, V. Galagan, and M. Dombrougov, "Development and European integration of Ukrainian research and academic network (URAN) for provision of high speed services to science and education," in: Proc. 21st Int. CODATA Conf., 2008, Kyiv, Ukraine (2008).
7. "Create & manage data," http://www.data-archive.ac.uk/create-manage/life-cycle/.
8. C. Shearer, "The CRISP-DM model: The new blueprint for data mining," J. Data Warehousing, No. 5, 13–22 (2000).
9. S. S. Rohanizadeh and M. B. Moghadam, "A proposed data mining methodology and its application to industrial procedures," J. Industr. Eng., No. 4, 37–50 (2009).
10. ICSU Annual Report 2008, http://www.icsu.org/publications/annual-reports/annual-report-2008/annual-report-2008-file.
11. K. Yefremov, "Integration of heterogeneous data sources of Russian–Ukrainian WDS segment based on ontology and agent-oriented approach," in: Proc. JpGU Intern. Symposium 2012, Makuhari Messe, Chiba, Japan (2012), p. 23.
12. K. Yefremov, "Agent-oriented approach for integration of WDC-Ukraine partner network resources," in: Proc. 22nd Int. CODATA Conf., 2010, Cape Town, South Africa (2010).
13. L. M. Olsen, G. Major, K. Shein, et al., GCMD's Science Keywords and Associated Directory Keywords (2007).
14. S. I. Shapovalova, K. V. Yefremov, and A. I. Glukhanik, "Organization of the integrated access to information resources," in: Proc. 11th Intern. Conf. "Intelligent Information Analysis," Prosvita, Kyiv (2011), pp. 101–108.
15. M. Shaimardanov, A. Gvishiani, M. Zgurovsky, et al., "Development of WDS Russian–Ukrainian segment," in: Proc. 1st ICSU-WDS Conf. "Global Data for Global Science", 2011, Kyoto, Japan (2011), pp. 19–28.
16. S. A. Aivazyan, V. M Buhshtaber, I. S. Enjukov, et al., Applied Statistics. Classification and Reduction of Dimension [in Russian], Finansy i Statistika, Moscow (1989).
17. R. D. Luce, D. H. Krantz, P. Suppes, et al., Foundation of Measurement. Vol. 3: Representation, Axiomatization and Invariance, Academic Press, San Diego (1990).
18. E. S. Ventsel', Probability Theory [in Russian], Vyssh. Shk., Moscow (1999).
19. S. A. Aivazyan (ed.), I. S. Enyukov, and L. D. Meshalkin, Applied Statistics: Research of Dependences [in Russian], Finansy i Statistika, Moscow (1985).
20. M. Z. Zgurovsky and A. A. Boldak, "System adjustment of data of various nature in multidisciplinary research," Cybern. Syst. Analysis, **47**, No. 4, 546–556 (2011).
21. A. N. Kolmogorov and S. V. Fomin, Elements of the Theory of Functions and Functional Analysis [in Russian], Fizmatlit, Moscow (2004).
22. J. H. Conway and D. Smith, On Quaternions and Octonions, AKPeters/CRC Press, Abington, Oxfordshire (2003).
23. V. A. Il'yin and E. G. Poznyak, Linear Algebra: A Textbook for High Schools [in Russian], Fizmatlit, Moscow (2007).
24. K. Shannon, Works on Information Theory and Cybernetics [Russian translation], Izd. Inostr. Lit., Moscow (2002).
25. N. Balakrishnan, Handbook of the Logistic Distribution, Marcel Dekker, New York (1992).
26. K. Krippendorff, Content Analysis: An Introduction to its Methodology, Thousand Oaks, Sage, CA (2004), pp. 219–250.
27. M. Z. Zgurovsky, A. O. Boldak, K. V. Yefremov, et al., Analysis of Sustainable Development: Global and Regional Aspects, Pt. 2, Ukraine in the Indices of Sustainable Development [in Ukrainian], NTUU KPI, (2012).