CONSTRUCTION OF THE MAPS OF CONGENITAL HEART DISEASES AND ONCOLOGICAL DISEASES OVER THE TERRITORY OF THREE REGIONS OF UKRAINE USING THE METHOD OF SELF-ORGANISATION OF MATHEMATICAL MODELS.

L.Tymoshevska V.Yeremin, S.Kalkuta, A.N.Timoshevskii,

Institute of Magnetism of National Academy of Sciences, 36-b Vernadsky St., 03142 Kiev, Ukraine, e-mail: liliya@ukron.kiev.ua

A new methodology for processing and analysis of highly noised statistical data that are defined on a strongly irregular grid of observations is proposed. This methodology is based on the method of self-organisation of mathematical models. We have developed algorithms and generated a software suit "Spacer". Using this methodic, statistical data on congenital heart diseases and oncological diseases over the territory of Ukraine have been processed and analyzed. It is shown that these data are inhomogeneous. The maps of the man caused contaminations of these territories have been built. The correlations between the people morbidity and distribution of the man caused contaminations are studied.

Introduction

In ecology and medicine, geology and seismology and many other different fields of science and technology there exists the same problem that concerns processing of a large sets of statistical data defined on a strongly an irregular grid of observation data. Very often these data are highly noised and the number of measurements is limited. Processing and analysis of these data imply extraction of maximum volume of information with a controlled accuracy and require excluding or minimising personalistic decisions. Thus we come to the conclusion that we need methodology that would be the base for the generation of self-adjusted methods for data processing, which could be applied to any measured data with any noise content. Self-adjustment implies generation of optimal mathematical models, which describe adequately the observations defined on an irregular grid of observation data and exclude an influence of researcher's personalistic decisions. Using these models, we can reconstruct the data fields of potentially different origin and forecast their space distribution with a controlled accuracy. Another constituent of the above mentioned methodology is the method for analysis of data, which should enable searching implicit rules that might characterize measured data. Important information could be invisible due to a high noise content, strongly irregular grid of observation and limited number of data themselves.

Method

The authors of this research work propose required methodology that is based on the principles of self-organization. We have developed the algorithms and generated the software suit (***"SPACER"***) that is based on the methods of self-organization of multidimensional mathematical models and enables definition of deterministic and noise components of data defined on a strongly irregular grid of noised observation data. In contrast to all presently existing methods, we can control accuracy of optimal model being able to calculate prognostic errors over a total region of existence of measured data, and also determination coefficient that characterizes quality of the model chosen.

The problem is to determine the values of the unknown function of many variables in any space point, basing on the known values of a function in a finite number of observation points. If the field to be recovered should coincide exactly with the initial function values, then we consider these tasks as interpolation; when recovering should be made with some deviations, such tasks are to be considered as approximation.

Four stages for building algorithms of the field recovering are proposed.

The **first stage** is the choice of the interpolation function form. At present, this procedure has not been formalized and therefore the function is being chosen depending on the information about the process under study. As a result, we have some function of many variables:

$y = F(\vec{x}, \vec{A})$, where $\vec{x} = \{x_1, x_2, ..., x_m\}$ - vector of current coordinates, $\vec{A} = \{a_1, a_2, ..., a_k\}$ - vector of unknown parameters.

At the **second stage**, the function definition boundary is determined. There are two possibilities, as follows: the values of the unknown parameters a_i $(i = 1, k)$ are determined, basing on the observation points and do not depend on the current coordinates, i.e. a_i =constant within the whole interpolation range. Parameters a_i are determined at each interpolation site or in some local range, i.e. a_i depend

on the current coordinates x_i. These cases correspond to the global approximation (1) and local approximation (2).

At the **third stage,** metrics (?) for interpolation functional are chosen, which is minimized at the determination of unknown $\vec{A}$ values.

At the **fourth stage**, the functions of the functional weights are chosen. In such a case the weight S_i of the point *i* does not depend on the current coordinates of the global interpolation. When carrying out a local interpolation, some negative and non-increasing function $u(\vec{x}_t, \vec{x}_i, \vec{B})$ that depends on the initial points and interpolation cites is introduced. In general case, the weight function can include the constants. The values of these constants are to be determined during the minimization of the function, or are to be taken basing on the physical considerations.

$$\left(\sum_{i=1}^{n} \left| y_i - F(\vec{x}_i, \vec{A}) \right|^p S_i u(\vec{x}_t, \vec{x}_i, \vec{B}) \right)^{1/p} \xrightarrow[A]{} \min$$

where: $\vec{x}_t = \{x_{t_1}, x_{t_2}, ..., x_{t_m}\}$ is a current interpolation point; $\vec{x}_i = \{x_{i_1}, x_{i_2}, ..., x_{i_m}\}$ is the *i*-th point of initial data; $\vec{B} = \{b_1, b_2, ..., b_l\}$ are the parameters that determine the properties of the weight function.

And finally, the choice of suitable interpolation scheme should be estimated using specific numerical measurements that depend on the quality on interpolation in prognostic points. Methods of the model self-organizations are used at the all of stages for making the best decision.

For analysis of measured data and search of implicit rules we have generate algorithms and programs that realize the method of mixture separation (the second part of the ***"SPACER"*** software suit). The procedure of a mixture separation belongs to parametric methods of the estimation of probability density function. It should be used in those cases when it is necessary to extract homogeneous groups of data and classify the results of observations. The premise for the method of a mixture separation is the fact that every homogeneous group could be presented by its probability density function *f(X,A)*, where *A* parameter is some vector of values, which defines the form of distribution. The problem could be stated as follows: Let we have observation data $?_1, \ldots, ?_N$, which should be classified. For that, the equation of a finite mixture of distribution density functions could be written as:

$$h(x) = \sum_{i=1}^{M} P_i f(X, A)$$

where: M – is a number of homogeneous groups in summary sampling; A_i –parameters of *i*-th density function; P_i – portion of *i*-th group (group probability).

Thus, the problem of a mixture separation and classification of the results by groups comes to the task of evaluation of unknown parameters $M, \{A_i, P_i\}$ (*i*=1, M).

There exist several methods of these parameters evaluation. The most prevailing is the method of maximum of plausibility.

Results

In this work, we have demonstrated application of ***"SPACER"*** suit for processing and analysis of statistical data for two kinds of congenital diseases, namely, cancers and heart diseases over three regions of Ukraine (Zhytomyr, Kiev and Tchernigov regions). Using the methods that we have developed, we can determine an influence of different environmental contaminations on those diseases. As the factors that characterize the rate of environmental contamination over three regions of Ukraine we have used pollution by radioactive Cs-137 and Sr-90 isotopes and pesticide pollution. The information about each kind of disease and pollution is shown by 81 points. Geometrically, these points correspond to the location of regional capital cities. Thus, the information in each point is integral and characterizes morbidity and ecological situation of a whole region.

On the first stage of data processing we have synthesized optimal three-dimensional models and built the maps of distribution of 2 kinds of morbidity (Fig.1-2) and maps of space distribution of Cs and Sr isotopes, total radiation doze (Fig.3a,4a,5a) and pesticide distribution (Fig.6a). Accuracy of the maps is within 70-80%, which show high reliability of the results obtained.

On the second stage we have used the method of the mixture separation (second part of ***"SPACER"***) for analysis of measured data. Our calculations have shown that the data for two kinds of diseases are not homogeneous and could be presented by 3 classes (Fig.1b,2b). On the map of oncological diseases (Fig.1a), these classes are marked with the red, green and yellow. The same calculations were carried out for classifying initial data on the heart diseases. The results of classification and classes space distribution are shown in Fig.2a and Fig.2b. From Fig.1a and Fig.2a it follows that the space distribution of oncological and heart diseases differs significantly. Application of the method of mixture separation to the investigation of initial data on the space distribution of isotopes and pesticides has shown that Sr-90 distribution contains 2 classes of data (Fig.3a, yellow and red), and Cs-137 distribution, total radioactive doze and pesticide distributions contain 3 classes of data (Fig.4a,5a,6a, red, green and yellow).

Figure 1: (a) - The map of cancer diseases; (b) - Division of mixture for cancer diseases.

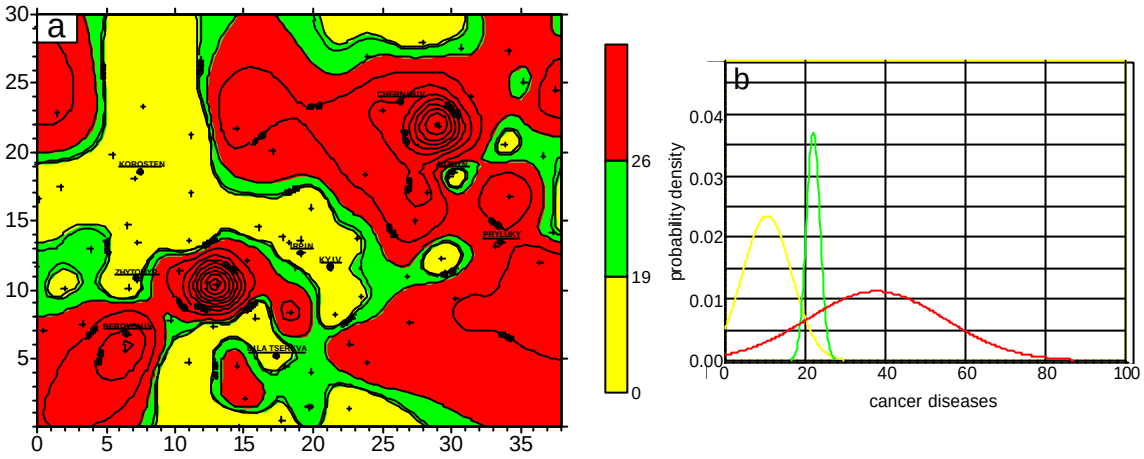
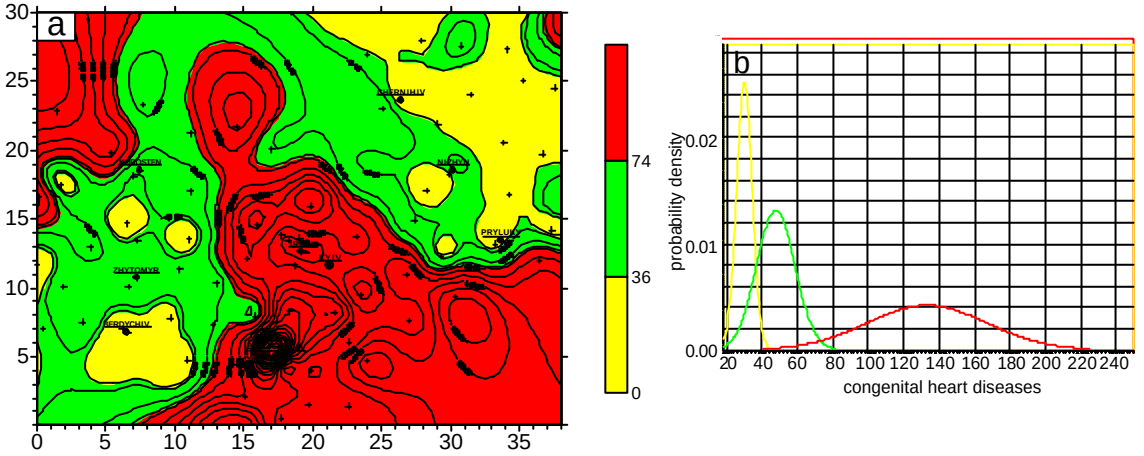


Figure 2: (a) - The map of congenital heart diseases; (b) - Division of mixture for congenital heart diseases.



Building of the maps of the space distribution of morbidity and different territory contaminations enables calculation of the maps of space distribution of coefficients of correlation between the morbidity and kind of pollution. The results of these calculations for oncological diseases are shown in Fig.3b,4b,5b,6b. The same results for the heart diseases distribution are shown in Fig. 3?,4?,5?,6?. From the results of our investigations we can conclude that environmental pollution has a stronger influence on congenital heart diseases then on oncological diseases. In order to clarify results in details it is necessary to enlist the service of medical experts and to continue further joint studies using additional medical data.

Figure 3:
(a) The map of aprox. average soil contamination of manned settlements with Sr-90 (Ci/km2), 1992
(b) The map of coefficient of correlation between cancer diseases and Sr-90
(c) The map of coefficient of correlation between heart diseases and Sr-90

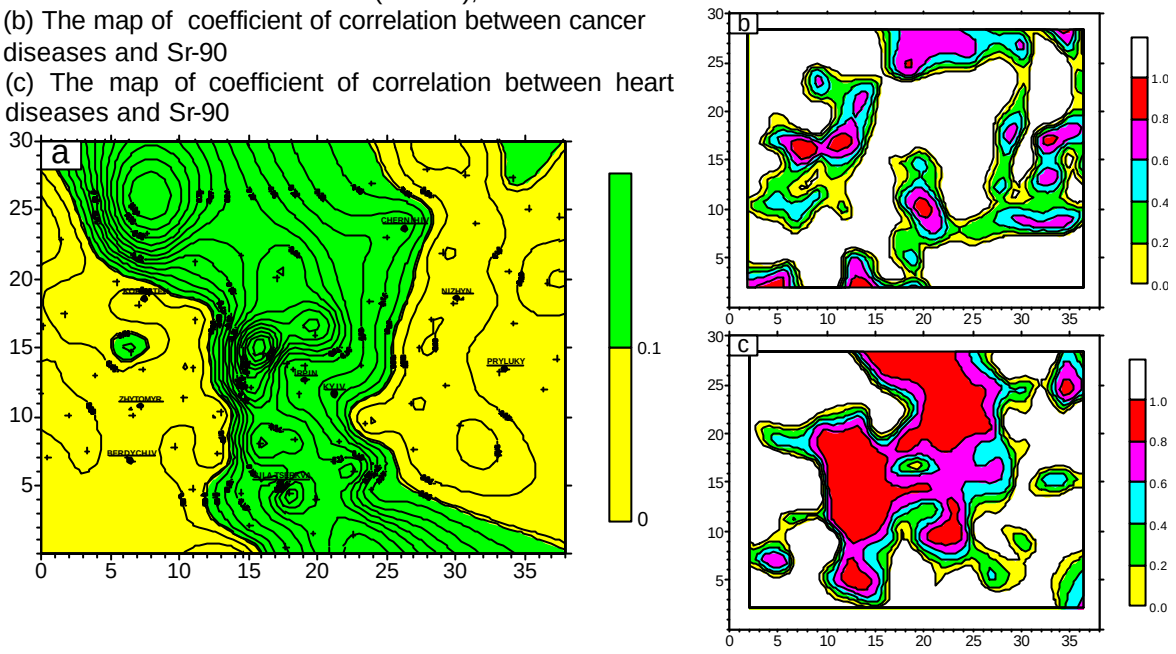


Figure 4:
(a) The map of aprox. average soil contamination of manned settlements with Cs-137 (Ci/km2), 1992.
(b) The map of coefficient of correlation between cancer diseases and Cs-137
(c) The map of coefficient of correlation between heart diseases and Cs-137

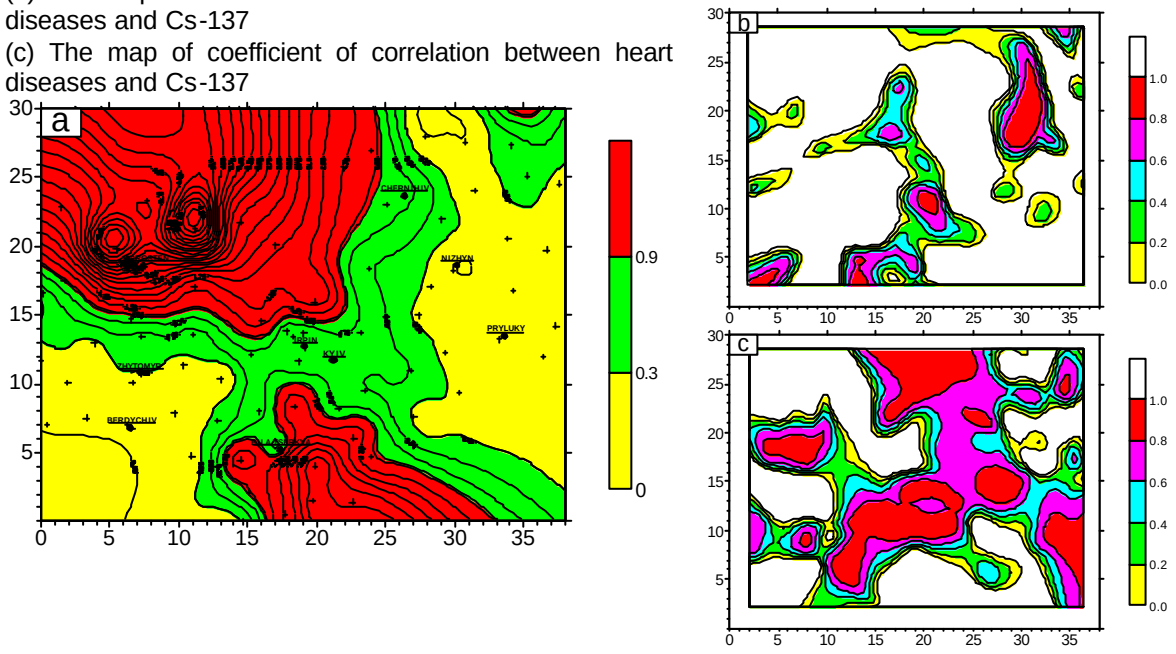


Figure 5:
(a) The map of the average designed all year total doze of Chernobil irradiation (mBer), 1992.
(b) The map of coefficient of correlation between cancer diseases and total doze of Chernobil irradiation
(c) The map of coefficient of correlation between heart diseases and total doze of Chernobil irradiation

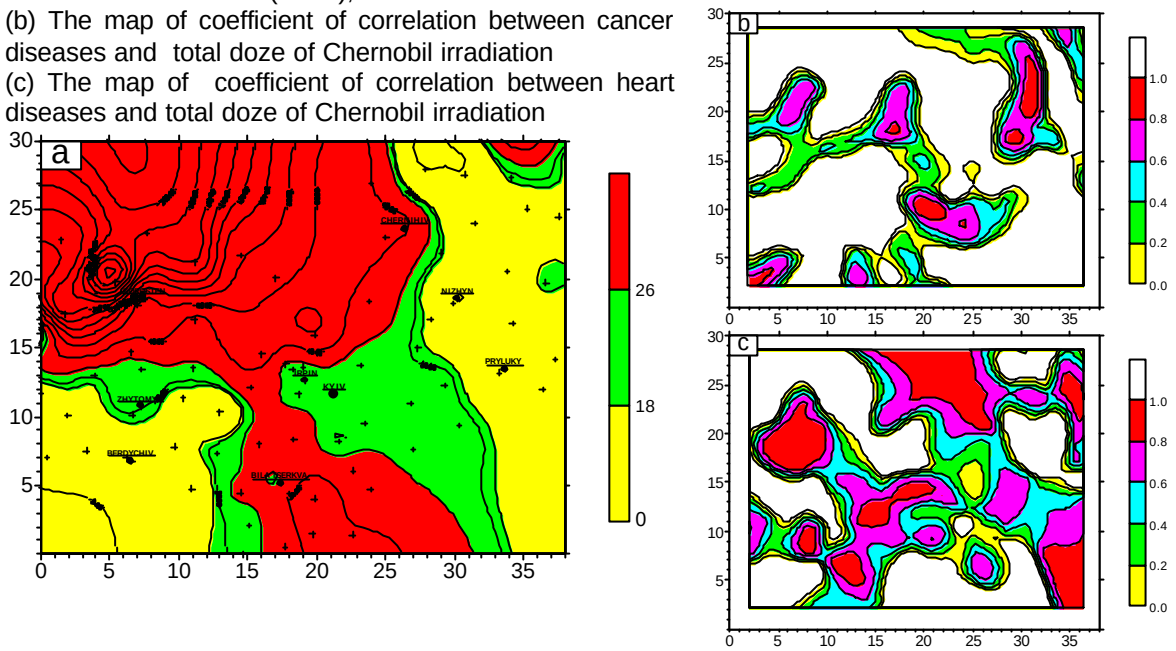
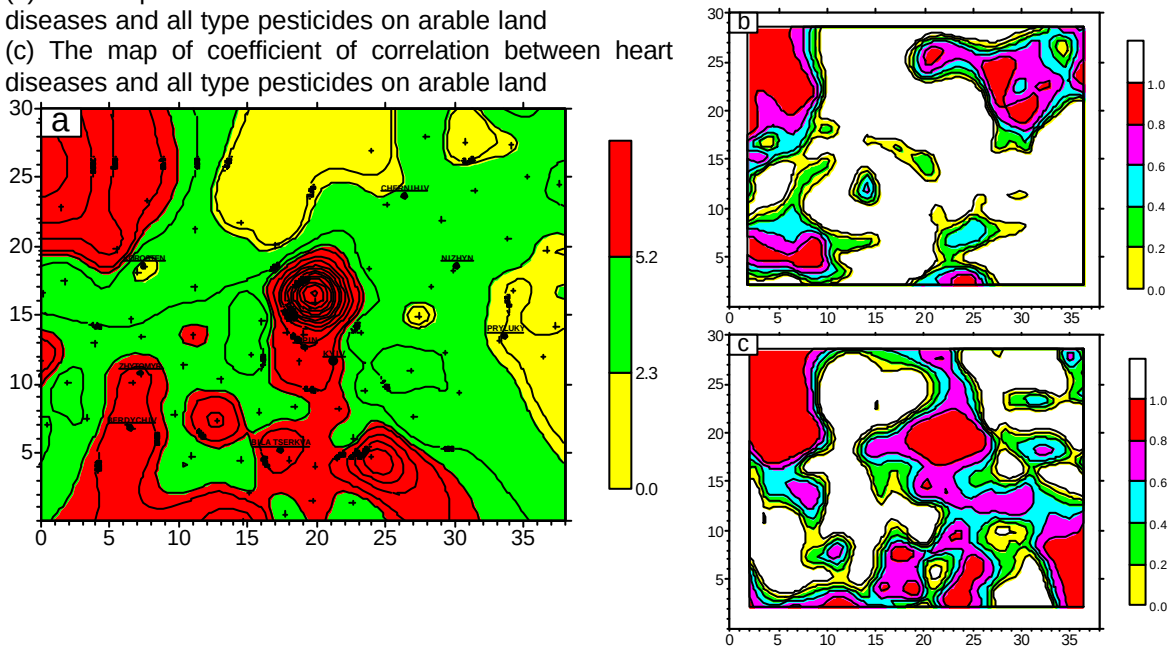


Figure 6:
(a) The map of the all year loading of all type pesticides on arable land (Kg/Hektar), 1988.
(b) The map of coefficient of correlation between cancer diseases and all type pesticides on arable land
(c) The map of coefficient of correlation between heart diseases and all type pesticides on arable land



References

(1) A. Timoshevski, V.I. Yeremin, S.A.Kalkuta "New Method of Environmental Assessment Based on the Methods of Self-Organization of Mathematical Models" The proceedings International Environmental Modeling and Software Society, Lugano, Switzerland, 24-27 June 2002, v. 3, p.542-547.

(2) A. Timoshevskii, V. Yeremin, S. Kalkuta "New method for ecological monitoring based on the method of self-organising mathematical models" Ecological Modelling, 162, ? 1-2, 2003, p. 1-13