Applied Interval Analysis of Big Data Using Linear Programming Methods

Nikolay Oskorbin[0000-0003-2902-0964]

Altai State University, Lenin Ave. 61, 656049 Barnaul, Russia
osk46@mail.ru

Abstract. The article describes the problems of mathematical modeling of processes using an experimental database and a knowledge base. This research relates to multidimensional dependency building. It uses regression analysis and machine learning techniques within the framework of probability theory and mathematical statistics. A large observation table often cannot be processed on a single computer. The analysis of such data requires parallel computations and in this article it is carried out by the method of interval mathematics, which allows performing such computations. The analysis of linear dependences on parameters is reduced to solving systems of interval linear algebraic equations. Among the approaches to systems study known in the literature, an approach was chosen that takes into account the so-called “single set of solutions”. This method provides a guaranteed estimate of the required dependencies and allows the use of linear programming in some cases. Using this method, interval forecasts of the output variable of the modeled process are calculated. Interval estimates of the parameters of the studied dependence were also obtained. Two methods of sequential and parallel analysis of a large database are proposed, using methods for solving large-scale linear programming problems. The optimality of the algorithms is substantiated using the well-known technique of removing constraints in optimization problems of large dimension.The research was carried out on model processes and on real data of statistics of road traffic accidents in England.

Keywords: Big Data, Applied Interval Analysis, Linear Programming, Guaranteed Estimation of Parameters of Linear Dependencies.

1 Introduction

In recent years, Big Data Analysis Methods have been significantly developed in connection with new opportunities for collecting, storing, transmitting situational information and the use of artificial intelligence technology to support decision-making based on them [1]. In practice, the analysis of large data tables requires the use of parallelization schemes in computational algorithms. The use of special high-performance computing systems and cloud technologies for solving complex problems is becoming effective.

Big data is characterized by a large number of observations and the lack of the necessary structure for representing factors and output variables. Therefore, the analysis of these data sets is carried out in two stages. The first stage is associated with solving various problems of structuring the information flow of data: identifying essential factors; digital coding of qualitative assessments of factors; transformation of descriptive characteristics of factors in natural language; data coding in order to obtain a digital table of input and output variables of the modeled process.

At the second stage, a study of causal relationships, assessing the adequacy and performance of the obtained empirical model is carried out. Among the methods of the second stage, both classical probability-theoretic methods [2, 3, 4, 5, 6] and relatively new ones, in particular, applied interval analysis [7, 8, 9] are used.

Currently, regression analysis based on the least squares method is traditionally used for big data analysis. Methodological approaches to this analysis are presented in [2, 3]. In [4, 5], the use of high-performance computing systems for solving problems of regression analysis of big data is considered.

The initial idea and the first applications of the interval approach to data analysis are presented in [7] to estimate the parameters of the linear dependence of the output variable, measured with interval error, on n input variables accurately measured in each of the *N* tests. The processing of the experimental data table of the model was carried out by solving special linear programming (LP) problems.

The development of this approach in science and practice has been associated in recent decades with the use of interval approximation of experimental data [8, 9, 10, 11]. New methods of interval mathematics had a significant impact on the formation of interval data analysis [12, 13]. Interval data analysis in a number of applications is more efficient in comparison with known methods and has significant development potential, including for solving big data processing problems

In the general case, the problem of constructing linear dependences with respect to parameters with interval measurement errors of all variables is reduced to solving interval systems of linear algebraic equations. Among the set of solutions for interval systems of linear algebraic equations known in the literature [12, 13], one set of solutions is selected that provides a guaranteed estimate of the desired dependencies. For the purposes of this work, it is useful to be able to solve mathematical problems by methods of mathematical programming. These methods can also be used to solve high-dimensionality optimization problems [14, 15].

The research is carried out in the following order:

- the theoretical foundations of applied interval analysis in the modeling of linear processes are considered;
- methods of sequential analysis of a large database of experimental data in one computer and parallel data analysis in high-performance computing systems are proposed;
- computer modeling of algorithms for big data analysis is carried out;
- the last section analyzes the real data of road traffic accidents statistics in England for the period of 2005 – 2017.

Using the considered approach in the analysis of big data, interval estimates of the output variable of the modeled process are obtained for given values of the input variables and interval estimates of the parameters of the studied dependence. Two algorithms are proposed for the implementation of the method using sequential reading of database rows and parallel computer calculations.

The optimality of the calculations is substantiated using the method of relaxation of constraints when solving optimization problems of large dimensions [14, 15]. Computer modeling of the proposed algorithms for analyzing big data has been carried out in order to study the possibility of their use in practice and to assess the errors of their limited implementations. The study was carried out on model processes under the conditions of the feasibility of the assumptions of interval data analysis and on real data, the source and description of which are presented in [16, 17].

2 Methods and Data

2.1 Optimization Methods in Interval Data Analysis

Mathematical models of processes are represented as a scalar function, the input and output variables of which are generally measured for each of N observations with interval errors. It is assumed that the systematic components of errors in measuring variables are equal to zero. This general case of analyzing such a database with all observation errors will be discussed later.

Further, we assume that the variables $\mathrm{x} = (\mathrm{x}_1, \dots, \mathrm{x}_\mathrm{n})$ are measured exactly (without measurement errors), and the measured value of the output variable is the interval:

$$Y_j = \left[y_j^H,\ y_j^V\right];\ y_j^H = y_j^M - \varepsilon_j^0;\ \ y_j^V = y_j^M + \varepsilon_j^0;\ j = 1, \dots, N. \tag{1}$$

Here $\mathrm{y}_\mathrm{j}^\mathrm{M}$ is the measurement of the output variable in the j-th observation; ε_j^0 – estimate of the maximum value of the modulus of the interval measurement error.

Then the unknown values of the true coefficients of the linear model:

$$y = a_1 x_1 + \dots + a_n x_n \tag{2}$$

satisfy the system of *N* two-sided inequalities:

$$y_j^H \le a_1 x_{1j} + \dots + a_n x_{nj} \le y_j^V;\ \ j = 1, \dots, N. \tag{3}$$

Let *M* denote the set of values of the vector $\mathrm{a} = (\mathrm{a}_1, \dots, \mathrm{a}_\mathrm{n})$, which satisfy the system of inequalities (3). In the literature, this set is called the "set of uncertainty" or "information set" [12, 13].

The goals and results of data analysis are considered complete when the set *M* is not empty, bounded and allows obtaining process estimates with a given accuracy.

In the case of an empty set *M*, the information of the knowledge bases and databases is inconsistent and requires their correction. The case when the set Mis not bounded indicates that there is insufficient information for analysis and it is required to expand the composition of databases or knowledge.

Let us consider the main applied problems that are solved when modeling processes.

1. *The problem of forecasting the output variable at a given point of the factor space* $x^P = (x_1^P, \dots, x_n^P)$. The interval estimate $[y^H(x^P), y^V(x^P)]$ is obtained by solving two LP problems:

$$y^H(x^P) = \min_{a\in \mathrm{M}}(a_1x_1^P + \cdots + a_nx_n^P); y^V(x^P) = \mathrm{m}ax_{a\in \mathrm{M}}(a_1x_1^P + \cdots + a_nx_n^P). \quad (4)$$

2. *Interval estimation of the parameters of the sought dependence*. In applied interval analysis, the set of true values of the vector $a$is specified by the information set, but for its visualization in practice, the hyper-rectangle approximation is traditionally used. This representation can be considered as independent interval estimates of the components of the vector a. To calculate them, it is enough to solve *2n* linear programming problems. For example, the guaranteed estimate of the coefficient a_1 belongs to the interval $[a_1^H,\ a_1^V]$:

$$a_1^H = min_{a\in \mathrm{M}}\, a_1\,; \qquad a_1^V = max_{a\in \mathrm{M}}\, a_1. \quad (5)$$

3. *The projection of the point a^P onto the set M*. The point a^P belongs to the set *M* if and only if δ is equal to zero, where δ is the solution to the following quadratic programming problem:

$$\delta = min_{a\in \mathrm{M}}\|a^P - a\|. \quad (6)$$

In practice, problem 3 is solved to study the properties of the information set and to test the feasibility of the initial assumptions of applied interval data analysis, including for assessing the significance of the selected input variables.

4. *Point estimation of parameters of linear models (method of the center of uncertainty).* In some cases, it is required to check the feasibility of the initial assumptions (for example, in the case when *M* is an empty set) or to obtain point estimates of the model parameters. One of the ways to solve this problem is associated with the "expansion" or with the "contraction" of the set *M*. Let us set the information set *M* (*k*) by the following system of inequalities:

$$y_j^{\text{И}} - k\varepsilon_j^0 \le a_1x_{1j} + \cdots + a_nx_{nj} \le y_j^{\text{И}} + k\varepsilon_j^0; \quad j = 1, \dots, N; \quad k > 0. \quad (7)$$

The system of inequalities (7) coincides with (3) for $k = 1$. This parameter is called the coefficient of expansion (*k*> *1*) or contraction (*k* <*1*). The next minimum problem is called the problem of finding the center of uncertainty:

$$k^* = min_{a\in \mathrm{M}(k)}\, k\,. \quad (8)$$

As the practice of solving problem (8) on real and model data shows, the minimum is reached at a single point, which can be considered as a point estimate of the parameters of the modeled process. Thek^* value is an indicator of the fulfillment of the initial assumptions of interval data analysis. Thus, the indices of active observations for $k^* > 1$ allocate a portion of the database observations, among which gross errors in data recording or underestimated errors in measuring variables are possible. Such information can be used to adjust the database and knowledge base.

5. *The task of eliminating insignificant factors of the modeled process*. One of the ways to solve this problem in applied interval analysis is based on the use of interval estimation of the model coefficients according to (5). If the zero value of the investigated parameter belongs to the found interval, then the corresponding input variable can be considered insignificant and the factor space can be reduced. In practice, there are other methods for solving this problem using, for example, the results of the analysis of complete and reduced databases.

It should be noted that the considered mathematical formulations of data analysis problems do not change in the general case of the presence of measurement errors for all variables. The changes concern systems of inequalities (3) and (7), which define an information set as a set of solutions to interval systems of linear algebraic equations.

We will consider these differences below in computer modeling of algorithms for analyzing big data.

2.2 Algorithms for Interval Analysis of Big Data

It is required that the computational process provides an optimal solution to one of the LP problems given above.

Let us introduce the following notation:

- $J = \{1, \dots, N\}$ – indexes of records of the complete database;
- J_l – partition of the set J – indices of the l-th piece of data; $l = 1, \dots, m;\ m$ is the number of chunks allocated when splitting big data;
- I_l – a set of observation indices that are active when analyzing the l-th chunk of data;
- M_l – information set when analyzing the l-th piece of data;
- I_0 – indices of observations that are heuristically allocated at the initial stage of calculations from the totality of all observations and provide non-emptiness and boundedness of the corresponding information set.

If it is not possible to single out such observations from the entire database, then the set of data is incomplete or inconsistent and, therefore, needs to be corrected. In particular, such a situation can arise for n>N or for an insufficient number of different points of the factor space. Further, we assume that such a situation does not arise when analyzing the data and the set I_0 selects n linearly independent constraints of the LP problem being solved ($|\mathrm{I}_0| = \mathrm{n}$).

Let us write the algorithm for sequential processing of a large base in the following form.

1. *Step 0.* Let us define a partition of the set *J* into subsets $J_1, \dots, J_m$, select the data analysis problem and the objective function of the corresponding LP problem. We put μ = 0.
2. *Step 1.* We form the set I_0 and set *l* = 1.

3. *Step 2.* Let us solve the analysis problem for observations with indices $J_l \cup I_{l-1}$, which define the set M_l. If this set is empty, then the calculations are stopped with the issuance of the corresponding message.
4. *Step 3.* Selecting the indices of active constraints in the LP problem solved at step 2. If $I_l \neq I_{l-1}$, then we set μ = 1. Further, we *put l* = *l* + 1. If$l \leq m$ go to step 2.
5. *Step 4.* If μ = 1, we put $I_0 = I_k;\ \ l = 1$, go to step 2. Otherwise, we analyze the results of solving the LP problem and the information set of the database, which coincides with the set M_m.

The possibilities of implementing this computational algorithm are determined by the acceptable dimension of the LP problem for the selected computer in step 2.

Let us consider an algorithm for parallel processing of big data, which is a variant of hierarchical algorithms for solving large-scale problems, similar to the algorithms in [18]. The task of the Center is to form the set I_0 and its sequential refinement until the condition of optimality of the solution of the selected LP problem is satisfied for the entire large database. We will present the algorithm according to a similar scheme given above.

1. *Step 0 - Task of the Center*. We define the partition of the set *J* into subsets $J_1, \dots, J_m$, select the data analysis problem and the objective function of the corresponding LP problem.
2. *Step 1 - Task of the Center*. We form the set I_0 and transfer the corresponding rows of the observation matrix to the data exchange buffer.
3. *Step 2 - Task of the Computers*. We solve in each computer *l* the analysis problem for observations with indices $J_l \cup I_0$, which determine the set $M_l;\ l = 1, \dots, m$. If there is a computer for which this set is empty, then the calculations are stopped with the issuance of the corresponding message. Next, we select the index sets I_l of active constraints, and transfer the corresponding rows of the observation matrix to the data exchange buffer.
4. *Step 3 - Task of the Center*. We compare the index sets I_0 and $I_l;\ l = 1, \dots, m$. If there is a computer *l* for which $I_l \neq I_0$, then using all active observations we form a new set I_0 and go with it to step 1. Otherwise, we analyze the results of processing the entire database, the composition of which is represented by the set I_0.

Specific implementations of the described computational algorithm are determined by the nature of the big data and the software of the used computing system. It should be noted that at step 3 of the algorithm, the Center can refine the composition of active constraints by solving the LP problem with constraints that correspond to the index sets I_0 and $\mathrm{I}_\mathrm{l};\ \mathrm{l} = 1, \dots, \mathrm{m}$. In the case when the LP problem turns out to be a problem of large dimension, the Center can use the first algorithm for sequential data processing.

In addition, it should be emphasized that in the considered algorithms there are no requirements for the number of observations when partitioning a large database into chunks.

3 Results and Discussion

3.1 Computer Modeling of Interval Analysis of Big Data

This section discusses the implementation of distributed computing, taking into account the limited capabilities of the selected software and hardware tools. Computer modeling of big data analysis processes is carried out in the Excel environment using the "Search for a solution" tool. The maximum size of the database for the number of model variables is determined by the capabilities of this tool.

We take into account (in our case of computer modeling) that in order to check the optimality of the calculations, it is necessary to find a solution to the LP problem for the entire database, and its dimension in the number of variables should not be more than 200, and in the number of main constraints – more than 100.

Let us consider the mathematical problems of computational experiments for interval data analysis in the general case of the presence of observation errors for all variables. These tasks relate to obtaining estimates of the information set. This section uses the notation accepted in the literature on the theory of interval systems of linear algebraic observations and the traditional notation of linear algebra and LP.

Interval systems of linear algebraic observations in matrix form are written by an interval (N × n) matrix of coefficients and an interval (N × 1) vector of the right-hand side in the following form:

$$\boldsymbol{A}x = \boldsymbol{B}\,. \tag{9}$$

Elements of matrices A and B are interval estimates of the measurement results of input and output variables in N observations and are conventionally represented by inequalities: $\mathrm{A}^{\mathrm{H}} \le \mathrm{A} \le \mathrm{A}^{\mathrm{V}};\ \mathrm{B}^{\mathrm{H}} \le \mathrm{B} \le \mathrm{B}^{\mathrm{V}}$.

The values of the vector $\mathrm{x} \in \mathrm{R}^{\mathrm{n}}$ in (9) correspond to the estimates of the parameters of the linear dependence, and the combined set of solutions Ξ_{uni} corresponds to the set of uncertainty described above. In work [13], it is argued that "computing for the combined set of solutions of external coordinate-wise estimates with any given absolute or relative accuracy is an NP-hard problem".

In the particular case of positive components of the solution to interval systems of linear algebraic observations, a single set of solutions is given by a system of linear inequalities, which we write in the following form:

$$\Xi_{uni} = \{x \in R_{+}^{n}: A^{V}x \ \ge B^{H};\ A^{H}x \ \le B^{V}\}. \tag{10}$$

Let us write down LP problems for the interval estimation of the value of the output variable b at a given point $\mathrm{a_p} \in \mathrm{R}^{\mathrm{n}}$ of the factor space:

$$b^{H}(a_p) = min_{x\in\Xi_{uni}}\, a_p x;\ b^{V}(a_p) = max_{x\in\Xi_{uni}}\, a_p x. \tag{11}$$

Thus, LP problems in the selected version of applied interval data analysis have n variables, and the number of constraints is equal to twice the number of observations of the analyzed portion of the data base.

Let us move on to the computer implementation of computational algorithms. In general, the number m and the size of the data portions must satisfy the inequality $2(\mathrm{n}+\mathrm{N}/\mathrm{m})<\mathrm{D}$, where D is the maximum number of constraints in the LP problem allowed by the optimization software package.

In our case (D = 100), n = 20 was chosen for the main variant of the experimental base, and the admissible 50 observations were cut into 5 portions. Note that the minimum number of portions in our case can be three, for example, with the number of observations 17, 17, 16, respectively. In computational experiments, other variants of the size of the observation tables were also investigated, including much larger (50x20), for linear processes of different parameters.

In all variants, the observation table was filled in the selected intervals for the input variables and for measurement errors with uniform pseudo-random numbers by the Excel function RAND(). The values of the output variable for the given parameters of the linear dependence were modeled with an interval error. For the basic version of the database, the intervals for the input variables were equal to [5, 100], for theirerrors – [-1, 1], for the observation error of the output variable – [-2, 2]. The value of the dependency coefficients for all variables was 10.

Computational experiments were carried out in one Excel workbook and the program scheme for the two computational algorithms was chosen the same: some of the sheets were occupied by the database generators, then one of the LP problem (10), the Center problem (step 1) and 5 computer tasks (step 2). The differences between the algorithms for sequential and parallel analysis of data pieces consisted in different transmission schemes for the index set I_0 formed and adjusted by the Center for observations, which determine the matrix of the basic variables of the LP problem.

For the main version of the database with its repeated updates, it is shown that the proposed computational technologies make it possible to obtain an optimal solution to the LP problem when analyzing big data. All other things being equal, parallel analysis is more preferable in terms of the number of LP subproblems for obtaining an exact solution.

The number of LP problems solved at Step 2 of the algorithms essentially depends on the initial set I_0 of the entire LP problem. The corresponding matrix of constraints, in addition to the absence of linearly dependent rows, must be well conditioned. In this case, it is shown that, on average, for m solved LP problems at step 2, it is possible to obtain the optimal or close to it value of the desired estimate. According to the experience of solving large LP problems [8, 9], in practice, one should expect a fast approximation (in 1, 2 runs) in the vicinity of the optimal estimate and slow motion to its exact value.

In computer modeling, the features of the application of linear programming methods to identify variables, the influence of which on the output variable of the process is absent or not significant, are considered. To solve this problem of data analysis, the hypothesis of the belonging of the zero value of the investigated coefficient to its interval estimation was tested by solving the LP problems (5).

In this case, you can use the solution of nonlinear programming problems (6) for the selected set of variables. The composition of variables for checking their significance is obtained by the method of the center of uncertainty by solving problem (8), a version of which in our case is reduced to two criterial nonlinear programming problem.

It should be noted that the efficiency of solving the problem of big data analysis in practice could be increased by modifying the parallelization schemes in the proposed algorithms, using additional methods of organizing the computational process and using high-performance computing systems.

3.2 Analysis of Road Accidents in England

The study is conducted using large baseline data of road traffic accidents (RTA) throughout England for 2005-2017. Sources from the Internet and a description of the accident data are presented in [16, 17]. The general database, including its main variables and records of the factors of each accident, obtained from the Internet, takes hundreds of megabytes and requires universal software for processing.

In our case, a sample of the records of road accidents with fatal outcomes was made. The total number of such accidents registered in the database is 26370, including 17010 on rural roads. For the analysis, data on accidents on rural roads in the county of Cornwall in the south-west of England were selected. This data was converted and processed in the MS Excel environment in three stages.

At the first stage of the transformation of records, some of the variables were excluded during the formation of the working database, including a number of fields were excluded. These are, for example, road class for 2-road accident participant, highway number for 2-road accident participant, highway number for 1 road accident participant, traffic control method, geographic Cartesian coordinates, speed limit level, weather conditions, astronomical time of an accident, etc.

These exceptions are caused, firstly, by the fact that for some road accidents there are no complete data sets for the excluded items, and, secondly, we considered them insignificant in our estimates. In particular, weather conditions are related to the condition of the road surface (dry, wet), and the time of theaccident is associated with the level of illumination (darkness, daylight).

The next stage of data analysis is converting the table into numerical formats in accordance with the requirements of applying formula (2). It is necessary to highlight the time period in the description of the causes of the accident. This time period is the number of the month, starting from January 2005 to December 2017. Summing up the number of accidents for each month, we obtain for analysis a data table for 13 years, the number of rows of which is 156. Thus, the resulting database for processing in MS Excel is large and exceeds the capabilities of the "Search for a solution" tool.

The last transformation of the data table is associated with the decision to take into account the time factor by encoding from 1 to 156 and take into account the quarter number of each record as an input variable. In the considered case, a mathematical model similar to (2) takes the following form:

$$y = a_0 + a_1 t_1 + K_1 d_1(t) + \cdots + K_4 d_4(t) + Prd_{Pr}(t) + Cvd_{Cv}(t) + Cpd_{Cp}(t). \quad (12)$$

In the mathematical model (12), the following variables are identified: $\mathrm{a}_0,\ \mathrm{a}_1$–free term of the regression equation and the coefficient of the time trend of the number of accidents in the selected area; t – month number; $\mathrm{K}_\mathrm{i}\mathrm{d}_\mathrm{i}(\mathrm{t})$ – contribution to the number of road accidents from the conditions of quarter i, $\mathrm{i} = 1, \dots, 4$; $\mathrm{Prd}_{\mathrm{Pr}}(\mathrm{t})$, $\mathrm{Cvd}_{\mathrm{Cv}}(\mathrm{t})$, $\mathrm{Cpd}_{\mathrm{Cp}}(\mathrm{t})$– contributions, respectively, of holidays, illumination and road conditions to the level of road accidents in the month t, $\mathrm{t} = 1, \dots, 156$.

Data analysis was carried out using the algorithms proposed in this work. The obtained estimates of the coefficients of the model (12) make it possible to solve applied problems of forecasting a selected class of accidents in the study area and assessing the influence of individual factors on their occurrence.

Below is a Table of the causal analysis of road accidents in the designated area by quarters 2017.

The base number of road accidents in each quarter was determined taking into account the quantitative variables in the model (12) and the base (zero) values of the qualitative risk factors. The contributions of qualitative factors are highlighted in separate lines: road accidents on holidays and at night.

Table 1. The results of the quarterly analysis of the factors of road traffic accidents for 2017.

Characteristics of road accident conditions in 2017	***Pr***	***Cv***	***Cp***	**The fact of the accident**	**Accident assessment**
Base number of road accident per quarter 1	-	Daylight	3		2.27
Increase in road accidents on rest days	1	Daylight	Dry		0.45
Increase in road accidents with Darkness or lights lit	-	1	Dry		0.27
Total for quarter 1 in 2017				**4**	**2.99**
Base number of road accident per quarter 2	-	Daylight	5		3.57
Increase in road accidents on rest days	3	Daylight	Dry		1.36
Increase in road accidents with Darkness or lights lit	-	2	Dry		0.53
Total for quarter 2 in 2017				**6**	**5.46**
Base number of road accident per quarter 3	-	Daylight	5		3.79
Increase in road accidents on rest days	3	Daylight	Dry		1.36
Increase in road accidents with Darkness or lights lit	-	5	Dry		1.33
Total for quarter 3 in 2017				**6**	**6.48**
Base number of road accident per quarter 4	-	Daylight	2		3.40
Increase in road accidents on rest days	3	Daylight	Dry		1.36
Increase in road accidents with Darkness or lights lit	-	1	Dry		0.27
Total for quarter 4 in 2017				**5**	**5.03**

The quality of analytical research can be assessed by the following indicator. The actual number of road accidents in 2017 in the allocated area is 21. An explanation of the causes of road accidents in this area is given – 19.96, i.e. the analysis error was 4.94%.

4 Discussion

The main result of this article is to substantiate the possibility of using applied interval analysis for big data. The main idea of the work is based on the fact that the mathematical problems of this analysis (estimating the parameters of the required dependence, identifying and eliminating outliers of observation results, interval forecasting of the resulting variable's values, etc.) can be solved by optimization methods. This property is distinguished by the interval approach and the regression analysis based on the method of least squares. In particular, simple schemes for separating data blocks and using average results (as in [2]) do not work in interval analysis. However, optimization methods of large dimensions become effective for interval analysis of big data [14, 15].

In section 2 of the article, using the example of linear processes, applied problems of interval analysis are systematized and the universality of the application of optimization methods for their solution is shown. This conclusion is also valid in the general case of modeling nonlinear processes. In the methodological section, two implementations of the method of relaxation of constraints are proposed [15] in the general case of measurement errors for all variables, which are directly generalized for the case of analysis of nonlinear processes. These studies have elements of scientific novelty.

Section 3 presents new results of the study of interval analysis of big data of model and real processes. On specific examples of big data analysis in the Excel environment, the organization of distributed computing is considered, an estimate of the final convergence rate is given, and empirical estimates of the error of linear calculations are obtained. It is shown that in practice the number of rows of the observation matrix is not limited, and the number of columns (the number of factors) is determined by the capabilities of the computer program for solving optimization problems. The ways of reducing the factor space of the modeled process are investigated. On the example of the analysis of big data of road traffic accidents (RTA) in England, all stages of interval data analysis are considered, including the tasks of structuring the information flow of data in order to obtain a digital table of input and output variables of the modeled process. Thus, the results of the work can be used for interval analysis of big data of similar applications.

5 Conclusions

The article proposes using an interval approach to solve the problem of big data analysis, in which LP methods are used in the study of dependences linear in parameters. Two methods of sequential reading of constraints and parallel computer computations are proposed for solving LP problems of large dimension. The optimality of the calculations is substantiated using the well-known technique of relaxation of constraints when solving optimization problems of large dimensions.

Computational experiments have shown the possibility of using applied interval analysis in practice. The research was carried out for model processes and for real data.

References

1. Müller, A., Guido, S.: Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly Media, Newton, MA, USA (2016).
2. Fan, T.-H, Lin, D.K.J., Cheng, K.-F.: Regression analysis for massive datasets. Data & Knowledge Engineering **61**(3), 554-562 (2007). https://doi.org/10.1016/j.datak.2006.06.017.
3. Adjout, M. R., Boufares, F. A.: Massively parallel processing for the multiple linear regression. In: Proceedings of International IEEE Conference on Signal-Image Technologies and Internet-Based System, 27-27 November 2014, Marrakech, Morocco, pp. 666-671 (2015). https://doi.org/10.1109/SITIS.2014.26.
4. Frank, A., Fabregat-Traver, D., Bientinesi, P.: Large-scale linear regression: development of high-performance routines. Applied Mathematics and Computation **275**, 411-421 (2016). https://doi.org/10.1016/j.amc.2015.11.078.
5. Khine, K.L.L., Nyunt, T.T.S. (2019): Predictive big data analytics using multiple linear regression model. Advances in Intelligent Systems and Computing, **744**, 9-19. https://doi.org/10.1007/978-981-13-0869-7_2.
6. Wang, K., Li, S.: Robust distributed modal regression for massive data. Computational Statistics & Data Analysis **160** 107225 (2021). https://doi.org/10.1016/j.csda.2021.107225.
7. Kantorovič, L. V. Towards novel approaches to computational methods and the processing of observed phenomena (O nekotoryh novyh podhodah k vychislitel'nym metodam i obrabotke nabljudenij). Siberian Mathematical Journal **3** 701–709 (1962) (in Russian).
8. Milanese, M., Norton, J., Piet-Lahanier, H., Walter, E. (eds.): Bounding Approaches to System Identification. Plenum Press, New York, NY, USA (1996). https://doi.org/10.1007/978-1-4757-9545-5.
9. Jaulin, L., Kieffer, M., Didrit, O., Walter, E.: Applied Interval Analysis. Springer-Verlag, London (2001).
10. Zhilin, S.I.: On fitting empirical data under interval error. Reliable Computing **11**, 433-442 (2005). https://doi.org/10.1007/s11155-005-0050-3.
11. Gutowski, M.W.: Interval experimental data fitting. In: Liu, J.P. (ed.): Focus on Numerical Analysis, pp. 27-70. Nova Science, New York, NY, USA (2006). https://doi.org/10.13140/2.1.5156.3520.
12. Shary, S.P.: Maximum consistency method for data fitting under interval uncertainty. Journal of Global Optimization, **66**(1), 111-126 (2016). https://doi.org/10.1007/s10898-015-0340-1
13. Shary, S.P.: Weak and strong compatibility in data fitting problems under interval uncertainty. Advances in Data Science and Adaptive Analysis, **12**(1), 2050002 (2020). https://doi.org/10.1142/S2424922X20500023.
14. Lasdon, L.S.: Optimization Theory of Large Systems. Collier-Macmillan, London (1970).
15. Geoffrion, A.: Reducing concave programs with some linear constraints. SIAM J. Appl. Math. **15**, 653-664 (1967).
16. Tsiaras, Th.: UK Road Safety: Traffic Accidents and Vehicles, Detailed dataset of road accidents and involved vehicles in the UK (2005-2017) by Kaggle, ver. 3. https://www.kaggle.com/tsiaras/uk-road-safety-accidents-and-vehicles, last accessed 12 February 2021.

17. UK Department for Transport: Road Safety Data. https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data, last accessed 12 February 2021.
18. Oskorbin, N.M.: Computational technologies for the synthesis of decentralized control systems for multistage technological processes. Journal of Physics: Conference Series **1615**, 012020 (2020). https://doi.org/10.1088/1742-6596/1615/1/012020.