ISSN 1407-5806

COMPUTER MODELLING AND NEW TECHNOLOGIES

Volume 11
No 1

2007

Transporta un sakaru institūts
(Transport and Telecommunication Institute)

Computer Modelling and New Technologies

Volume 11, No. 1 – 2007

ISSN 1407-5806
ISSN 1407-5814
(On-line: www.tsi.lv)

Riga – 2007

THE JOURNAL IS DESIGNED FOR PUBLISHING PAPERS CONCERNING THE FOLLOWING FIELDS OF RESEARCH:

- mathematical and computer modelling
- mathematical methods in natural and engineering sciences
- physical and technical sciences
- computer sciences and technologies
- semiconductor electronics and semiconductor technologies
- aviation and aerospace technologies
- electronics and telecommunication
- navigation and radar systems
- telematics and information technologies
- transport and logistics
- economics and management
- social sciences

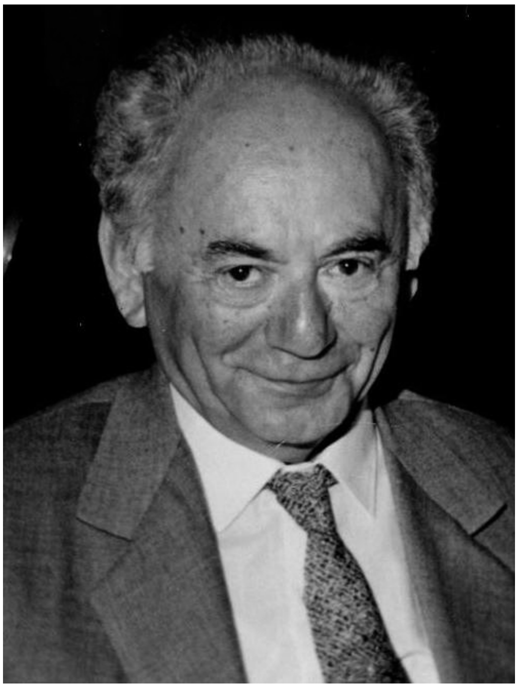
Articles can be presented in journal in English. All articles are reviewed.

EDITORIAL CORRESPONDENCE
Transporta un sakaru institūts (Transport and Telecommunication Institute)
Lomonosova 1, Riga, LV-1019, Latvia. Phone: (+371)7100593. Fax: (+371)7100535.
E-mail: journal@tsi.lv, http:// www.tsi.lv

COMPUTER MODELLING AND NEW TECHNOLOGIES, 2007, Vol. 11, No. 1
ISSN 1407-5806, **ISSN** 1407-5814 (on-line: www.tsi.lv)
Scientific and research journal of Transport and Telecommunication Institute (Riga, Latvia)
The journal is being published since 1996.

CONTENTS

Editors' Remarks



Khaim B. Kordonsky (1919-1999)

This volume of journal comprises the papers selected at the *First International Workshop "Modern Statistical Methods and Models"*, which was hold in Riga on October 25-28, 2006. This Workshop has been affiliated in the program of the *6th International Conference "Reliability and Statistics in Transportation and Communication"* devoted to memory of an outstanding mathematician and pedagogue Khaim B. Kordonsky (1919-1999).

Professor Khaim B. Kordonsky has made significant contribution in the development and application of probabilistic and statistical methods. In the Soviet Union his monograph [1] is the first book considering probability methods, which has been addressed to engineers. Khaim Kordonsky devoted his life in advancing not only the science of reliability and mathematical statistics, but also the application to the solution of important problems in safety and operational efficiency. Main fields of his research activity are as follows: statistical quality control, statistical fatigue theory, theory of accuracy of machines, statistical reliability theory, probability methods in airline scheduling, statistical medical research. Last two fields will be considered further on.

He was the research supervisor of computer systems development for the Soviet company "Aeroflot" which in the 20th century was the biggest aviation company in the world. Under the direction of professor Kh. Kordonsky Aeroflot Computer System of Central Airline Scheduling was created [4]. Amount the main results of Kh. Kordonsky's research in medicine we can point out statistical analysis of the leucosis cattle diseases rate in Latvia and the sinus rhythm mathematical model creation.

Almost all papers of the given volume are prepared by Khaim B. Kordonsky's pupils or pupils of his pupils. The volume consists of two parts: *"Probabilistic Models"* and *"Statistical Inferences"*.

The first part of issue begins with the paper by *Yu. Paramonov* and *J. Andersons*. It is purposed to reliability calculation of the system with complex structure, which is the main subject matter in Kordonsky's works. Appropriate models and methods have continuation in the tasks of Inventory Control and Financial Risks. These tasks are considered correspondingly in the papers by *E. Kopytov, L. Gringlaz, A. Muravjov*, *E. Puzinkevich* and *A. Sverchenkov*. The interesting probabilistic treatment of the task of Spatial Arrangement of Service Stations is presented in the paper by *A. Andronov* and *A. Kashurin*.

In the second part of the journal tasks purposed to statistical problems are gathered. As Kh. Kordonsky persistently has mentioned, probabilistic models are not of great importance, if good statistical estimates of them are absent. Therefore Khaim B. Kordonsky loved mathematical statistics very much. Many statistical tasks, which he has solved in order to apply them in aviation [2, 3, 5-8], subsequently have become popular in different mathematical and statistical researches, for instance, Censored Samples, the Best Time Scale for Reliability, etc.

The second part of the issue begins with the paper by *Michael S. Tikhov, Dmitriy S. Krishtopenko* and *Marina V. Yarochuk*, which presents profound theoretical results on the kernel estimator of unknown distribution function on the base of indirect observations. The last is some generalization of the censored sample. Application of the Resampling approach in regression tasks is considered in the paper by *A. Andronov*. It should be noted that at present Resampling approach is one of the most popular intensive computer statistical methods. Other modern methods of regressions estimation on the base of the Generalized Linear Model and the Single Index Model are considered in the papers by *Catherine Zhukovskaya and Diana Santalova*. Numerical examples examined in these papers show the obvious preference of the proposed approaches, in other words the suggested approaches have given better results than the classical approach in cases of smoothing and forecasting.

We hope that the presented issue of the journal will display to reader an idea of modern tendencies of the development and practical application of probabilistic and statistical methods. We also invite readership to take part in the next Workshops and in similar issues.

Selected Bibliography

1. Kordonsky Kh. B. *Applications of Probability Theory in Engineering*. Moscow-Leningrad: Physics & Math. Literature Publishing, 1963. (In Russian)
2. Gertsbakh I. and Kordonsky Kh. *Models of Failures.* Berlin-Heidelberg-New York: Springer, 1969.
3. Artamanovsky A. V. and Kordonsky Kh. B. Maximum Likelihood Estimation for Simple Data Grouping, *Probability Theory and it Application,* **1**, 1970. (In Russian)
4. Kordonsky Kh., Venyavcev V. and others. Central Airline Scheduling as Part of Air Traffic Control. In: *Proceedings of the 1st International Symposium on Traffic Control*. Versailles, 1970.
5. Kordonsky Kh., Gertsbakh I. Choice of the Best Time Scale for Reliability Analysis, *Europ. J.Operat. Res*., **65,** 1993.
6. Kordonsky Kh., Gertsbakh I. Best Time Scale for Age Replacement, *Inter. J. of Reliab., Quality and Safety Engineering*, **1**, 1994.
7. Kordonsky Kh., Gertsbakh I. System State Monitoring and Lifetime Scales. I, II, *Reliab*. *Engineering and System Safety,* **47**, **49**, 1994.
8. Kordonsky Kh., Gertsbakh I. Multiple Time Scales and the Lifetime Coefficient of Variation: Engineering Applications, *Lifetime Data Analysis,* **3**, 1997.

July 2, 2007

EDITORS

A.M. Andronov

E.A. Kopytov

I.V. Kabashkin

PART I

PROBOBABILYSTIC MODELS

Computer Modelling and New Technologies, 2007, Vol.11, No.1, 8-20
Transport and Telecommunication Institute, Lomonosova 1, Riga, LV-1019, Latvia

ANALYSIS OF FIBER STRENGTH DEPENDENCE ON LENGTH USING AN EXTENDED WEAKEST-LINK DISTRIBUTION FAMILY

Yury Paramonov[1], Janis Andersons[2]

[1]Aviation Institute, Riga Technical University
Lomonosova Str. 1, Riga, LV-1019, Latvia
Tel.: +371 7255394; Fax: +371 7089990; E-mail rauprm@junik.lv

[2]Institute of Polymer Mechanics, University of Latvia
Aizkraukles Str. 23, Riga, LV-1006, Latvia
Tel.: +371 7255394; Fax: +371 7089990; E-mail rauprm@junik.lv

An extended family of the weakest-link models based on the assumption of a two-stage failure process of a fiber specimen was developed in [1, 2]. A generalization of this family is presented in this paper. As in [1, 2] we consider the specimen as a chain of *n* elements (links). The fracture4 process is modelled as follows: in the first stage initiation of defects (before loading or during loading), and in the second stage a specimen fracture takes place. As distinct from our previous publications, the strength of items without defects is taken into account and two types of the influence of defect number on the specimen strength are considered. The comparison of the models and the choice of the best one are made using cross validation method. The offered models sometimes describe more adequately the experimentally observed fiber strength scatter and the strength dependence on fiber length than the traditional models do.

Keywords: *distribution function, composite, static strength*

1. Introduction

The significant dependence of static strength of a composite on the scatter of static strength of its components can be illustrated by the following example. Let us consider three parallel items with 10 N, 15 N and 30 N strength and identical stiffness. It may seem surprising that they will fail at the applied load of 30 N, as if the strength of every item is equal to 10 N. Why?

The reason is that under 30 N load at first the weakest item will fail because its strength is equal to 10 N. At the uniform distribution of total loads, its load is equal to 10 N also. Now the load acting on each "surviving" item is equal to 15 N. So the second item, the strength of which is equal to the same value of 15 N, will fail. Now the load for the last strongest item will be equal to 30 N. It will fail also because its strength is just equal to this load. This process ("domino phenomenon") is shown on Fig.1. The same phenomenon takes place if element strengths are proportional to the terms of harmonic series: 1, ½, 1/3, …, 1/n, see Fig. 2.

So we see that the composite strength dependence on the strength scatter of its constituents can be very significant.

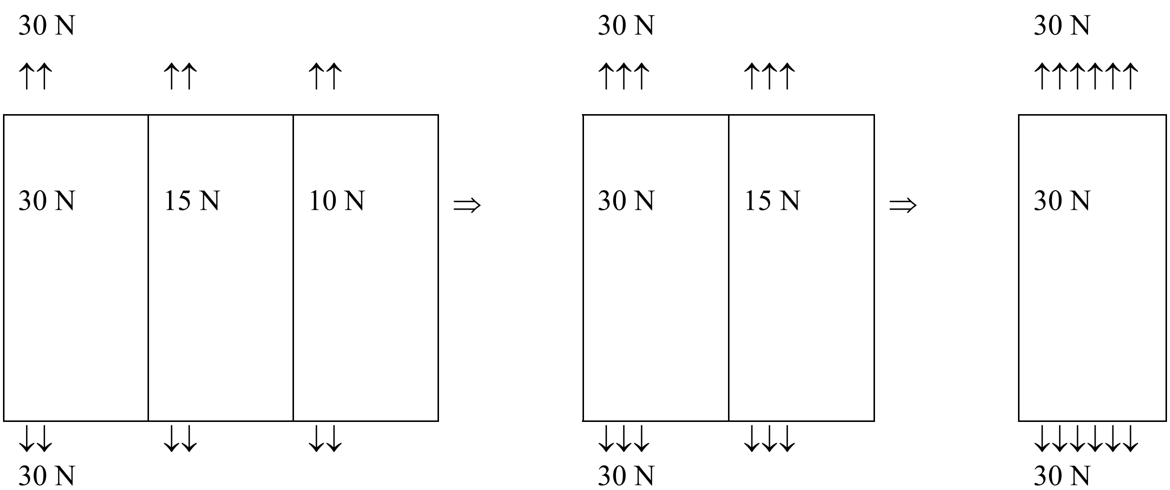


Figure 1

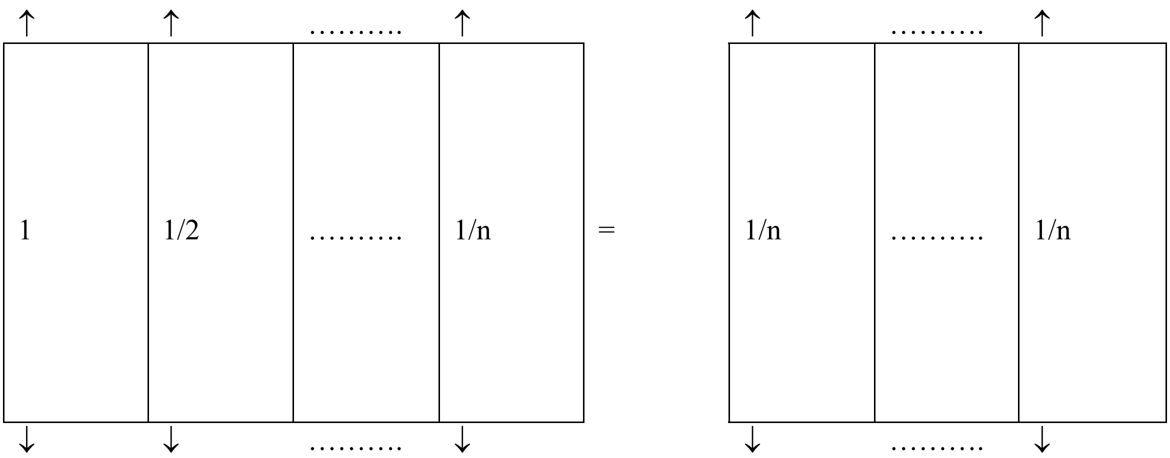


Figure 2

Power-Weibull (PW) model of distribution [3, 4, 5]

$$F(s) = 1 - \exp(-(L/l_1)^{\gamma}(s/\beta_1)^{\alpha}), \qquad (1)$$

which has been intensively studied in literature, while providing a good empirical fit to the strength data of specimens with different length, L, lacks the theoretical appeal of the weakest-link models (It should be noted that here parameter β_1 corresponds to $L = l_1$, β_1 changes if l_1 changes.). We derive a new weakest-link model family (WLMF) based on the assumption of a two-stage failure process. For modelling purposes we consider a specimen (fiber) as a chain of n elements (links) of length l_1. First, the process develops along the specimen and defects appear in K elements. Here K is integer random variable, $0 \le K \le n$. Two types of the second stage will be considered in this paper. First type: in every element (containing defects or intact) the development of fracture process takes place and the strength of the weakest item (link) defines the strength of the specimen. Second type: development of fracture process takes place only in one, critical element. Then only the probability that the second stage will take place depends on the number of elements but the strength distribution of this element (the process of accumulation of elementary damages in crosswise direction up to specimen failure) does not depend on this number.

We consider two different versions of the first stage also. First version: defects appear before the loading and their number does not depend on the subsequent loading. Second version: defects appear during loading (instantly or gradually) and their number depends on the load.

2. General Description of the Model Family

2.1. The Fracture Process Takes Place in Every Element

2.1.1. Models of instant fracture

Let K, $0 \le K \le n$ be the number of elements in which defects appear. Let $Y_1, Y_2, ..., Y_K$ be independent random variables which are the strengths of these elements with the same cumulative distribution function (cdf) $F_Y(x)$; $Z_1, Z_2, ..., Z_{n-K}$, $F_Z(x)$ are the same for the elements without defects. It seems reasonable to assume that the random strength of the specimen is the strength of the weakest item

$$X = \min(Y_1, ..., Y_K, Z_1, ..., Z_{n-K}), \qquad (2)$$

with the corresponding cdf

$$F(x) = 1 - (1 - F_{Z1,n}(x)) \sum_{k=0}^{n} p_k \delta^k , \qquad (3)$$

where

$$\delta(x) = (1 - F_Y(x))/(1 - F_Z(x)), \tag{4}$$

$$F_{Z_{1,n}}(x) = 1 - (1 - F_Z(x))^n. \tag{5}$$

Several different assumptions can be made here. Let us consider first the case of defects appearing before loading. It can be assumed that the probability of defect in one item, p, is some constant (and it is a parameter of the model). Then the corresponding binomial probability mass function (pmf) is

$$p_k = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}. \tag{6}$$

If $\lambda = np$ is large enough we can use (as an approximation) Poisson pmf:

$$p_k = \exp(-\lambda)\lambda^k / k!. \tag{7}$$

In this case the equation (3) (approximately) can be written in the following way:

$$F(x) = 1 - (1 - F_{Z_{1,n}}(x))\exp(-\lambda(1 - \delta(x)). \tag{8}$$

If initiation of the defects depends on the applied load then it can be assumed that $p = F_0(x)$, where $F_0(x)$ is the cdf of defect initiation stress.

2.1.2. Models of gradual accumulation of defects

We consider the process of accumulation of defects as an inhomogeneous finite Markov's chain (MC) with finite state space $I = \{i_1, i_2, ..., i_{n+1}, i_{n+2}\}$. MC is in state i_k if there are $(k-1)$ defects, $k = 1, ..., n + 1$. State i_{n+2} is an absorbing state corresponding to the fracture of specimen. Usually we suppose that the Markov's chain starts in state i_1 but in general case the initial distribution is represented by a row vector π given by $\pi = (\pi_1, \pi_2, ..., \pi_{n+1,} \pi_{n+2})$. We further assume that the loading (i.e. the process of nominal stress increase in the specimen cross section) is described by an ascending (up to infinity) sequence $\{x_1, x_2, ..., x_t, ...\}$ and the process of MC state change is described by the transition probabilities matrix

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{23} & \cdots & p_{1(n+1)} & p_{1(n+2)} \\ 0 & p_{22} & p_{23} & p_{24} & \cdots & p_{2(n+1)} & p_{2(n+2)} \\ 0 & 0 & p_{33} & p_{34} & \cdots & p_{3(n+1)} & p_{3(n+2)} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & p_{(n+1)(n+1)} & p_{(n+1)(n+2)} \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix},$$

which at the tth-step is a function of x_t, $t = 1, 2, ...$ Let the sequence $\{x_t\}$ be fixed, then P is a function of t. Let us note that if $n = \infty$ then the subscript $(n + 2)$ is not a number but only a symbol, corresponding to the absorbing state i_{n+2}.

In the new model the number of defects and the strength of specimens are random functions of time,

$K(t)$, and

$$X(t) = \min(Y_1, Y_2, ..., Y_{K(t)}, Z_1, Z_2, ..., Z_{n-K(t)}) \tag{9}$$

correspondingly. The specimen fracture occurs when the strength of the specimen becomes equal to or less than the current load (stress). Ultimate strength

$$X = x_{T^*}, \tag{10}$$

where

$$T^* = \max(t : X(t) > x_t) . \quad (11)$$

Cdf of *X* is defined by equation

$$F(x_m) = \pi(\prod_{j=1}^{m} P(j))u , \quad (12)$$

where $P(j)$ is the transition matrix for step number *j,* column vector $u = (0,...,0,1)'$ where only the last component is equal to 1 but all the others are equal to 0.

2.1.3. Specifying models. The specimen strength without defect is very large

For the purpose of specification of the models, the general description of which was given in the previous section, we additionally have to specify the cdf $F_Y(x)$, $F_0(x)$ (for models with defect number dependence on load), $F_Z(x)$, and, additionally, for Markov models, a prior distribution, π, which, of course, in general case can differ from binomial or Poisson distribution. For Markov models we need to specify also the matrix *P* as a function of current stress, x_t.

In this paper we assume that $F_Y(x)$ and $F_0(x)$ are the smallest extreme value (sev) distributions. For the case when location parameter $\theta_0 = 0$ and scale parameter $\theta_1 = 1$ it is assumed that

$$F_Y(x) = 1 - \exp(-\exp(x)) , \quad (13)$$

$$F_0(x) = F_Y(x - \delta_0) , \quad (14)$$

where $x = \log(s)$, *s* is the strength (expressed in MPa). If $\delta_0 > 0$ then at the same probability of events the stress required for new defect initiation is larger than the stress required for the failure of an element with defect.

For $F_Z(x)$ we consider two assumptions in this paper. First, sev distribution can be assumed again:

$$F_Z(x) = F_Y(x - \delta_Z) . \quad (15)$$

Again we can say that if $\delta_Z > 0$ then $F_Z(x) < F_Y(x)$.

But the simplest is the assumption that

$$F_Z(x) = \begin{cases} 0, x < C, \\ 1, x \geq C, \end{cases} \quad (16)$$

where C is a very large constant.

Then instead of (2) we have

$$X = \begin{cases} \min(Y_1,....Y_K),\ K > 0, \\ C,\ \ K = 0. \end{cases} \quad (17)$$

The equation (3) can be written in the form

$$F(x) = \begin{cases} 1 - \sum_{k=0}^{n} p_k \delta^k ,\ x < C, \\ 1,\ x \geq C \end{cases} \quad (18)$$

where $\delta = 1 - F_Y(x)$. But equation (8) now has the following form

$$F(x)=\begin{cases}1-\exp(-\lambda F_Y(x)), & x<C,\\ 1, & x\ge C.\end{cases} \tag{19}$$

In [1, 2] was shown that the cdf

$$F(x)=\sum_{k=0}^{\infty} p_k\{1-(1-F_Y(x))^{k+1}\} \tag{20 a}$$

or

$$F(x)=1-(1-F_Y(x))\exp(-\lambda F_Y(x)), \tag{20 b}$$

where p_k is defined by (7) , $\lambda = np,$ $p = F_Y(x)$, $F_Y(x)$ is sev cdf, provides a good empirical fit to the strength data of specimens with different length, *L*. Equation (20 b) can be considered as modification of (8): $F_Y(x)$ is used here instead of $F_{Z_{1,n}}(x)$. But now it is not only an approximation of the "binomial" model. Now we can consider the specimen as continuous one and define λ by equation

$$\lambda=\lambda_1(L/l_1),$$

where *L* is the specimen length, λ_1 is the intensity of defects (the defect number per length l_1). Then function $F_Y(x)$ can be regarded as an element-length-independent cdf of strength distribution in the cross section with defect, where the number of defective cross sections has the corresponding Poisson distribution.

For Markov models we should specify the matrix *P*. In the case when parameter *C* is very large (the theoretical strength is much higher than the real strength) the probability that in some element the defect appears at the stress x_t under the condition that it has not appeared at the stress x_{t-1} is

$$b(t)=(F_0(x_t)-F_0(x_{(t-1)}))/(1-F_0(x_{(t-1)})).$$

Consider the case of *s* defects present. The probability that *r* new defects appear, $0\le r\le k=n-s$, and the total number of defects is equal to $m = s + r$

$$\widetilde{p}_{sm}(t)=(b(t))^r(1-b(t))^{k-r}k!/r!(k-r)!$$

Conditional probability of element fracture at the nominal stress x_t

$$q(t)=(F_Y(x_t)-F_Y(x_{(t-1)}))/(1-F_Y(x_{(t-1)})).$$

Corresponding probability that none of the elements fail when there are defects in *m* elements is

$$u_m(t)=(1-q(t))^m.$$

The probability of coincidence of these events, which we consider as independent, is the probability of transition from state $i = s + 1$ to state $j = i + r$

$$p_{ij}(t)=\widetilde{p}_{(i-1)(j-1)}(t)u_{j-1}(t),$$

where $i\le j\le (n+1)$.

Conditional fracture probability at state *i*

$$p_{i(n+2)}(t)=1-\sum_{j=i}^{n+1}p_{ij}(t).$$

Of course, $p_{ij}(t)=0$, if $j<i$, and $p_{(n+2)(n+2)}(t)=1$.

2.2. The Fracture Process Takes Place Only in One Element

2.2.1. The models of instantaneous failure

In the previous models it is assumed that defects are uniformly distributed along the specimen length. But it is plausible that such uniformity is retained only at the initial stage of loading. More precisely, upon formation of the weakest link in a chain, the development of failure proceeds only in this link, and the specimen length is of no importance any more. The simplest variant of such a model corresponds to the assumption that the law of strength distribution in the element where this process proceeds (in the cross section where the critical defect is formed) is independent of specimen length, which determines only the probability of formation of an element with defect. The mathematical formulation of this hypothesis is as follows

$$X = \begin{cases} Y, K > 0, \\ Z, K = 0. \end{cases} \tag{21}$$

Here, Y and Z are random variables, which are the strength of element where the failure process proceeds with or without defect, correspondingly.

In this case

$$F(x) = \{1 - (1 - F_0(x))^n\} F_Y(x) + (1 - F_0(x))^n F_Z(x) . \tag{22}$$

2.2.2. Model of successive formation of at least one defect

The corresponding Markov's chain has only three states. The first state corresponds to the absence of defective elements, the second one means the presence of at least one defective element, and the third, absorbing one, means failure of the specimen. The corresponding probabilities at a *t*th step are determined by the formulae

$$p_{11}(t) = [1 - b(t)]^n, \quad p_{12}(t) = (1 - p_{11}(t))(1 - q(t)), \quad p_{13}(t) = (1 - p_{11}(t))q(t),$$

$$p_{21}(t) = 0, \quad p_{22}(t) = 1 - q(t), \quad p_{23}(t) = p_{32}(t) = 0, \quad p_{33}(t) = 1 .$$

Specification of the cdf and of elements of the matrix P can be made in the same manner as in section 2.1.3.

3. The Processing of Test Data

The maximum likelihood method can be used for parameter estimation but it is excessively labor-consuming. The estimates of parameters θ_0 and θ_1 (at fixed other parameters) can be found easily using regression analysis of order statistics. Our purpose here is only the investigation of the possibility of using the considered models for prediction of fiber strength distribution changes when the fiber length is varied and the comparison of the models has been done as well. So we have limited ourselves by the use of regression analysis.

Let x_{ij} be *j*th order statistic, $j = 1,2,...,n_i$, n_i is the number of specimens with $L = L_i$, $i = 1,2,...,k_L$, k_L is number of different L_i, $E(X_{ij})$ is the expected value of random order statistic X_{ij}, $E(\overset{0}{X}_{ij})$ is the same but for $\theta_0 = 0$ and $\theta_1 = 1$.

Then we have the following linear regression model

$$E(X_{ij}) = \theta_0 + \theta_1 \, E(\overset{0}{X}_{ij}), \tag{23}$$

where $E(\overset{0}{X}_{ij})$ is a function of L_i, n_i and *j*.

This equation can be used for estimation of θ_0 and θ_1 if all the other parameters are fixed.

We compare the above-mentioned models with the PW model (see equation (1)) and LW model (it is the original Weibull model: PW model with $\gamma = 1$). If *S* is random strength of specimen with cdf (1) then for $X = \log(S)$,

$$F_X(x) = 1 - \exp(-\exp((x - \theta_0)/\theta_1), \tag{24}$$

where

$\theta_0 = \log(\beta_1) - (\gamma/\alpha)\log(L/l_1)$, $\theta_1 = 1/\alpha$.

So for PW model we have equation with three unknown parameters $\theta_{00} = \log(\beta_1)$, $\theta_{01} = -\gamma/\alpha$ and θ_1,

$$E(X_{ij}) = \theta_{00} + \theta_{01} \log(L_i/l_1) + \theta_1 E(\overset{0}{X}_{ij}). \tag{25}$$

For LW model we have an equation with two unknown parameters θ_0 and θ_1

$$E(X_{ij}) = \theta_0 + \theta_1 (-\log(L_i/l_1) + E(\overset{0}{X}_{ij})). \tag{26}$$

In (25) and (26) the value of $E(\overset{0}{X}_{ij})$ is the expected value of *j*th order statistic for sample from sev distribution with sample size n_i .

It is assumed that roughly $E(\overset{0}{X}_{ij}) = \overset{0}{F}{}^{-1}(\hat{F}(x_{ij}))$, where $\hat{F}(x_{ij}) = (j - 0.3)/(k_L + 0.4)$ is an estimate of $F(x_{ij})$.

For comparison of different models, the glass fiber dataset described in [1,2] is used (four samples with specimen lengths (L_1, L_2, L_3, L_4)=(10, 20, 40,80 mm), sample sizes (n_1, n_2, n_3, n_4)=(78,74,50,60)). For parameter estimation a version of the cross validation method is applied. At the fixed nonlinear parameters ($l_1, \ldots$) for the linear regression (LR) estimation of parameters θ_0 and θ_1 we use only the dataset corresponding to *L*= 10 mm and *L*=20 mm. We calculate also two additional statistics

$$Q_1 = (\sum_{i=1}^{k_L} (\bar{x}_i - \hat{x}_i)^2 / \sum_{i=1}^{k_L} (\bar{x}_i - \bar{x})^2)^{1/2}, \tag{27}$$

where $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$; $\hat{x}_i = \sum_{i=1}^{n_i} \hat{x}_{ij}/n_i$; $\hat{x}_{ij} = \hat{\theta}_0 + \hat{\theta}_1 E(\overset{o}{X}_{i.j})$; $\hat{\theta}_0$ and $\hat{\theta}_1$ LR estimates of θ_0 and θ_1, $\bar{x} = \sum_{i=1}^{k_L} \bar{x}_i / k_L$,

and

$$\overline{R}_{LR} = (1 - R^2)^{1/2}, \tag{28}$$

where R^2 is standard statistic of LR analysis (the coefficient of determination).

As nonlinear parameter estimates, the values of the parameters which correspond to the minimum of statistics OSPPt (Order Statistics Probability Plot Test) are taken. OSPPt is the measure of the error of order statistics prediction for sample with L_4= 80 mm:

$$\text{OSPPt} = (\sum_{j=1}^{n4} (x_{4j} - \hat{x}_{4j})^2 / \sum_{j=1}^{n4} (x_{4j} - \bar{x}_4)^2)^{1/2} . \tag{29}$$

For the convenience of the following references let us list the full number of specifications and assumptions which define the specific model in the considered family and make specific notations of the corresponding assumptions.

We have to specify the conditions under which the initiation of defects takes place. By symbol '**T**' we denote the assumption that the process of initiation of defects is a function of **technology** only, but symbol '**L**' is used if this initiation depends on **load**.

A **prior distribution** of defects needs to be specified for the models in which Markov's chains theory is used. In general case we denote a prior distribution by π but we use symbols '**B**' or '**P**' if binomial or Poisson distribution is used.

If we consider the **instantaneous fracture** of specimen we use symbol '**B**' for binomial distribution of defect number, *K*, symbol '**P**' for Poisson distribution and symbol **'Pm'** for 'truncated' Poisson distribution.

(Remark. We use the words 'truncated in m (discrete) distribution' if instead of discrete rv *X* we consider the rv

$$Xm = \begin{cases} X, \text{ if } X < m+1, \\ m+1, \text{if } X > m. \end{cases}$$

The use of it can be convenient for calculation of the cdf of steps to absorption using formulae of finite Markov's chains theory).

We use simbols '**MB**', '**MBm**' (for truncated binomial distribution), '**MP**' and '**MPm**' (for truncated Poisson distribution) if the Markov's chain is used for description of defect initiation process (Note. Formulae for transition probability matrix in this section are given only for MB case).

F_Z , F_Y and F_0 have to be specified:

- the cdf of strength of elements without defects, F_Z ;
- the cdf of strength of elements with defects, F_Y ;
- the cdf of defect initiation stress F_0 (if the process of defect initiation is assumed to be a function of load). In general case we use symbols F_Z , F_Y and F_0 correspondingly but they should be specified by specific equations or by specific definitions. In this paper (Fig. 3-9) we use symbol **S** if cdf is defined by equation (13), symbol **St** if cdf is defined by equation (14); symbol **Zt**, if cdf is defined by equation (15); symbol **C**, if cdf is defined by equation (16).

If the Markov's chain is used then the sequence of loads (stresses) $\{x_t\}$ should be specified also, but in this paper, as a rule, $\{x_t\}$ is a sequence of numbers uniformly distributed in some interval, which can be seen on Figures with $f(x)$ and $F(x)$ (see Fig. 3, …).

We consider six models in total, but already preliminary investigation shows that the first two (T.B.Zt.S and T.P.C.S) are not appropriate for fiber strength distribution description although it seems that both are very natural. These models correspond to assumptions that there is binomial or Poisson distribution of technological defects which can appear during preparation of specimens. We show this by presenting some examples of calculations.

The model T.B.Zt.S corresponds to assumption that during production of fiber specimens in every element (with length l_1) of specimen one defect can appear with probability *p*. (Here and later on we presume that the ratio L_i/l_1 is integer and it is equal to the number of elements in specimen with length L_i (for every i = 1, …, 4)). The results of calculations of $f(x)$, $F(x)$ for θ_0 = 0 and θ_1 = 1, estimates of order statistics, $\hat{x}_{4j}$, as function of x_{ij}, estimates of mean, $\hat{x}_i$, as function of L_i (using LR estimates of θ_0 and θ_1) are shown on Fig. 3, which corresponds to l_1 = 10 mm (it is the length of the shortest specimen), p = 0.5, δ_Z = 7, θ_0 = 7.5326, θ_1 = 0.0562. We see that although the estimates of mean, $\hat{x}_i$, are acceptable, the estimates of $\hat{x}_{4j}$ are less than satisfactory.

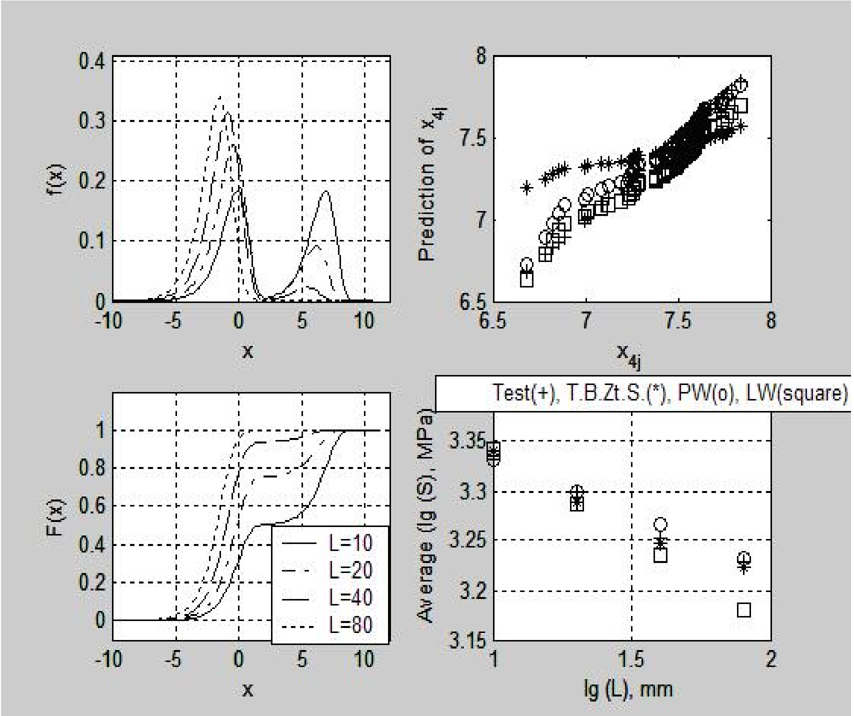


Figure 3

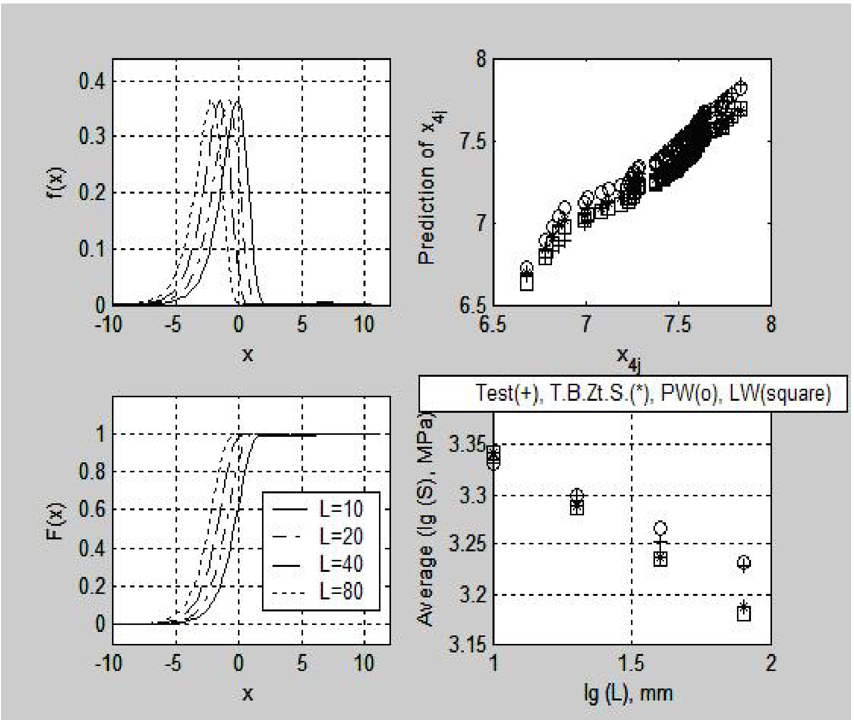


Figure 4

But if better agreement of the estimates $\hat{x}_{4j}$ is achieved (using parameters $p = 0.99$ (for $l_1 = 10$ mm), $\delta_Z = 7$, $\theta_0 = 7.7944$, $\theta_1 = 0.1660$), then the estimates $\hat{x}_i$ deteriorate considerably (see Fig. 4).

In accordance with the model T.P.C.S it is assumed that the number of defects in specimen with length L_i has Poisson distribution with parameter $\lambda_1 L_i / l_1$. Cdf $F(x)$ is defined by (19). Results of calculation for $\lambda_1 = 0.65$ (for $l_1 = 10$ mm), $\theta_0 = 7.5107$, $\theta_1 = 0.0389$ and $C = 10$ are shown on Fig. 5.

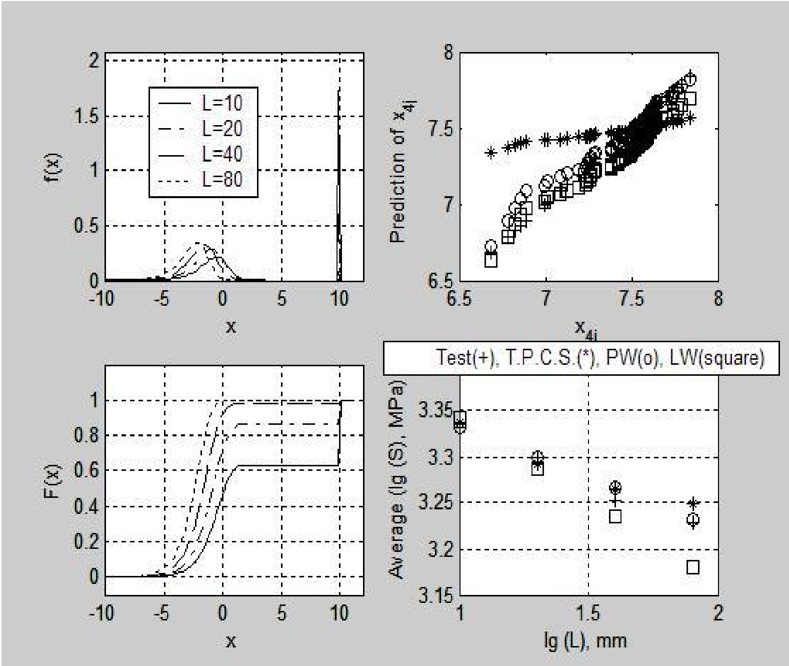


Figure 5

Again we see reasonable estimation of mean , $\hat{x}_i$, but the estimates $\hat{x}_{4j}$ are less than satisfactory. And just as in the previous model, we can improve $\hat{x}_{4j}$, but then the estimates of mean, $\hat{x}_i$, deteriorate.

More detailed search of parameter estimates was made for four models.

For the Model Lmod.P.C.S.S. (see equation (20)), which in [2] was denoted by p-sev-sev), the following parameter estimates were determined (for $C=\infty$): $\hat{\lambda}$ = 1.1 (for l_1 = 1 mm), $\hat{\theta}_0$ = 8.1406, $\hat{\theta}_1$ = 0.2743. These estimates correspond to the minimum of $\overline{R}_{LR}$. Estimates $\hat{x}_{4j}$ and $\hat{x}_i$ are shown on Fig. 6. Although for this model the values of $\overline{R}_{LR}$, Q_1 are better than for PW and LW model (see Table 1), the statistics OSPPt for prediction for L_4 = 80 mm is better than for LW but worse than for PW.

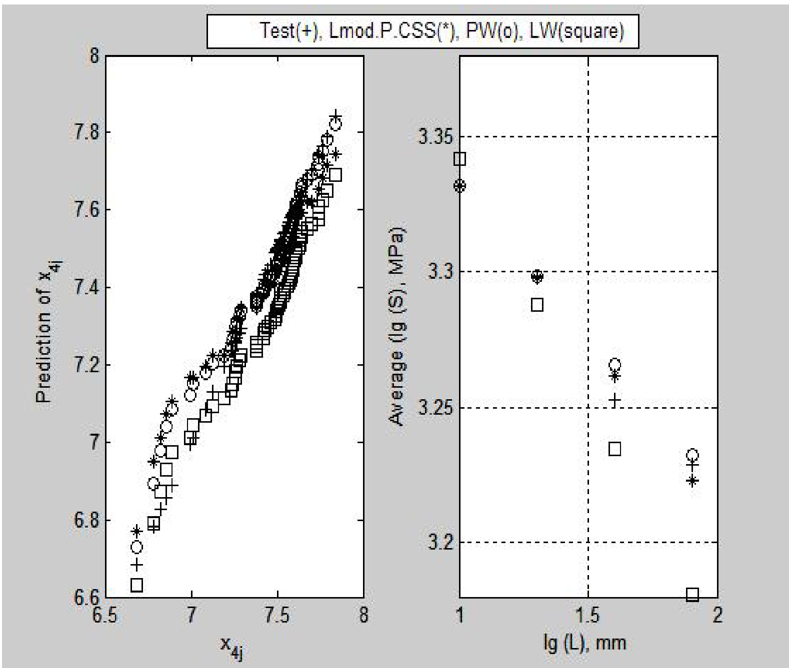


Figure 6

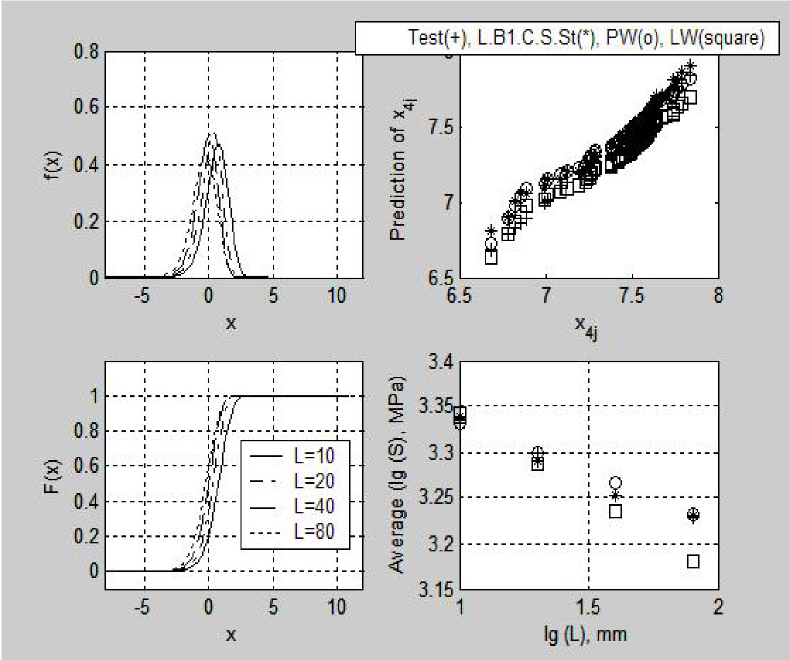


Figure 7

The Model L.B1.C.S.St (which in [1] for $\delta_0 = 0$ is denoted by D1) corresponds to equations (21) for $F(x)$ and (16) for $F_Z(x)$ with C = ∞. For parameter estimates: p = 0.9 (for l_1 = 10), θ_0 = 7.5398, θ_1 = 0.2605, δ_0 = 0.9, corresponding results (which are very close to the results of Lmod.P.C.S.S.) can be seen in Fig. 7 and Table 1.

The same can be said concerning the Model L.π.MB.C.S.S (see equation (12) for $F(x)$ and (16) with C = ∞ for $F_Z(x)$), which was denoted by MB in [1]. The estimates of the model parameters are: l_1 = 5mm, θ_0 = 7.7578, θ_1 = 0.236, π = (0,1,0,…,0)). The corresponding results (which are very close to the results of Lmod.P.C.S.S.) can be seen in Fig. 8 and Table 1.

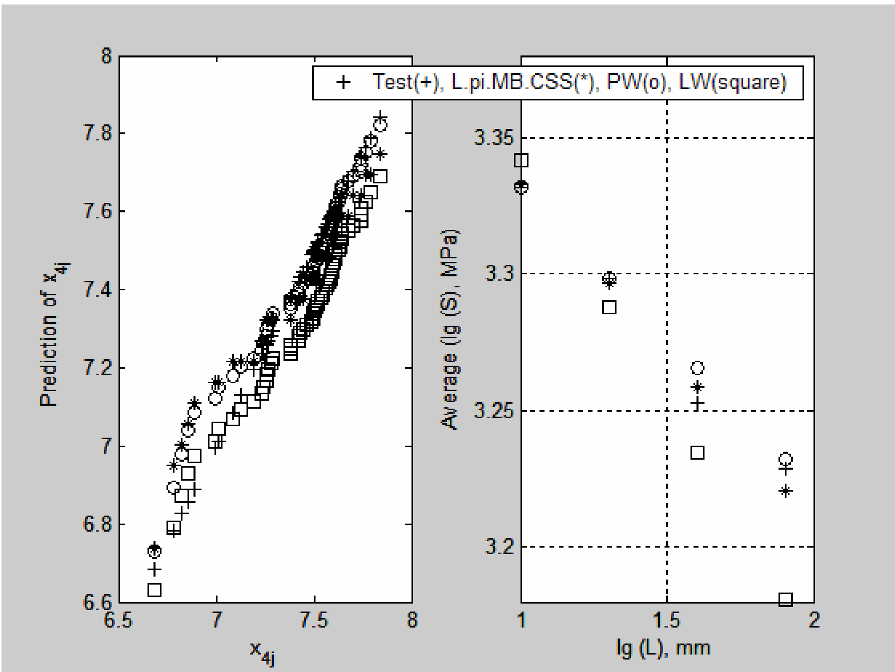


Figure 8

The best results, which are better than results of both LW and PW models (see Fig. 9 and Table 1), we obtained using L.Pm.MBm.C.S.S model (see equation (12) for $F(x)$ and (16) with C = ∞ for $F_Z(x)$). For this model a prior distribution of defect number is the (truncated at m = 2) Poisson distribution with $\lambda = \lambda_1 L / l_1$, where λ_1 is the defect intensity (defect number per specimen length unit), Parameter estimates of the model are λ_1 = 0.15, l_1 = 5, θ_0 = 7.7578, θ_1 = 0.2346. In this paper we did not estimate the parameter δ_0. It was assumed that δ_0 = 0.

We see that the L.Pm.MBm.C.S.S model ensures the minimum of all three statistics.

TABLE 1. The comparison of models

Statistics	L.Pm.MBm.C.S.S	Lmod.P.S.S	L.B1.C.S.St	L.π.B.C.S.S	PW	LW
OSSPt	0.1574	0.3094	0.2630	0.3202	0.2155	0.4760
Q_1	0.1032	0.1279	0.1441	0.1303	0.1644	0.6702
$\overline{R}_{LR}$	0.1479	0.1509	0.2274	0.1545	0.1525	0.1855

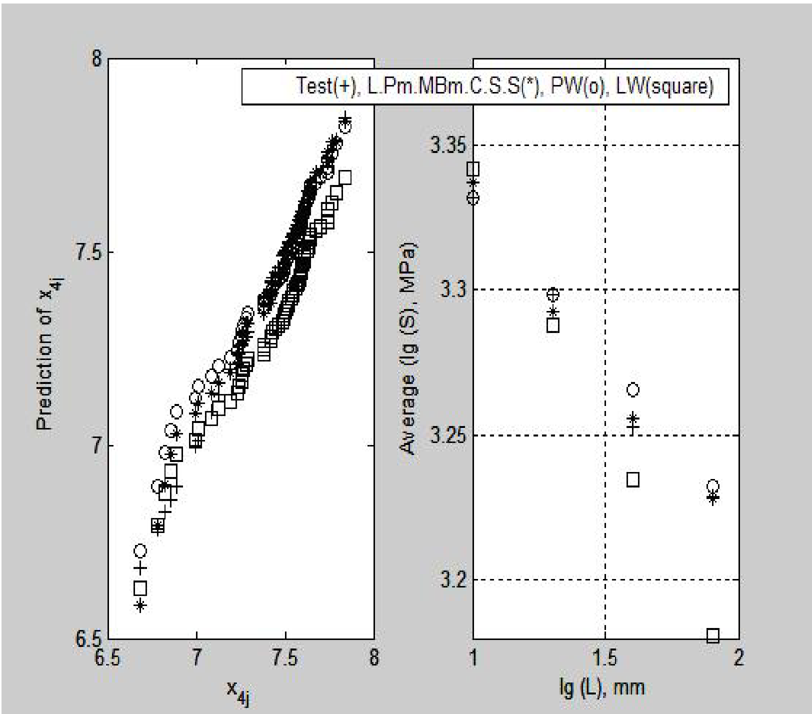


Figure 9

4. Resume

The Model L.Pm.MBm.C.S.S provides the best estimates of fiber strength for L = 80 mm using test data for L = 10 mm and L = 20 mm. All four WLMF models, including those in the Table 1, are providing better estimates of fiber strength dependence on specimen length than both LW and PW models (see statistics Q_1). Common feature of these models is the presence of some form of limitation of this dependence. It is obvious for Model L.B1.C.S.St (C = ∞), where (see (21)) only the probability of defect initiation depends on specimen length (or number of elements, $n = L / l_1$). It can be seen also for Model Lmod.P.S.S. The equation (20 a) corresponds to the assumption that initially there is one defect in specimen regardless of its length. The same is true for L.π.B.C.S.S where $\pi = (0,1,0,...,0)$. For the Model L.Pm.MBm.C.S.S the number of possible defects is deliberately limited by the number ($m+1$). Both models T.B.Zt.S and T.P.C.S have no similar limitation and fail to capture the strength dependence on specimen length.

The Model L.Pm.MBm.C.S.S provides the best agreement with the experimental dataset among the considered models. But we should take into account that it has five unknown parameters: θ_0, θ_1, l_1, λ_1 and m. PW model has three parameters only. Evidently we have random conclusions because we have random dataset. But it seems that the presented distribution family has great potential (for example, we have wide choice of $F_Z(x), F_Y(x)$, $F_0(x), \ldots$) and deserves to be studied much more thoroughly using much more test data. We should mention also that the considered distribution family can be applied not only to the fiber strength analysis but to the analysis of reliability of any series system with two types of elements as well.

References

1. Paramonov, Yu. and Andersons, J. A new model family for the strength distribution of fibers in relation to their length, *Mechanics of Composite Materials*, Vol. 42(2), 2006, pp. 179-192.
2. Paramonov, Yu. and Andersons, J. A family of weakest link models for fibers strength distribution, *Composite: Part A,* Vol. 38, 2007, pp. 1227-1233.
3. Watson, A.S., Smith, R.L. An examination of statistical theories for fibrous materials in the light of experimental data, *J. Mater. Sci.*, Vol. 20, 1985, pp. 3260-3270.
4. Padgett, W. J., Durham, S. D., Mason, A. M. Weibull analysis of the strength of carbon fibers using linear and power law models for the length effect, *J. Comp. Mater.,* Vol. 29, No 14, 1995, pp. 1873-1884.
5. Gutans, Yu. A., Tamuzh, V. P. To scale distribution of Weibull's effect of durability of fibres, *Mechanics of Composite Materials*, No 6, 1984, pp. 1107-1109. (In Russian)

Computer Modelling and New Technologies, 2007, Vol.11, No.1, 21-30
Transport and Telecommunication Institute, Lomonosova 1, Riga, LV-1019, Latvia

MODELLING OF TWO STRATEGIES IN INVENTORY CONTROL SYSTEM WITH RANDOM LEAD TIME AND DEMAND

***Eugene Kopytov*[1]*, Leonid Greenglaz*[2]*, Aivar Muravyov*[1]*, Edvin Puzinkevich*[1]**

[1]*Transport and Telecommunication Institute*
Lomonosova Str. 1, Riga, LV-1019, Latvia
Ph.: (+371)9621337, fax: (+371) 7383066, e-mail: kopitov@tsi.lv

[2]*Riga International School of Economics and Business Administration*
Meza Str. 1, build. 2, Riga, LV-1048, Latvia
E-mail: gringlaz@riceba.lv

The paper considers two multiple period single-product inventory control models with random parameters. These models are of interest because they illustrate real situations of the business. The first model is a model with fixed reorder point and fixed order quantity. The second model is the model with fixed period of time between the moments of placing neighbouring orders. Order quantity is determined as difference between the fixed stock level and quantity of goods in the moment of ordering. The considered models are realized using analytical and simulation approaches. The numerical examples of problem solving are presented.

Keywords: *inventory control, demand, lead time, order quantity, reorder point, analytical model, simulation*

1. Introduction

Most inventory control situations of significance are complex. Decision-maker's need to understand this complexity depends on his role within the business and the way he chooses to solve the problems. Mathematical models can provide a description of business situations that are difficult to examine in any other way.

The search of the effective solutions of stock control in transport company should be based on a number of economic, social and technical characteristics [4]. In practice we have to investigate the stochastic models for different situations characterizing inventory control systems; a set of stochastic models are available to solve the inventory control problem [1, 5]. In the given paper two multiple period single-product inventory control models with random demand and lead time are considered.

The first model is a model with fixed reorder point and fixed order quantity. This model describes dependency of average expenses for goods holding, ordering and losses from deficit per time unit on two control parameters – the order quantity and reorder point. The description of this model and analytical method of problem solving are examined in the previous authors' work [2]. We have solved this problem using regenerative approach.

The second model is a model with fixed time interval between the moments of placing neighbouring orders. In this model the order quantity is determined as difference between the fixed stock level and quantity of goods in the moment of ordering. The analytical description of the second model is considered in the given paper. Note that in the second model we have used the same economical criteria – minimum of average total cost in inventory system.

So, we have two inventory control models with continuously review inventory position (permanent stock level monitoring). The strategy of each model selection is based on the real conditions of the business. Thus, the first model can be used for the system with arbitrary time moment of placing the order; this situation takes place in inventory system used own means of transportation for order delivery. The second model is suggested for the system with fixed moment of placing the orders, where the order transportation depends on schedule of transport departure.

The considered models can be realized using analytical and simulation methods. As it was shown in the previous works of these authors the analytical models are fairly complex. An alternative to solution by mathematical manipulation is simulation [3]. In the given paper analytical and simulation approaches are investigated. The numerical results of problem solving are obtained in simulation package Extend.

2. Description of the Models

2.1. Model 1

We consider a single-product stochastic inventory control model under following conditions. The demand for goods is a Poisson process with intensity λ . In the moment of time, when the stock level falls till certain level *R*, a new order is placed (see Figure 1). The quantity *R* is called as reorder point. The order quantity *Q* is constant. We suppose that $Q \geq R$ .The lead time *L* (time between placing an order and receiving it) has a normal distribution with a mean μ_L and a standard deviation σ_L . There is the possible situation of deficit, when demand D_L during lead time L exceeds the value of reorder point *R*. We suppose that in case of deficit the last cannot be covered by the expected order.

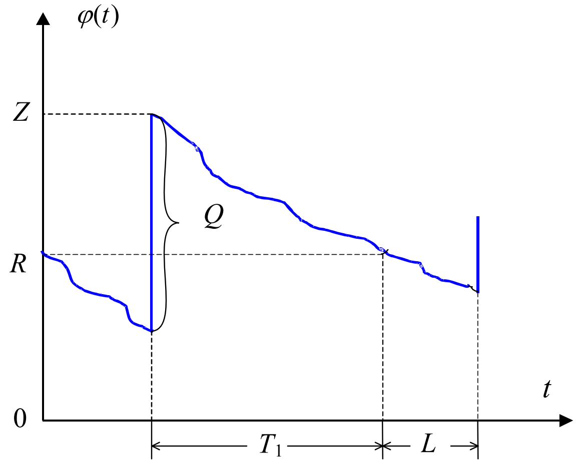


Figure 1. Dynamics of inventory level during one cycle for Model 1

Denote as *Z* the quantity of goods in stock in the time moment immediately after receiving of order. We can determine this quantity of goods *Z* as function of demand D_L during lead time *L*:

$$Z = \begin{cases} R + Q - D_L & \text{if} \quad D_L < R; \\ Q, & \text{if} \quad D_L \geq R. \end{cases} \tag{1}$$

Formula (1) is basic. It allows expressing different economical indexes of the considered process.

Let *T* is the duration of a cycle. Length of the cycle consists of two parts: time T_1 between receiving the goods and placing a new order and lead time *L*, i.e. $T = T_1 + L$.

We suppose that next economic parameters of the model are known:

- the ordering cost C_0 is known function of the order quantity *Q*, i.e. $C_0 = C_0(Q)$;
- the holding cost is proportional to quantity of goods in stock and holding time with coefficient of proportionality C_H ;
- the losses from deficit are proportional to quantity of deficit with coefficient of proportionality C_{SH} .

Let us denote D_τ as demand for goods within period of time τ .

Principal aim of the considered model is to define the optimal values of order quantity *Q* and reorder point *R*, which are control parameters of the model. Criteria of optimization are minimum of average total expenses (costs) per time unit. We solve this problem using regenerative approach [5].

2.2. Model 2

Let us consider the Model 2 with fixed time *T* of the cycle, i.e. with fixed time between neighbouring moments of placing the orders. It is a single-product stochastic inventory control model under the following conditions. The demand for goods is a Poisson process with intensity λ . The lead time L has a normal distribution with a mean μ_L and a standard deviation σ_L . We suppose that lead time is essentially less as time of the cycle: $\mu_L + 3\sigma_L << T$.

There exists the possible situation of deficit, when the demand during time between neighbouring moments of receiving of order exceeds the quantity of goods in stock Z in the time moment immediately after receiving of order. Analogy Model 1 we suppose that in case of deficit the last cannot be covered by expected order.

In Figure 2 the cycle with number k is presented. Let R_k is the rest of goods in stock at the start of the k-th period and R_{k+1} is the rest of goods at the end of k-th cycle (or the rest at start of cycle with number k+1). We denote as S the goods quantity which is needed "ideally" for one period and it equals to the sum $S = \overline{D}_T + S_0$, where $\overline{D}_T$ is the average demand for cycle time; S_0 is the safety stock. In the given sentence we suppose that "ideally" S gives us in future the minimum of total expenditure for ordering, holding and loses from deficit per unit of time.

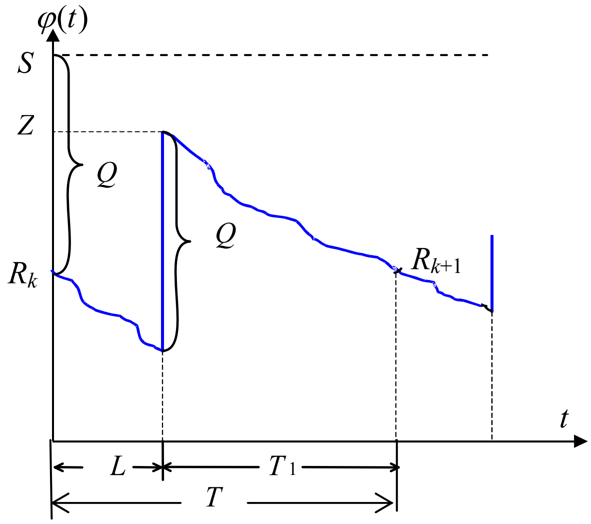


Figure 2. Dynamics of inventory level during *k*-th cycle for Model 2

So, in the suggested model period of time T and stock level S are *control parameters*.

The order quantity Q is the difference

$$Q = S - R_k \, . \tag{2}$$

We suppose that in the moment of time when a new order has to be placed it may be situation, when the stock level is so big that a new ordering doesn't occur. However for generality of model we'll keep the conception of lead time and quantity of goods at the time moment immediately after receiving of order in such case too. It corresponds to real situation when the customer uses the transport means, which depart at the fixed moments of time not depending on existence of the order and which have the random lead time; for example, transportation by trailers, which depart each first and fifteenth day of each month.

Taking into account that in case of deficit it can't be covered by the expected order, we can obtain the expression for goods quantity at the moment of time immediately after receiving of order

$$Z = \begin{cases} R_k + Q - D_L & \text{if} \quad D_L < R_k; \\ Q, & \text{if} \quad D_L \geq R_k. \end{cases} \tag{3}$$

and using (2) we have:

$$Z = \begin{cases} S - D_L & \text{if} \quad D_L < R_k; \\ S - R_k, & \text{if} \quad D_L \geq R_k. \end{cases} \tag{4}$$

The rest R_k at the start of the k-th period and the goods quantity Z at the moment of time immediately after receiving of order take values from interval [0; S]:

- $R_k = 0$, if in the previous cycle the demand during the time T_1 between the receiving of order and placing of the new order is more or equal Z, i.e. $D_{T_1} \geq Z$;

  $R_k = S$, if in the previous cycle Z is equal S and there isn't the demand during the time period T_1, i.e. $Z = S \wedge D_{T_1} = 0$.
- $Z = 0$, if the rest R_k to the moment of ordering is S (i.e. order quantity Q is 0) and demand D_L during lead time L is more or equal S, i.e. $R_k = S \wedge D_L \geq S$;

  $Z = S$, if the rest R_k to the moment of ordering is 0 or demand D_L during lead time L is absent, i.e. $R_k = 0 \vee D_L = 0$.

In the next section we should determine the average total cost per cycle for the fixed rest of product in the moment of ordering.

3. Analytical approach to creation of the models

The analytical description of the Model 1 is presented in the previous paper of the authors [2]. In the given section we consider a detailed creation of Model 2 with fixed period of time between the moments of placing the neighbouring orders.

3.1. Distribution of Demand during Lead Time

As demand for goods is a Poisson flow with intensity λ, we can determine distribution for demand within fixed period of time τ

$$P(D_\tau = i) = \frac{(\lambda\tau)^i}{i!} e^{-\lambda\tau}, \quad i = 1, 2, ... \tag{5}$$

If $f_L(\tau)$ is a density function for lead time L, then distribution for demand D_L within time L can be calculated by formula

$$P(D_L = i) = \int_0^\infty P(D_\tau = i) \cdot f_L(\tau) d\tau . \tag{6}$$

In the case of normal distribution for L we obtain the formula

$$P(D_L = i) = \int_0^\infty \frac{(\lambda\tau)^i}{i!} e^{-\lambda\tau} \frac{1}{\sigma_L\sqrt{2\pi}} e^{-\frac{(\tau-\mu_L)^2}{2\sigma_L^2}} d\tau = \frac{\lambda^i}{i!\sigma_L\sqrt{2\pi}} \int_0^\infty \tau^i e^{-\lambda\tau} e^{-\frac{(\tau-\mu_L)^2}{2\sigma_L^2}} d\tau . \tag{7}$$

3.2. Holding Cost during One Cycle

Calculation process of the holding cost during one cycle is divided in two stages: calculation for lead time L and calculation for time T_1 between receiving the goods and placing a new order.

Let τ is the length of time from the last ordering and $\tau < T$. If the demand D_τ during the time τ equals i, then the holding cost during the time interval (τ, $\tau + d\tau$) is

$$TC_H(D_\tau = i, (\tau, \tau + d\tau)) = C_H(R_k - i)d\tau \tag{8}$$

and expected holding cost during the lead time L is

$$E(TC_{H,L}) = C_H \int_0^\infty f_L(\tau)d\tau \int_0^\tau \sum_{j=0}^{S}\sum_{i=0}^{j} (j-i) \cdot P(D_x = i) \cdot P(R_k = j)dx , \tag{9}$$

where $P(D_x{=}i)$ is define by formula (5).

Let consider the expected holding cost during the time T_1 . If j is the goods quantity Z at the moment of time immediately after receiving of order, τ is time interval after the receiving of order and

$\tau \leq T_1$, and the demand D_τ during this time τ equals *i*, then the holding cost during the time (τ , $\tau + d\tau$) equals

$$TC_H(T_1 > \tau, Z = j, D_\tau = i, (\tau, \tau + d\tau)) = C_H(j - i)d\tau . \tag{10}$$

Let's note that $i \leq j$ and *j* takes values from interval ($S - R_k$, *S*). So, expected holding cost during the time T_1 is

$$E(TC_{H,T_1}) = C_H \int_0^T f_L(T-\tau)d\tau \int_0^\tau \sum_{j=S-r}^{S} \sum_{i=0}^{j} (j-i) \cdot P(Z=j) \cdot P(D_x = i)dx , \tag{11}$$

where condition $T_1 > \tau$ is equivalent to condition $L < T - \tau$;

$P(Z = j) = P(D_L = S - j)$, if $S - R_k < j \leq S$;

$$P(Z = S - r) = 1 - \sum_{j=S-R_k+1}^{S} P(Z = j) . \tag{12}$$

Average holding cost $E(TC_H)$ within cycle *T* is the sum of the corresponding addendums:

$$E(TC_H) = E(TC_{H,L}) + E(TC_{H,T_1}) . \tag{13}$$

3.3. Losses from Deficit

Similar to previous point the calculation process of the losses from deficit during one cycle is divided into two stages: calculation for lead time *L* and calculation for time T_1 between receiving the goods and placing a new order.

If within lead time L the demand D_L exceeds the value of reorder point R_k , then deficit of goods is present. Let D_L=*i* and $i > R_k$, then losses from deficit are $C_{SH} \cdot (i - R_k)$. So, average shortage cost within lead time

$$E(TC_{SH,L}) = C_{SH} \sum_{i=R_k+1}^{\infty} P(D_L = i) \cdot (i - R_k) . \tag{14}$$

Let demand for time T_1 equals *i*, $D_{T_1}=i$, and the goods quantity *Z* at the moment of time immediately after receiving of order is *j* and $i > j$. Then losses from deficit are $C_{SH} \cdot (i - j)$. Thus an average shortage cost during the time T_1 is

$$E(TC_{SH,T_1}) = C_{SH} \sum_{j=S-R_k}^{S} P(Z = j) \sum_{i=j+1}^{\infty} (i - j) \cdot P(D_{T_1} = i) , \tag{15}$$

where $P(D_{T_1} = i) = \int_0^T P(D_\tau = i) \cdot f_L(T - \tau)d\tau$ and probability $P(Z = j)$ is calculated by formula (12).

An average shortage cost $E(TC_{SH})$ within cycle is the sum of the corresponding addendums:

$$E(TC_{SH}) = E(TC_{SH,L}) + E(TC_{SH,T_1}) . \tag{16}$$

Finally an average total cost for a cycle is

$$E(TC) = E(TC_H) + E(TC_{SH}) + C_0 , \tag{17}$$

where $E(TC_H)$ and $E(TC_{SH})$ is calculated by formulas (13) and (16) accordingly, and average total cost per time unit in inventory system is

$$E(AC) = \frac{E(TC)}{T} . \tag{18}$$

Using the known distributions of demand and lead time and formula (4), applying recurrence method we can find the conditional distribution of the rest of product R_{k+1} at the end of *k*-th cycle (start of cycle with number *k*+1) for the known value of rest R_k *:*

$$R_{k+1} = \begin{cases} Z - D_{T-L}, & \text{if} \quad D_{T-L} < Z; \\ 0, & \text{if} \quad D_{T-L} \geq Z \end{cases} \tag{19}$$

and combining expressions (4) and (19) we have

$$R_{k+1} = \begin{cases} S - D_L - D_{T-L} = S - D_T, & \text{if } D_L < R_k,\, D_{T-L} < Z = S - D_L; \\ S - R_k - D_{T-L}, & \text{if } D_L \geq R_k,\, D_{T-L} < Z = S - R_k; \\ 0, & \text{if } D_L < R_k,\, D_{T-L} \geq Z = S - D_L; \\ 0, & \text{if } D_L \geq R_k,\, D_{T-L} \geq Z = S - R_k. \end{cases} \tag{20}$$

In accordance with (19) we can calculate probability of event $R_{k+1} = j$ for condition that the rest at the beginning of cycle equals to R_k . As it is evident from (20) the rest R_{k+1} takes values from interval [0; *S*]. In particular $R_{k+1} = S$ if demand D_T during period of cycle *T* is absent, i.e. $D_T = 0$.

At first let's consider the case $R_{k+1} = j$, where $j > 0$. According to condition of the task we can write $L \leq T$. Let $L = \tau$ and demand D_τ during time τ equals to *x* and $x < R_k$. In this case $Z = S - x$ and the request $R_{k+1} > 0$ is equivalent to the condition $D_{T-\tau} = S - x - j$ accordingly the first line of the formula (20). Then probability of event that the rest of product R_{k+1} at the end of cycle equals to *j* (where $j > 0$) under the condition that $x < R_k$ and $L \in (\tau, \tau + d\tau)$ is calculated by the formula:

$$P(R_{k+1} = j / x < R_k \wedge L \in (\tau, \tau + d\tau)) = \sum_{x=0}^{R_k - 1} P(D_\tau = x) \cdot P(D_{T-\tau} = S - x - j) \cdot f_L(\tau) d\tau \,. \tag{21}$$

Accordingly

$$P(R_{k+1} = j / x < R_k) = \int_0^T \left(\sum_{x=0}^{R_k - 1} P(D_\tau = x) \cdot P(D_{T-\tau} = S - x - j) \right) \cdot f_L(\tau) d\tau \,. \tag{22}$$

Similarly, if $L = \tau$ and $D_\tau = x$, where $x > R_k$, then $Z = S - R_k$ and request $R_{k+1} > 0$ is equivalent to condition $D_{T-\tau} = S - R_k - j$, the probability $P(R_{k+1} = j / x > R_k)$ is calculated by formula

$$P(R_{k+1} = j / x > R_k) = \int_0^T \left(\sum_{x=R_0}^{\infty} P(D_\tau = x) \cdot P(D_{T-\tau} = S - R_k - j) \right) f_L(\tau) d\tau \,. \tag{23}$$

Finally, if $j > 0$, then

$$P(R_{k+1} = j) = \int_0^T \left[\sum_{x=0}^{R_k - 1} P(D_\tau = x) \cdot P(D_{T-\tau} = S - x - j) + \sum_{x=R_k}^{\infty} P(D_\tau = x) P(D_{T-\tau} = S - R_k - j) \right] f_L(\tau) d\tau \,. \tag{24}$$

Reasoning by analogy it can be shown, that, if $j = 0$, then

$$P(R_{k+1} = 0) = \int_0^T \left[\sum_{x=0}^{R_k - 1} P(D_\tau = x) \cdot P(D_{T-\tau} > S - x) + \sum_{x=R_k}^{\infty} P(D_\tau = x) \cdot P(D_{T-\tau} > S - R_k) \right] \cdot f_L(\tau) d\tau \,. \tag{25}$$

For analytical solving of the considered problem we have created a complex of programs realized on the base of programming system DELPHI. For calculation there were used standard quantitative methods.

4. Simulation Approach

As it was shown in the previous section the analytical inventory control model is rather complex. As alternative to analytical approach the authors have used simulation models realized in the simulation package Extend [3].

4.1. Model 1

Let us consider the model with two fixed control parameters: reorder point *R* and order quantity *Q*. The schema of the task simulation is shown in Figure 3.

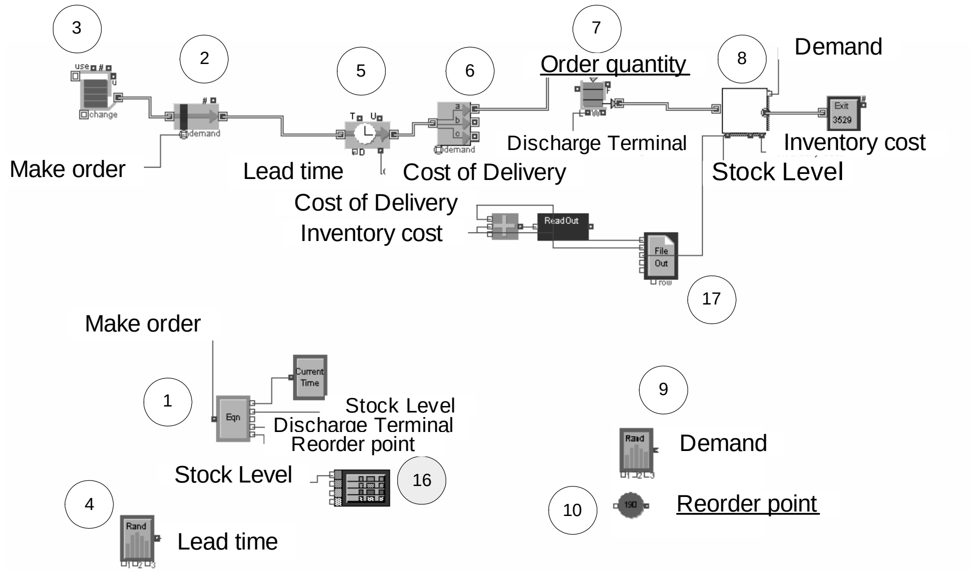


Figure 3. Simulation model overview: inventory control with fixed reorder point and fixed order quantity

Let us consider the main blocks of the simulation schema. In the block #1 the decision of a new ordering (*Make Order*) is generated using data about *Reorder point* (block #10) and quantity of goods in stock (*Stock level*). As the result variable *Make Order* takes value 1, it is transmitted to connector of block #2, and a new goods ordering is executed. In block #5 the process of order delivery is simulated. The value of random lead time is generated in block #4 (*Input Random Number*) using parameters μ_L and σ_L of normal distribution. The demand for goods is generated in block #9 as random value with Poisson distribution and known parameter λ. The warehouse is realized in hierarchical block #8, which schema is shown in Figure 4. Process of goods realization is simulated in block #11. Block #12 (dummy source of goods) and block #13 (*Set Attribute*) are used for good deficit calculation. The results of simulation are printed out in text file (block #17) and on the screen (block #16).

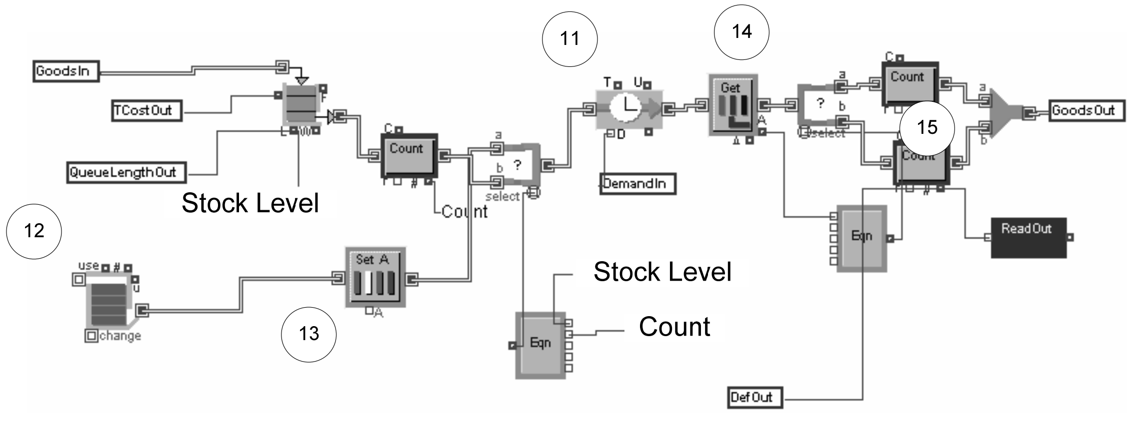


Figure 4. Warehouse simulation model overview

Using the created simulation model we can find the optimal solution for inventory control problem with two control parameters – reorder point R and order quantity Q (see Example 1).

Example 1. Let demand D for goods is a Poisson process with intensity 10 units per day; lead time L has a normal distribution with a mean 11 days and a standard deviation 3,5; ordering cost C_0 equals to 200 EUR, holding cost C_H equals to 2 EUR per unit per year, losses from deficit C_{SH} equals to 8 EUR per unit; unit time is 1 year. The period of simulation is one year and a number of realizations are 100.

The results of simulation are shown in Table 1 and in Figure 5. Note that for the given steps of the control parameters changing the best result is achieved at the point $Q = 950$ units and $R = 150$ units, where for 100 realizations an average total cost for one year period equals 1889,34 EUR.

TABLE 1. Average total cost per year in inventory system with fixed reorder point and order quantity (Model 1)

	Reorder point, units				
Order quantity, units	*100*	*150*	*200*	*250*	*300*
850	2430,32	1988,34	2026,22	2113,90	2209,30
900	2224,99	2001,77	2051,28	2141,19	2235,84
950	2241,90	**1889,34**	1953,86	2092,33	2236,53
1000	2267,96	1960,65	1993,83	2071,15	2153,34
1050	2387,28	2030,89	2048,83	2135,75	2216,93

Model 2. Let us consider second strategy of inventory control with fixed period of time T between the moments of placing neighbouring orders. Note that in the suggested model period of time T and required stock level S are *control parameters.*

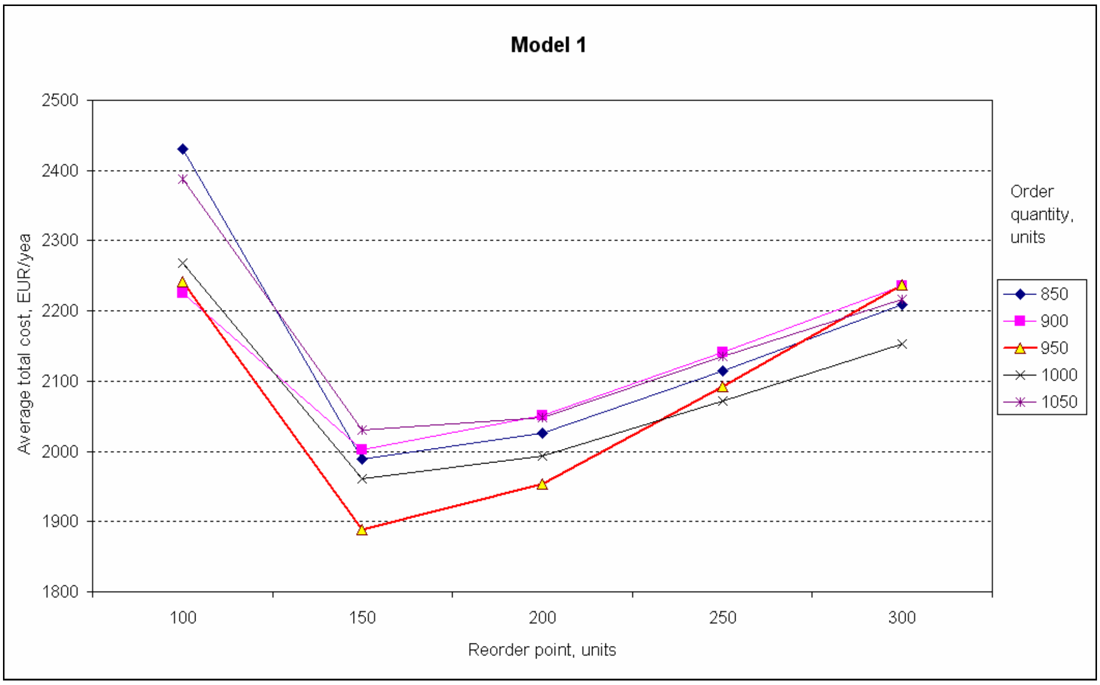


Figure 5. Average total cost per year in inventory system with fixed reorder point and fixed order quantity

For simulation of inventory control process we have created the schema shown in Figure 6. Let us consider the main blocks of schema. Block #1 generates the transactions in the fixed moments of time; these transactions are used for simulation of goods ordering during the considered time period. Block #2 calculates the *Order quantity* using data about *Stock level* in the moment of ordering and *Required stock level* (quantity of goods which is needed "ideally" for one period); this result is saved in block #3 (*Set Attribute*). Block #4 determines the moment of order delivery using the value of lead time generated in block #5 (*Input Random Number*) as random variable with normal distribution and known parameters. The demand for goods is generated in block #11 as random value with Poisson distribution and known parameter. Process of goods realization is simulated in block #10. Blocks #8 and #9 are used for goods deficit calculation. The results of simulation are printed out in text file and are shown on the screen.

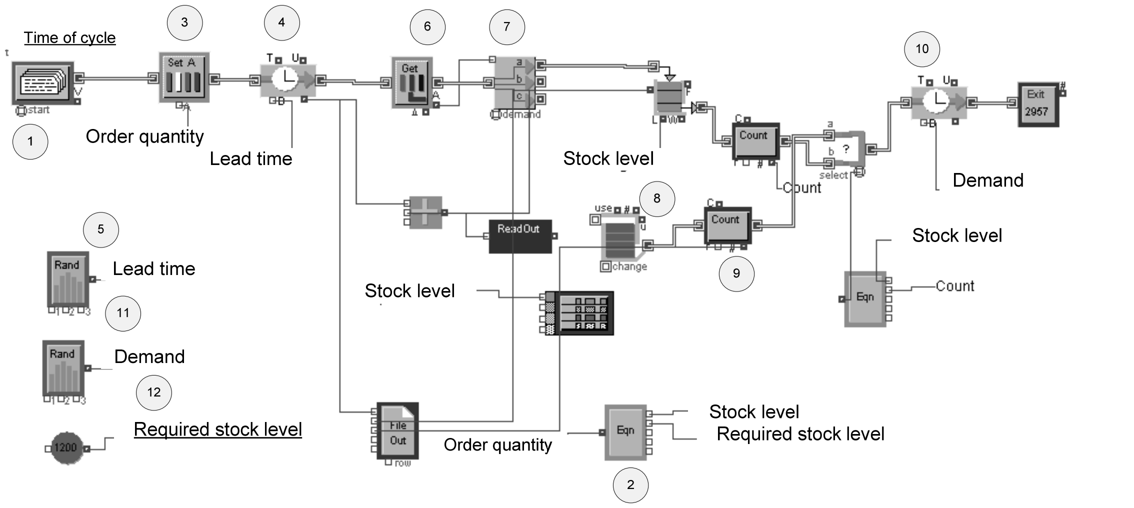


Figure 6. Simulation model overview: inventory control with fixed time interval between placing neighbouring orders

Example 2. Let us consider another strategy of inventory control accordingly Model 2 using initial data from Example 1. For problem solving we have used the simulation model shown in Figure 6. The results of simulation are shown in Table 2 and in Figure 7. For the given steps of control parameters changing the best result is achieved at the point $S = 900$ units of goods and $T = 75$ days, where for 100 realization an average total cost for one year period equals 1965,9 EUR.

TABLE 2. Average total cost per year in inventory system with fixed time interval between placing neighbouring orders (Model 2)

	Time interval between placing neighbouring orders, days						
Level up to order, units	*70*	*75*	*80*	*85*	*90*	*100*	*110*
850	2091,40	2206,92	2826,08	3512,02	3891,42	5213,66	7489,90
900	2108,88	***1965,99***	2287,16	2287,16	3237,74	4365,35	6352,34
950	2203,41	1985,33	2022,51	2341,89	2552,83	3643,92	5308,60
1000	2300,46	2076,96	2044,00	2069,62	2212,28	2945,29	4403,56
1050	2396,41	2179,56	2144,75	2079,61	2075,02	2497,04	3655,52

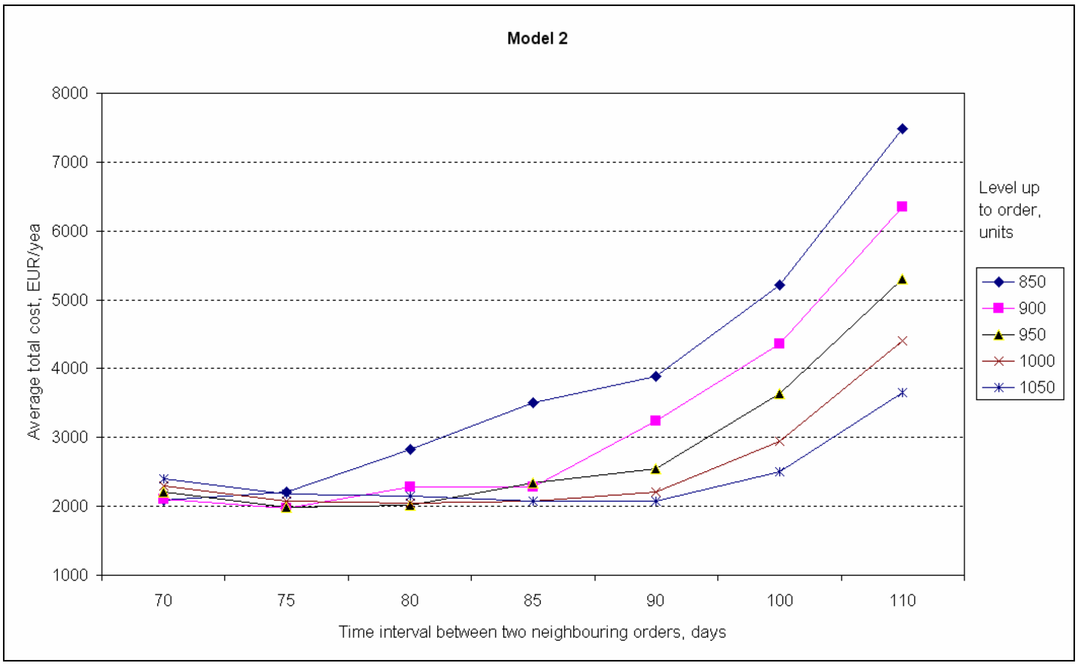


Figure 7. An average total cost per year in inventory system with fixed time interval between placing neighbouring orders

Conclusions

Principal aim of the proposed models is to define the exact order quantity and time of the ordering to achieve minimum expenses for holding, ordering goods and losses from deficit per time unit for transport companies.

Two considered models of inventory control, based on different principles of ordering, give the closely related results near optimum solution.

The main advantages of the considered methods of solving the inventory control problems for the suggested models are as follows:

- simulation approach gives
  - the clearness of the presentation of results; firstly, it touches the case of analysis of total expenses dependence on one control parameter with fixing others;
  - the possibility of finding optimum solution of an inventory problem in the case when realization of analytical model is rather difficult;
- analytical approach gives
  - the mathematical model of situation;
  - the various possibilities of analysis;
  - universality of usage.

In the examined paper single-product inventory control models are considered. In the present research the authors investigate multi-product model with random correlated demands for different goods. In this research we use the simulation modelling in inventory system with a fixed moment of placing the order. In particular the random demand vector is generated using demand statistics and Holecky decomposition of correlation matrix.

References

1. Chopra, S. and P. Meindl. *Supply Chain Management*. London: Prentice Hall, 2001.
2. Kopytov, E. and L. Greenglaz. On a task of optimal inventory control. In: *Proceedings of XXIV International Seminar on Stability Problems for Stochastic Models, Sept. 9-17, 2004, Jurmala, Latvia.* Riga: Transport and Telecommunication Institute, 2004, pp. 247-252.
3. Kopytov, E., Muravjov, A., Greenglaz, L. and Burakov, G. Investigation of two strategies in inventory control system with random parameters. In: *Proceedings of the 21st European Conference on Modelling and Simulation (ECMS 2007). June 4-6, 2007, Prague, Check Republic.* Thomas Bata University in Zlin, 2007, pp. 566-571.
4. Kopytov, E., Tissen, F., Greenglaz, L. Inventory Control Model for the Typical Railways Company. In: *Proceedings of the International Conference "Reliability and Statistics in Transportation and Communication" (RelStat'03), 16-17 October 2003, Riga, Latvia,* Riga: Transport and Telecommunication Institute, Vol. 5 (1), 2004, pp. 39-45.
5. Ross, Sh. M. *Applied Probability Models with Optimization Applications*. New York: Dover Publications, Inc., 1992.

Computer Modelling and New Technologies, 2007, Vol.11, No.1, 31-37
Transport and Telecommunication Institute, Lomonosova 1, Riga, LV-1019, Latvia

ON A PROBLEM OF SPATIAL ARRANGEMENT OF SERVICE STATIONS

Alexander Andronov, Andrey Kashurin

Riga Technical University
Kalku Str.1, Riga, LV-1658 Latvia
E-mail: Aleksandrs.Andronovs@rtu.lv, andrejs.kasurins@sp.gov.lv

A problem of service station arrangement in spatial space is considered. A density function of serviced object location and a function that describes the corresponding loss are known. As criteria of arrangement is an average total loss. For the optimisation the gradient method is used. Numerical examples illustrate the suggested approach to setting the problem solution

Keywords: *spatial arrangement, service stations, gradient method*

1. Introduction

Let us consider a real space *X* for that concrete *point* will be marked by ***x***, for plane it is two-dimensional vector (it is available to consider another dimension too). *A distance* $l(x, x^*)$ is determined for points ***x*** and ***x****, that satisfies usual conditional of distance axioms: $l(x,x)=0$, $l(x,x^*)\geq 0$, $l(x,x^*)\leq l(x,x')+l(x',x^*)$.

Some *objects* are arranged in the space (for example men, animals, stationers). Let us name as ***x***–*object,* the object that is at the point ***x***. The density of object arrangement is described by known density function $f(\boldsymbol{x}) \geq 0$, so

$$\int_{x\in X} f(x)x = 1.$$

Some *service stations* must be arranged in the space, their number is *k*. It is necessary to determine those coordinates $x^{(1)}, x^{(2)}, ..., x^{(k)}$. If a ***x***–object is serviced by *i*–th station then corresponding loss is equal to $g_x\left(x^{(i)}\right)$, for example $g_x\left(x^{(i)}\right) = g(x - x^{(i)})$. Let us call $g_x(\circ)$ as *loss function* and suppose that it is a symmetry according to zero ($g_x(x^{(i)}) = g_x(-x^{(i)})$) and convex (down).

All amount of service for the ***x***–object is deviated between various service stations according to inverse proportion of the distances from the ***x***–object and the station. Most precisely, a part of ***x***–object service that belongs to the *i-th* station is

$$\delta_i(x) = \frac{\left(l\left(x, x^{(i)}\right)\right)^{-1}}{\sum_j \left(l\left(x, x^{(j)}\right)\right)^{-1}}. \qquad (1)$$

Now a problem can be formulated as follows: to find coordinates $x^{(1)}, x^{(2)}, ..., x^{(k)}$ of station arrangement that minimizes the total loss:

$$D\left(x^{(1)}, x^{(2)}, ..., x^{(k)}\right) = \int \frac{1}{\sum_{i=1}^{k}\left(l\left(x, x^{(i)}\right)\right)^{-1}} \sum_{i=1}^{k} \left(l\left(x, x^{(i)}\right)\right)^{-1} g_x(x^{(i)}) f(x) dx. \qquad (2)$$

The article is organized in the following way. At first one-dimensional case of space *X* and the corresponding example is considered. Then we consider two-dimensional case. The article ends by some conclusion remarks. The Appendix contains the analytical investigation of the simplest one-dimensional case when *k* = 1 and density $f(\boldsymbol{x})$ and loss function $g_x(\circ)$ are symmetric functions.

2. One-Dimensional Case

At first we consider a case when space X is real axis R = (- ∞, ∞). We will use a gradient method for the minimization of criterion (2). For that aim let us calculate a corresponding gradient. For a partial derivative of (2) we have the following expression:

$$\frac{\partial}{\partial x^{(j)}} D\left(x^{(1)}, x^{(2)},..., x^{(k)}\right) = \frac{\partial}{\partial x^{(j)}} \int \frac{1}{\sum_i \left(l\left(x, x^{(i)}\right)\right)^{-1}} \sum_{i=1}^{k} \left(l\left(x, x^{(i)}\right)\right)^{-1} g_x(x^{(i)}) f(x) dx =$$

$$= \int \frac{1}{\sum_i \left(l\left(x, x^{(i)}\right)\right)^{-1}} \left(\left(g_x(x^{(j)}) \left(-\left(l\left(x, x^{(i)}\right)\right)^{-2} \right) \frac{\partial}{\partial x^{(j)}} l\left(x, x^{(j)}\right) + \frac{1}{l\left(x, x^{(j)}\right)} \frac{\partial}{\partial x^{(j)}} g_x(x^{(j)}) \right) \right) f(x) dx -$$

$$- \int \left(\sum_i \left(l\left(x, x^{(i)}\right)\right)^{-1} \right)^{-2} \left(\left(-\left(l\left(x, x^{(j)}\right)\right)^{-2} \right) \frac{\partial}{\partial x^{(j)}} l\left(x, x^{(j)}\right) \sum_i \left(l\left(x, x^{(i)}\right)\right)^{-1} g_x(x^{(i)}) \right) f(x) dx. \qquad (3)$$

Now we are able to rewrite the gradient of D as

$$\nabla D(x^{(1)}, x^{(2)},..., x^{(k)}) = \left(\frac{\partial}{\partial x^{(1)}} D(x^{(1)}, x^{(2)},..., x^{(k)}),..., \frac{\partial}{\partial x^{(k)}} D(x^{(1)}, x^{(2)},..., x^{(k)}) \right)^T. \qquad (4)$$

To accelerate a convergence of the gradient method we use a two-stage procedure. At the first stage we use component-wise (coordinate-wise) modification of the gradient method. It means that a sequence of cycles is preformed. Each cycle contains k iterations. During the i-th iteration (j = 1, 2, …, k) function (2) is minimized with respect to coordinate $x^{(j)}$, at the same time other coordinates do not change. For that minimization the gradient method with gradient (3) is used. The cycles end when the change of function (2) is mall. In the second stage we calibrate the obtained result by using the usual gradient method with gradient (4).

3. Example of One-Dimensional Case

Let density function be a mixture of normal distributions with means $\mu^{(1)}, \mu^{(2)},..., \mu^{(r)}$ and variances $\left(\sigma^{(1)}\right)^2, \left(\sigma^{(2)}\right)^2,..., \left(\sigma^{(r)}\right)^2$, and weighted coefficients $p_1, p_2,..., p_r$ $\left(p_1 + p_2 + ... + p_r = 1\right)$:

$$f(x) = \sum_{i=1}^{r} p_i \frac{1}{\sqrt{2\pi}\sigma^{(i)}} \exp\left(-\frac{1}{2} \left(\frac{x - \mu^{(i)}}{\sigma^{(i)}} \right)^2 \right), -\infty < x < \infty. \qquad (5)$$

Further let us use the following distance function and loss function:

$$l(x,z) = |x - z|, \qquad (6)$$

$$g(x,z) = (x - z)^2. \qquad (7)$$

Then we have the following derivatives:

$$\frac{\partial}{\partial z} l(x,z) = \begin{cases} 1 & if\ x < z, \\ -1 & otherwise, \end{cases} \qquad (8)$$

$$\frac{\partial}{\partial z} g(x,z) = -2(x - z). \qquad (9)$$

Now we are able to use formula (3) for optimization.

Let us consider the following numerical data: $k = 4$, $r = 9$ and

$p = (0.1 \quad 0.2 \quad 0.15 \quad 0.05 \quad 0.1 \quad 0.06 \quad 0.04 \quad 0.13 \quad 0.17)^T$,

$\mu = (0 \quad 3.15 \quad 0.05 \quad 4.05 \quad 6.1 \quad 7.06 \quad 7.74 \quad 8.13 \quad 10.17)^T$,

$\sigma = (0.2 \quad 1.2 \quad 1.15 \quad 1.05 \quad 0.71 \quad 2.06 \quad 1.74 \quad 2.13 \quad 1.17)^T$.

The Figure 1 contains an according graphic of density function *f*(*x*).

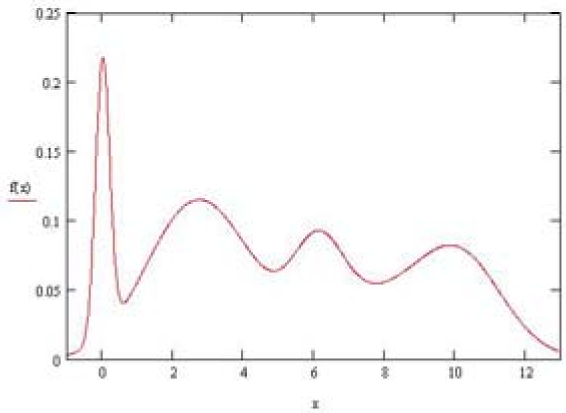


Figure 1. Plot of function f(x) for one-dimensional case

We begin the first stage of the optimization procedure with the values of coordinates $x = \left(x^{(1)}, x^{(2)}, .x^{(3)} x^{(4)}\right)^T = (1 \quad 2 \quad 4 \quad 7)^T$. It corresponds to $D = 9.766$ value of criterion (2). Table 1 contains the results of sequential cycles.

TABLE 1. Results of sequential cycles for one-dimensional case

Iteration number	0	1	2	3	4	5
$x^{(1)}$	1	1	1	1	0.174	0.174
$x^{(2)}$	2	2	2	3.259	3.259	3.259
$x^{(3)}$	4	4	6.207	6.207	6.207	6.207
$x^{(4)}$	7	9.809	9.809	9.809	9.809	10
$D\left(x^{(1)}, x^{(2)}, .x^{(3)} x^{(4)}\right)$	9.766	9.327	8.818	7.727	7.494	7.468

We see that minimal value of criterion (2) is equal to $D = 7.468$ that is calculated by

$x = (0.174 \quad 3.259 \quad 6.207 \quad 10)^T$.

Further we perform the second stage of the optimization procedure and finally get minimal value $D = 7.445$ that corresponds to coordinates

$x = (0.193 \quad 3.176 \quad 6.428 \quad 10.035)^T$.

4. Two-Dimensional Case

Now we consider a case when space X is real plane $R^2 = (-\infty, \infty) \times (-\infty, \infty)$. Then the coordinates of an object are $x = (x_1 \quad x_2)^T$, coordinates of the *j*-st station are $x^{(j)} = \left(x_1^{(j)} \quad x_2^{(j)}\right)^T$. Now instead of scalar derivative (3) we have two-dimensional vector

$$\frac{\partial}{\partial x^{(j)}} D\left(x^{(1)}, x^{(2)}, ..., x^{(k)}\right) = \left(\frac{\partial}{\partial x_1^{(j)}} D\left(x^{(1)}, x^{(2)}, ..., x^{(k)}\right) \quad \frac{\partial}{\partial x_2^{(j)}} D\left(x^{(1)}, x^{(2)}, ..., x^{(k)}\right) \right)^T .$$

Analogously to (3) we have for partial derivative (i = 1, 2):

$$\frac{\partial}{\partial x_q^{(j)}} D\left(x^{(1)}, x^{(2)}, ..., x^{(k)}\right) = \frac{\partial}{\partial x_q^{(j)}} \int \frac{1}{\sum_i \left(l\left(x, x^{(i)}\right)\right)^{-1}} \sum_{i=1}^{k} \left(l\left(x, x^{(i)}\right)\right)^{-1} g_x(x^{(i)}) f(x) dx =$$

$$= \int \left(\sum_i \left(l\left(x, x^{(i)}\right)\right)^{-1} \right)^{-1} \left(\left(g_x(x^{(j)}) \left(-\left(l\left(x, x^{(j)}\right)\right)^{-2} \right) \frac{\partial}{\partial x_q^{(j)}} l\left(x, x^{(j)}\right) + \left(l\left(x, x^{(j)}\right)\right)^{-1} \frac{\partial}{\partial x_q^{(j)}} g_x(x^{(j)}) \right) \right) f(x) dx -$$

$$- \int \left(\sum_i \left(l\left(x, x^{(i)}\right)\right)^{-1} \right)^{-2} \left(\sum_i \left(l\left(x, x^{(i)}\right)\right)^{-1} g_x(x^{(i)}) \left(-\left(l\left(x, x^{(j)}\right)\right)^{-2} \right) \frac{\partial}{\partial x_q^{(j)}} l\left(x, x^{(j)}\right) \right) f(x) dx. \qquad (10)$$

Now instead of (4) we have the ($2 \times k$)-matrix of the partial derivatives

$$\nabla D(x^{(1)}, x^{(2)}, ..., x^{(k)}) = \begin{pmatrix} \frac{\partial}{\partial x_1^{(1)}} D(x^{(1)}, x^{(2)}, ..., x^{(k)}) & \cdots & \frac{\partial}{\partial x_1^{(k)}} D(x^{(1)}, x^{(2)}, ..., x^{(k)}) \\ \frac{\partial}{\partial x_2^{(1)}} D(x^{(1)}, x^{(2)}, ..., x^{(k)}) & \cdots & \frac{\partial}{\partial x_2^{(k)}} D(x^{(1)}, x^{(2)}, ..., x^{(k)}) \end{pmatrix}. \qquad (11)$$

For the optimization we again use the two-stage procedure. At the first stage the component-wise (coordinate-wise) modification is used as follows. During the j-th iteration (j = 1, 2, …, k) function (2) is minimized with respect to both coordinates of the j-st station $x^{(j)} = \left(x_1^{(j)} \quad x_2^{(j)}\right)$, at the same time other coordinates do not change. According to the gradient method we move along the gradient with respect to $\left(x_1^{(j)} \quad x_2^{(j)}\right)$, recalculating the one continually. At the second stage we work with the full gradient (11).

5. Example of Two-Dimensional Case

As before, let density function be a mixture of two-dimensional normal distributions with means $\mu^{(1)} = \left(\mu_1^{(1)} \quad \mu_2^{(1)}\right)^T, ..., \mu^{(r)} = \left(\mu_1^{(r)} \quad \mu_2^{(r)}\right)^T$ and variances $\sigma^{(1)} = \left(\sigma_1^{(1)} \quad \sigma_2^{(1)}\right)^T, ..., \sigma^{(r)T}, = \left(\sigma_1^{(r)} \quad \sigma_2^{(r)}\right)^T$ and weighted coefficients $p_1, p_2, ..., p_r$ $(p_1 + p_2 + ... + p_r = 1)$:

$$f(x) = f\left(\left(x_1 \quad x_2\right)^T\right) = \sum_{i=1}^{r} p_i \frac{1}{\sqrt{2\pi\left(1-\rho^2\right)}\sigma_1^{(i)}\sigma_2^{(i)}} \times$$
$$\times \exp\left(-\frac{1}{2\left(1-\rho^2\right)} \left(\left(\frac{x_1 - \mu_1^{(i)}}{\sigma_1^{(i)}} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1^{(i)}}{\sigma_1^{(i)}} \right) \left(\frac{x_2 - \mu_2^{(i)}}{\sigma_2^{(i)}} \right) + \left(\frac{x_2 - \mu_2^{(i)}}{\sigma_2^{(i)}} \right)^2 \right) \right). \qquad (12)$$

Further let us use the following distance function and loss function:

$$l(x, z) = \sqrt{\left(x_1 - z_1\right)^2 + \left(x_2 - z_2\right)^2}\,, \qquad (13)$$

$$g(x, z) = \left|x_1 - z_1\right| + \left|x_2 - z_2\right|. \qquad (14)$$

Then we have the following derivatives (i = 1, 2):

$$\frac{\partial}{\partial z_q} l(x, z) = -\frac{1}{\sqrt{\left(x_1 - z_1\right)^2 + \left(x_2 - z_2\right)^2}} \left(x_q - z_q\right), \qquad (15)$$

$$\frac{\partial}{\partial z_q} g(x,z) = \frac{\partial}{\partial z_q} g\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}\right) = \begin{cases} 1 \ \text{if} \ x_q < z_q, \\ -1 \ \text{otherwise.} \end{cases} \tag{16}$$

Now we are able to use formula (3) for optimization.

Let us consider the following numerical data: $k = 4$, $r = 9$ and

$$p = \begin{pmatrix} 0.1 & 0.2 & 0.15 & 0.05 & 0.1 & 0.06 & 0.04 & 0.13 & 0.17 \end{pmatrix},$$

$$\mu = \begin{pmatrix} 0 & 2 & 3.15 & 4.05 & 6.1 & 7.06 & 7.74 & 8.13 & 10.2 \\ 3 & 2 & 1 & 8 & 3 & 6 & 3.1 & 6 & 7 \end{pmatrix},$$

$$\sigma = \begin{pmatrix} 0.2 & 1.2 & 1.15 & 0.7 & 2.1 & 1.76 & 1.74 & 2.13 & 1.2 \\ 1.3 & 0.4 & 1 & 0.8 & 1.3 & 1.7 & 1.5 & 1.6 & 2.7 \end{pmatrix},$$

$$\rho = \begin{pmatrix} 0 & 0.2 & -0.15 & -0.7 & 0.6 & 0.06 & -0.74 & 0.13 & -0.17 \end{pmatrix}.$$

The Figure 2 contains an according graphic of density function $f(x)$.

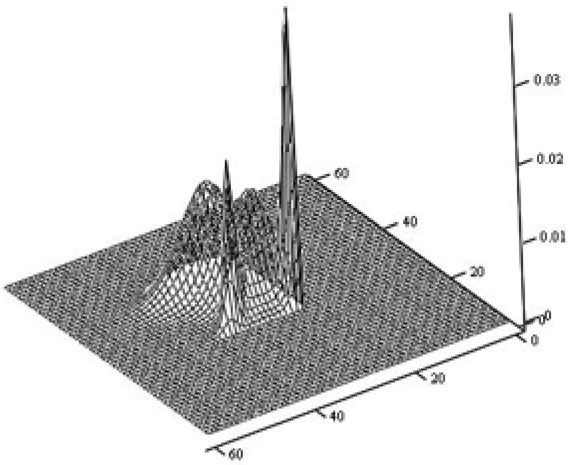


Figure 2. Plot of function f(x) for two-dimensional case

Table 2 contains the results of sequential cycles of the optimization procedure.

TABLE 2. Results of sequential cycles for two-dimensional case

Iteration number	0	1	2	3	4	5
$x_1^{(1)}$	0	0.961	1.099	1.134	1.212	1.190
$x_2^{(1)}$	0	1.545	2.240	1.874	2.072	1.945
$x_1^{(2)}$	3	2.703	2.944	3.130	3.099	3.079
$x_2^{(2)}$	3	0.845	1.444	2.221	1.813	2.085
$x_1^{(3)}$	6	6.115	6.170	6.303	6.361	6.487
$x_2^{(3)}$	6	5.398	5.206	4.889	4.737	4.404
$x_1^{(4)}$	8	8.141	8.194	8.303	8.335	8.381
$x_2^{(4)}$	8	7.251	7.053	6.742	6.640	6.466
D	5.173	4.616	4.391	4.328	4.304	4.296

From the Table we can see how the gradient method improves the criterion of value continually.

Conclusion

A problem of service station arrangement in spatial space is considered. The elaborated algorithm of the problem solution is based on the gradient method. The considered numerical examples show its efficiency. The authors intend to apply the suggested approach to solving the practical arrangement problems.

APPENDIX

Now we will consider the simplest one-dimensional case when $k = 1$ (one station only) and density $f(x)$ and loss function $g_x(\circ)$ are symmetric functions. Let $f(x)$ have a maximal value at the symmetry point $x = m$ and $g_x\left(x^{(1)}\right) = g(x - x^{(1)})$ have minimal value g^* at the symmetry point 0. Now instead of (2) we have the following criteria:

$$D\left(x^{(1)}\right) = \int_{-\infty}^{\infty} g(x - x^{(1)}) f(x) dx = \int_{-\infty}^{\infty} g(m + x - x^{(1)}) f(m + x) dx. \qquad (17)$$

As f is a symmetric function respectively m, $f(m + x) = f(m - x)$, then for the sum of two points $m + x$ and $m - x$, we have the following sum of the integral expression in (17):

$$g(m + x - x^{(1)}) f(m + x) + g(m - x - x^{(1)}) f(m - x) =$$
$$= f(m + x)\left(g(m + x - x^{(1)}) + g(m - x - x^{(1)})\right).$$

The convexity of function g gives us

$$g(m + x - x^{(1)}) + g(m - x - x^{(1)}) = 2\left(\frac{1}{2} g(m + x - x^{(1)}) + \frac{1}{2} g(m - x - x^{(1)})\right) \geq$$
$$\geq 2g\left(\frac{1}{2}\left(m + x - x^{(1)}\right) + \frac{1}{2}\left(m - x - x^{(1)}\right)\right) = 2g(m - x^{(1)}) \geq 2z.$$

The lower limit is obtained if $x^{(1)} = m$. Therefore,

$$D\left(x^{(1)}\right) = \int_{-\infty}^{\infty} g(x - x^{(1)}) f(x) dx = \int_{0}^{\infty} g(m + x - x^{(1)}) f(m + x) dx + \int_{0}^{\infty} g(m - x - x^{(1)}) f(m - x) dx =$$
$$= \int_{0}^{\infty} f(m + x)\left(g(m + x - x^{(1)}) + g(m - x - x^{(1)})\right) dx \geq z,$$

which is obviously clear.

Taking derivative with respect to $x^{(1)}$ and equate one to zero, we get

$$\frac{\partial}{\partial x^{(1)}} D\left(x^{(1)}\right) = -\int \frac{\partial}{\partial x^{(1)}} g(x - x^{(1)}) f(x) dx = \int \frac{\partial}{\partial x^{(1)}} g(m + x - x^{(1)}) f(m + x) dx = 0 .$$

As function g has the minimum at the point 0, and then the derivative from $g(x)$ is negative for $x < 0$ and is positive for $x > 0$. Therefore,

$$\int_{-\infty}^{x^{(1)}-m} \left| \frac{\partial}{\partial x^{(1)}} g(m + x - x^{(1)}) \right| f(m + x) dx = \int_{x(1)-m}^{\infty} \frac{\partial}{\partial x^{(1)}} g(m + x - x^{(1)}) f(m + x) dx.$$

Obviously, the unique solution is $x^{(1)} = m$. Therefore, for optimal value we have the following expression:

$$D(m) = \int_{-\infty}^{\infty} g(x-m) f(x) dx = \int_{-\infty}^{\infty} g(x) f(m+x) dx.$$

References

1. Srivastava, M. S. *Methods of Multivariate Statistics. Wiley Series in Probability and Statistics.* New York: John Wiley & Sons, Inc., 2002.
2. Turkington, D. A. *Matrix Calculation and Zero-One Matrices: Statistical and Econometric Applications*. Cambridge: Cambridge University Press, 2002.

Computer Modelling and New Technologies, 2007, Vol.11, No.1, 38-43
Transport and Telecommunication Institute, Lomonosova 1, Riga, LV-1019, Latvia

PRACTICAL METHOD OF RUIN PROBABILITY CALCULATION FOR FINITE TIME INTERVAL

Andrey Svirchenkov

Transport and Telecommunication Institute
Lomonosova Str. 1, Riga, LV-1013, Latvia
E-mail: Secretary@lateko.lv

A modification of the classical ruin problem is considered. Novelty consists of a consideration of nongomogeneous Poisson flow of claims, arbitrary distribution of claim costs and existence of lower level of necessary capital for any time moment *t*. The problem is to calculate probability that this lower level is not to be passed. A numerical method has been elaborated for the probability evaluation. The considered method is based on Markov's chain theory and Edgeworth expansion for the probabilistic density function.

Keywords: *ruin problem, Edgeworth expansion, Markov's chain, numerical method*

1. Introduction

We consider some modification of the classical *ruin problem* (Grandell (1991), Ross (1992)). An insurance company has *initial capital u*. The claims occur according to *nonhomogeneous Poisson process* with intensity function $\lambda(t)$ [Ross: 24]. The costs of the claims are independent and identically distributed random variables $\{\eta_i, i = 1,2,...,\}$, having *distribution function F*(*x*). The Poisson process and sequence $\{\eta_i, i = 1,2,...,\}$ are assumed to be independent. The *premium income* density function of the company is defined by positive function *c*(*y*) of current capital value *y*.

The *risk process*, *Y*, is defined by

$$Y(t) = \int_0^t c(Y(\varsigma))d\varsigma - \sum_{i=1}^{N(0,t)} \eta_i \,, \tag{1}$$

where *N*(*s*, *t*) is a number of Poisson events that have place in the interval (*s*, *t*), *Y*(0) = *u*, *u* is initial capital.

The *ruin probability till time moment t* for the company having initial capital *u* is defined by

$$\psi(u) = P\{Y(z) < 0 \quad for \ some \ moment \ z \in (0,t)\}. \tag{2}$$

Now let *l*(*t*) be the known function of time *t* that determines lower level of the necessary capital. Then we define *the probability to have overdue payment in the interval* (0, *t*) as

$$\Psi(t,u) = P\{Y(z) < l(z) \ \ for \ some \ moment \ z \in (0,t)\}. \tag{3}$$

Our aim is to calculate this probability.

2. Some Useful Relationships

Let us remind some known relationships. Those will be used below. Let $m(t,t+\Delta) = E(N(t,t+\Delta))$ be an average number of the events of Poisson process on interval (*t*, *t* + Δ). It is well known that

$$E(N(t,t+\Delta)) = m(t,t+\Delta) = \int_t^{t+\Delta} \lambda(z)dz \tag{4}$$

and *N*(*s*, *t*) has the Poisson distribution with this parameter:

$$P\{N(t,t+\Delta)=n\}=\frac{1}{n!}m(t,t+\Delta)^n \exp(-m(t,t+\Delta)), n=0,1,\ldots. \quad (5)$$

Let us remind that

$$X(t)=\sum_{i=1}^{N(0,t)}\eta_i . \quad (6)$$

from formula (1) is said to be a *compound nonhomogeneous Poisson process*.

We generalize the last notation, letting

$$X(t,t+\Delta)=\sum_{i=N(0,t)+1}^{N(t,t+\Delta)}\eta_i . \quad (7)$$

Obviously it is a *total cost of the claims* during interval (t, t + Δ).
Let us rewrite a generating function of the moments for $X(t,t+\Delta)$:

$$M(s;X(t,t+\Delta))=E(\exp(sX(t,t+\Delta))), \quad (8)$$

using a generating function of the moments for η:

$$M(s;\eta)=E(\exp(s\eta)). \quad (9)$$

Repeating proof from [Ross,1992: 22] we can rewrite

$$M(s;X(t,t+\Delta))=\exp\left\{[M(s;\eta)-1]\int_t^{t+\Lambda}\lambda(z)dz\right\}, \quad (10)$$

Now we are able to write an expression for cumulant generating function:

$$K(s;X(t,t+\Delta))=\ln(M(s;X(t,t+\Delta)))=(M(s;\eta)-1)\int_t^{t+\Lambda}\lambda(z)dz . \quad (11)$$

By differentiation of the above, we obtain cumulants of $X(t,t+\Delta)$ distribution:

$$k_r(t,t+\Delta)=\frac{\partial^r}{\partial s^r}K(0;X(t,t+\Delta))=\nu_r m(t,t+\Delta)=\nu_r\int_t^{t+\Lambda}\lambda(z)dz , \quad (12)$$

where $\nu_r=E(\eta^r)=\frac{\partial^r}{\partial s^r}M(0;\eta)$ is the r-th order moment of η.
For example,

$$\mu(t,t+\Delta)=E(X(t,t+\Delta))=k_1(t,t+\Delta t)=\nu_1 m(t,t+\Delta), \quad (13)$$

$$\sigma^2(t,t+\Delta)=D(X(t,t+\Delta))=k_2(t,t+\Delta)=\nu_2 m(t,t+\Delta). \quad (14)$$

These cumulants will be used for approximation of density f and distribution F functions for $X(t,t+\Delta)$. For that we use Edgeworth expansion [Barndorf-Nielsen; Cox, 1989]. For normalized random variable

$$\widetilde{X}(t,t+\Delta)=\frac{1}{\sqrt{D(X(t,t+\Delta))}}(X(t,t+\Delta)-E(X(t,t+\Delta))). \quad (15)$$

the Edgeworth expansion is determined as

$$f\left(x;\widetilde{X}(t,t+\Delta)\right)=\phi(x)\left\{1+\frac{1}{6}\rho_3(t,\Delta)H_3(x)+\frac{1}{24}\rho_4(t,\Delta)H_4(x)+\frac{1}{72}\rho_3^2(t,\Delta)H_6(x)+\ldots\right\}$$

where φ is density function of standard normal distribution, $\{H_i(x)\}$ are Edgeworth polynomials:

$$H_2(x)=x^2-1,\quad H_3(x)=x^3-3x,\quad H_4(x)=x^4-6x^2+3,$$

$$H_5(x)=x^5-10x^3+15x,\quad H_6(x)=x^6-15x^4+45x^2-15,$$

$\{\rho_r\}$ are normalized cumulants:

$$\rho_r(t,\Delta)=k_r\left(\sigma(t,t+\Delta)\right)^{-r}. \tag{16}$$

The Edgeworth expansion for density function of $X(t,t+\Delta)$ is

$$f\left(x;X(t,t+\Delta)\right)=\frac{1}{\sigma(t,t+\Delta)}f\left(\frac{x-\mu(t,t+\Delta)}{\sigma(t,t+\Delta)};\widetilde{X}(t,t+\Delta)\right)=$$

$$=\frac{1}{\sigma(t,t+\Delta)}\phi\left(\frac{x-\mu(t,t+\Delta)}{\sigma(t,t+\Delta)}\right)\left\{1+\frac{1}{6}\rho_3(t,t+\Delta)H_3\left(\frac{x-\mu(t,t+\Delta)}{\sigma(t,t+\Delta)}\right)+\right.$$

$$\left.+\frac{1}{24}\rho_4(t,t+\Delta)H_4\left(\frac{x-\mu(t,t+\Delta)}{\sigma(t,t+\Delta)}\right)+\frac{1}{72}\rho_3^2(t,t+\Delta)H_6\left(\frac{x-\mu(t,t+\Delta)}{\sigma(t,t+\Delta)})\right)+\ldots\right\}. \tag{17}$$

We will use the received expressions to evaluate ruin probability.

3. Suggested Approach

As earlier *Y*(*t*) is a capital at the time moment t. Let $G(x,t)$ be a probability that *Y*(*t*) not greater then *x* and till time *t* overdue payment absents:

$$G(x,t)=P\left\{Y(t)\le x, Y(z)\ge l(z)\quad for\quad z\in(0,t)\right\}.$$

We will consider a described process at the time moments $t_0=0<t_1<\ldots<t_n=t.$

By that we assume the following conditions:

1. A premium for the interval $(t_i\ ,t_{i+1})$ is calculated with respect to initial capital at the time t_i: one equals $c(t_i)(t_{i+1}-t_i)$.

2. A random event “overdue payment” during interval $(t_i\ ,t_{i+1})$ is determined for the final time moment t_{i+1}: $Y(t_{i+1})=Y(t_i)+c(Y(t_i))(t_{i+1}-t_i)-X(t_i,t_{i+1})$ less then $l(t_{i+1})$.

According to our assumptions, sequence $\{Y(t_i)\}$ produces the Markov’s chain. For original time moment $t=0$ we suppose $l(0)<u$ and have $Y(0)=u$:

$$G(x,0)=\begin{cases}0, x\le u,\\ 1, x>u.\end{cases}$$

Other values of $G(x,t_i)$ are calculated by using ordinary Markovian technique:

for t_1:

$$G(x,t_1)=1-F(u-x+c(u)t_1), x\ge l(t_1), \tag{18}$$

for t_{i+1}, i = 1, 2, …; $x > l(t_{i+1})$:

$$G(x, t_{i+1}) = \int_{l(t_i)}^{\infty} (1 - F(z - x + c(z)(t_{i+1} - t_i))) dG(z, t_i). \tag{19}$$

Obviously

$$\Psi(t; u) = 1 - G(t; \infty). \tag{20}$$

4. Numerical Example

Our aim is to illustrate wide possibilities of the considered model. At first we take the following input date: $\lambda(t) = 1$; $c(y) = c\,y = 0.15y$; $l(t) = 5$. Distribution function of claim cost $F(x)$ is described by the Edgeworth expansion for density function (17). Here $m(t, t + \Delta) = E(N(t, t + \Delta)) = \lambda\Delta$ and analysis of statistical date gives the following values of $\nu_r = E(\eta^r)$: $\nu_1 = 2$; $\nu_2 = 12$, $\nu_3 = 48$; $\nu_4 = 384$. It allows getting such values of cumulants $k_r(t, t + \Delta) = \nu_r \lambda\Delta$ and normalized cumulants $\rho_r(t, \Delta) = k_r(\sigma(t, t + \Delta))^{-r}$:

$$\mu(t, t + \Delta) = E(X(t, t + \Delta)) = k_1(t, t + \Delta t) = \nu_1 m(t, t + \Delta) = 2\Delta,$$

$$\sigma(t, t + \Delta) = \sqrt{D(X(t, t + \Delta))} = \sqrt{k_2(t, t + \Delta)} = \sqrt{\nu_2 m(t, t + \Delta)} = 3.464\Delta,$$

$$k_3(t, t + \Delta) = \nu_3 m(t, t + \Delta) = 48\Delta, \; k_4(t, t + \Delta) = \nu_4 m(t, t + \Delta) = 384\Delta,$$

$$\rho_3(t, t + \Delta) = k_3(t, t + \Delta) / \sigma(t, t + \Delta)^3 = 1.155\Delta^{-1/2},$$

$$\rho_4(t, t + \Delta) = k_4(t, t + \Delta) / \sigma(t, t + \Delta)^4 = 2.667\Delta^{-1}.$$

Using this data we wish to calculate values of the probability $\Psi(t; u)$ to have overdue payment in the interval (0, t), see formula (3). For that we use the suggested approach and consider time moments $t_0 = 0$, $t_i = i\Delta$, i = 1, 2, … .We set $\Delta = 1$.

Table 1 contains probabilities $G(x; 9, u)$ that current capital $Y(t)$ at the time moment $t = 9$ not greater then x and till time $t = 9$ overdue payment absents, if initial capital equals u. The last row contains probability $\Psi(9; u) = 1 - G(\infty; t; u)$. Analogous data corresponds to the initial capital values u from 11 till 16.

Table 2 contains analogously probabilities for a case when the lower level of necessary capital at the time moment t is defined as $l(t) = 5 + t$. We see that for this case the probabilities are changed essentially.

TABLE 1. Probabilities $G(x; 9, u)$ and $\Psi(9; u)$ (the last row) as functions of initial capital u ($l(t) = 5$)

x	u = 11	u = 12	u = 13	u = 14	u = 15	u = 16
5	0.009	0.009	0.008	0.008	0.007	0.006
7	0.035	0.035	0.033	0.030	0.027	0.024
9	0.068	0.067	0.064	0.059	0.054	0.047
11	0.105	0.104	0.099	0.092	0.084	0.074
13	0.143	0.142	0.137	0.128	0.117	0.104
15	0.181	0.182	0.176	0.166	0.152	0.136
17	0.219	0.221	0.216	0.205	0.189	0.171
19	0.255	0.260	0.256	0.245	0.228	0.207
21	0.289	0.289	0.296	0.285	0.267	0.244
23	0.321	0.334	0.335	0.325	0.307	0.283

The continuation of Table 1

x	u = 11	u = 12	u = 13	u = 14	u = 15	u = 16
25	0.351	0.368	0.372	0.365	0.348	0.323
27	0.378	0.400	0.408	0.403	0.388	0.363
29	0.401	0.429	0.442	0.441	0.427	0.404
31	0.422	0.455	0.473	0.476	0.466	0.444
33	0.440	0.478	0.501	0.509	0.502	0.483
35	0.454	0.498	0.526	0.539	0.537	0.521
37	0.467	0.514	0.549	0.567	0.570	0.557
39	0.476	0.529	0.568	0.592	0.599	0.592
41	0.484	0.540	0.584	0.613	0.626	0.623
43	0.490	0.549	0.597	0.632	0.650	0.652
45	0.494	0.556	0.608	0.647	0.671	0.678
47	0.497	0.562	0.617	0.660	0.689	0.701
49	0.499	0.566	0.624	0.670	0.703	0.721
Ψ(9; u)	**0.495**	**0.424**	**0.359**	**0.299**	**0.246**	**0.200**

TABLE 2. Probabilities $G(x; 9, u)$ and Ψ(9; u) as functions of initial capital u ($l(t)$ = 5 + t)

x	u = 11	u = 12	u = 13	u = 14	u = 15	u = 16
13	0	0	0	0	0	0
15	0.019	0.021	0.021	0.021	0.020	0.019
17	0.045	0.049	0.051	0.051	0.049	0.046
19	0.076	0.083	0.087	0.087	0.084	0.079
21	0.108	0.119	0.125	0.126	0.123	0.116
23	0.139	0.154	0.163	0.165	0.162	0.154
25	0.168	0.188	0.200	0.205	0.202	0.194
27	0.195	0.219	0.235	0.243	0.242	234
29	0.218	0.248	0.269	0.280	0.282	0.274
31	0.239	0.274	0.300	0.315	0.320	0.314
33	0.256	0.297	0.328	0.348	0.357	0.353
35	0.271	0316	0.353	0.379	0.391	0.391
37	0.283	0.333	0.375	0.406	0.424	0.428
39	0.293	0.347	0.395	0.431	0.453	0.462
41	0.300	0.359	0.411	0.452	0.480	0.494
43	0.306	0.368	0.424	0.471	0.504	0.523
45	0.310	0.375	0.435	0.486	0.525	0.549
47	0.313	0.380	0.444	0.499	0.543	0.572
49	0.315	0.384	0.450	0.509	0.558	0.592
51	0.317	0.387	0.455	0.518	0.570	0.609
53	0.318	0.389	0.456	0.524	0.580	0.624
55	0.318	0.391	0.462	0.528	0.587	0.635
57	0.319	0.391	0.464	0.532	0.593	0.644
65	0.320	0.393	0.467	0.538	0.604	0.664
67	0.320	0.393	0.467	0.538	0.605	0.666
69	0.320	0.394	0.467	0.539	0.606	0.667
71	0.321	0.394	0.467	0.539	0.607	0.668
73	0.321	0.394	0.468	0.539	0.607	0.669
75	0.321	0.394	0.468	0.540	0.607	0.669
77	0.321	0.394	0.468	0.540	0.607	0.670
79	0.321	0.394	0.468	0.540	0.608	0.670
Ψ(9; u)	**0.679**	**0.606**	**0.532**	**0.460**	**0.392**	**0.330**

Conclusion

A modification of the classical ruin problem has been considered. The suggested modifications allow taking into account additional factors. A numerical method of a probability of interest has been elaborated. Examples show that various factors influence on this probability essentially.

References

1. Barndorf-Nielsen, O.E., Cox, D.R. *Asymptotic Methods for Use in Statistics.* London: Chapmen and Hall, 1989. (The translation in Russian)
2. Barndorf-Nielsen, O., Cox, D. *Asymptotic Methods in Mathematical Statistics.* Moscow: Mir, 1999.
3. Grandell, J. *Aspects of Risk Theory*. New York: Springer-Verlag, 1991.
4. Ross, Sh.M. *Applied Probability Models with Optimization Applications.* New York: DoverPublications, INC., 1992.

PART II

STATISTICAL INFERENCES

Computer Modelling and New Technologies, 2007, Vol.11, No.1, 46-56
Transport and Telecommunication Institute, Lomonosova 1, Riga, LV-1019, Latvia

ASYMPTOTIC NORMALITY OF THE INTEGRATED SQUARE ERROR AT THE FIXED PLAN OF EXPERIMENT FOR INDIRECT OBSERVATIONS

***Mikhail S. Tikhov*[1], *Dmitriy S. Krishtopenko*[2]**
***Marina V. Yarochuk*[3]**

Nizhny Novgorod State University
Gagarina av., 23, Nizhny Novgorod, 603600, Russia
E-mail: [1] tikhovm@mail.ru,[2] krisdima@mail.ru, [3] marina.ya@list.ru

The goal of this paper is to establish the asymptotic normality of the L_2-deviation of the kernel distribution function estimator $F_n(x)$ defined by $I_n = \int (F_n(x) - F(x))^2 \omega(x)\,dx$, where $F(x)$ is the unknown distribution function of a random variable X, $\omega(x)$ is the weight function in dose-response dependence on the sample $U^{(n)} = \{(W_i, Y_i), 1 \le i \le n\}$, $W_i = I(X_i < u_i)$ is the indicator of even $(X_i < u_i)$ and Y is a random variable, u_i is fixed values. This result is useful for constructing the test goodness-of-fit for the distribution function $F(x)$.

Keywords: *dose-response dependence, indirect observation, integrated square error*

1. The Nonparametric Estimation of Distribution Function

Let $X_1, X_2, \ldots X_n$ be a random sample with a distribution function $F(x)$. We consider a sample $U^{(n)} = \{(W_i, Y_i), 1 \le i \le n\}$, where $W_i = I(X_i < u_i)$ is the indicator of event $(X_i < u_i)$ and Y_i is charactetistic metering error of u_i. We shall illustrate this case taking $Y_1 = u_i + \varepsilon_i$ so measured error is collided for u_i as additive; $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent distributed random variables with a density $q(x) > 0$, $x \in R^1$, further $\{X_i\}$ and $\{Y_i\}$ are independent distributed random variables. In [1], [4] u_i is treated as inserted for organism nonrandom dose decides beforehand and X_i is treated as minimal working dose, which the organism response begins.

The most nonparametric $U^{(n)}$ – sample estimator of $F(x)$ may be written in the form (see [2], [3]).

$$\widetilde{F}_{\mathrm{n}}(x) = \frac{S_{2n}(x)}{S_{1n}(x)},$$

where

$$S_{1n}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(Y_i - x), \quad S_{2n}(x) = \frac{1}{n}\sum_{i=1}^{n} W_i K_h(\mathrm{Y}_i - x), \tag{1}$$

and $K(\cdot) \ge 0$ is a kernel function, $h = h(n) > 0$ is a sequence of constants ($h \to 0$ and $nh \to \infty$ as $n \to \infty$) and $K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right)$.

We use the following conditions (**A**).

(**A0**) $\max_i |u_i - u_{i-1}| = O(n^{-1})$, as $n \to \infty$.

(**A1**) $K(x) \ge 0$ is a bounded even function on R and $\|K\|^2 = \int K^2(z)dz < \infty$.

(**A2**) $K(x) = 0$ for $x \notin [-1,1]$.

(**A3**) $\int K(x)dx = 1$, $\upsilon^2 = \int z^2 K(z)dz < \infty$.

(**A4**) $\int x^k K(x)dx = 0$, $k = 1,3$.

(**A5**) $f'(x)$ is a continuous function, $\int (f'(x))^2 dx < \infty$ and $f''(x)$ is a bounded function.

(**A6**) $f(x)/F(x)$, $f'(x)f(x)/F(x)$ are bounded integrand and $\int (f'(x))^4 dx < \infty$.

It is known (see [4]), that under conditions (*K*) $F_n(x)$ is asymptotically normal. Observed that, $F_n(x)$ is an asymptotically biased estimator.

In [5] a simple two step estimator of distribution function with zero asymptotic bias and the same asymptotic variance as usual kernel smoother in dependence dose-response is produced.

We investigate behaviour of these statistics in this case. Consider the first $S_{1n}(x)$. Observe that

$$ES_{1n}(x) = \frac{1}{nh}\sum_{i=1}^{n} EK\left(\frac{x-u_i-\varepsilon_i}{h}\right) = \frac{1}{nh}\sum_{i=1}^{n}\int q(y)K\left(\frac{x-u_i-y}{h}\right)dy .$$

On condition **(A0)** this expression is an integral sum, so

$$ES_{1n}(x) = \frac{1}{h}\int dt \int q(y)K\left(\frac{x-t-y}{h}\right)dy + O\left(\frac{1}{n}\right) = \int dt \int q(x-t+zh)K(z)dz + O\left(\frac{1}{n}\right) =$$

$$= \int dt \int \{q(x-t) + q'(x-t)zh + q''(x-t)\frac{z^2h^2}{2} + q'''(x-t)\frac{z^3h^3}{6} + q^{IV}(\zeta)\frac{z^4h^4}{24}\}K(z)dz +$$

$$+O\left(\frac{1}{n}\right) = \int dt \int \{q(x-t) + q'(x-t)zh + q''(x-t)\frac{z^2h^2}{2}\}K(z)dz + O(h^4) + O\left(\frac{1}{n}\right),$$

since $| q^{IV}(x) | \le C_1$.

We notice that $\int q(x-t)dt = 1$ hence it follows that $\int q''(x-t)dt = 0$.

As a result and in conditions (**A1-A3**), we get

$$ES_{1n}(x) = \int q(x-t)dt + O(h^4) + O\left(\frac{1}{n}\right) = 1 + O(h^4) , \text{ if } h = Cn^{-1/5} .$$

As $DS_{1n}(x) \to 0$, so $S_{1n}(x) = 1 + O_p(h^4)$ as $n \to \infty$.

We consider the $S_{2n}(x)$, determined by (1). The expectation value of $S_{2n}(x)$ has got the following presentation as $n \to \infty$

$$E(S_{2n}(x)) = \frac{1}{nh}\sum_{i=1}^{n} F(u_i)EK\left(\frac{x-u_i-\varepsilon_i}{h}\right) \sim \frac{1}{h}\int F(t)dt \int q(y)K\left(\frac{x-t-y}{h}\right)dy =$$

$$= \int F(t)dt \int q(x-t+zh)K(z)dz = \int F(t)dt \int \{q(x-t) + zhq'(x-t) + \frac{z^2h^2}{2}q''(x-t) +$$

$$+ q'''(x-t)\frac{z^3h^3}{6} + q^{IV}(\zeta)\frac{z^4h^4}{24}\}K(z)dz = \int F(t)q(x-t)dt + \frac{\upsilon^2h^2}{2}\int F(t)q''(x-t)dt + o(h^2) .$$

Set $\Phi(x) = \int F(t)q(x-t)dt$, so $\Phi''(x) = \int F(t)q''(x-t)dt$.

Hence

$$ES_{2n}(x) = \Phi(x) + \frac{\upsilon^2h^2}{2}\Phi''(x) + o(h^2) .$$

The variance of $S_{2n}(x)$ has the following asymptotic presentation

$$DS_{2n}(x) = D\{\frac{1}{nh}\sum_{i=1}^{n} W_i K\left(\frac{x-u_i-\varepsilon_i}{h}\right)\} = \frac{1}{n^2h^2}\sum_{i=1}^{n} F(i/n)(1-F(i/n))K\left(\frac{x-u_i-\varepsilon_i}{h}\right) \sim$$

$$\sim \frac{1}{nh^2}\int F(t)(1-F(t))DK\left(\frac{x-t-\varepsilon_i}{h}\right),$$

$$DK\left(\frac{x-t-\varepsilon_i}{h}\right)=\int q(y)K^2\left(\frac{x-t-y}{h}\right)dy-\left(\int q(y)K\left(\frac{x-t-y}{h}\right)dy\right)^2=$$

$$=h\int q(x-t+zh)K^2(z)dz-h^2\left(\int q(x-t+zh)K(z)dz\right)^2\sim$$

$$\sim hq(x-t)\,\|K\|^2-h^2q^2(x-t)+o(h^2)\,.$$

Thus,

$$DS_{2n}(x)=\frac{\|K\|^2}{nh}\int F(t)(1-F(t))q(x-t)dt+O(n^{-1})\,.$$

As $S_{1n}(x)=1+O_p(h^4)$, so the $U^{(n)}$ – sample estimator of the averaging distribution function may be written in the form $F_n(x)=S_{2n}(x)$.

Theorem 1. *Under the conditions (A) and* $h=Cn^{-1/5}$,

$$\sqrt{nh}(F_n(x)-\Phi(x))\xrightarrow[n\to\infty]{d}N(a(x),\sigma^2(x)),$$

where $a(x)=(1/2)C^{5/2}\upsilon^2\Phi''(x)$ *and* $\sigma^2(x)=\|K\|^2\int F(t)q(x-t)dt$.

Proof. You may see the proof of this theorem in [4]. This result follows immediately using the theorem Lindeberg-Feller.

We use the following estimation:

1) Set $h_0=C_2n^{-\alpha}$ and compute

$$\tilde{\varphi}(x)=\frac{1}{nh_0}\sum_{i=1}^{n}W_iK\left(\frac{x-Y_i}{h_0}\right).$$

2) Define $h_1=C_3n^{-1/5}$ and estimate $\beta(x)$ by

$$\tilde{\beta}(x)=\frac{1}{nh_1}\sum_{j=1}^{n}W_jK\left(\frac{x-Y_j}{h_1}\right)\frac{1}{\tilde{\varphi}(Y_j)}.$$

3) Consider next estimator

$$F_n^*(x)=\tilde{\varphi}(x)\tilde{\beta}(x)=\frac{1}{nh_1}\sum_{j=1}^{n}W_jK\left(\frac{x-Y_j}{h_1}\right)\frac{\tilde{\varphi}(x)}{\tilde{\varphi}(Y_j)}.$$

Let

$$\bar{f}_{n,h}=\bar{f}_{n,h}(x)=n^{-1}\sum_{i=1}^{n}K_h(x-U_i)\,.$$

We reduce the next result from [7], which will be necessary to us in the future.

Theorem 2 [7]**.** *Assuming (A) and* $f(x)$ *is a bounded function on* R*, we have for any* $c>0$*, with probability 1:*

$$\varlimsup_{n\to\infty}\ \sup_{c\ln n/n\le h\le 1}\frac{\sqrt{nh}\left\|\bar{f}_{n,h}-E\bar{f}_{n,h}\right\|_\infty}{\sqrt{\max(\ln(1/h),\ln\ln n)}}=k(c)<\infty\,,$$

where $\|K\|_\infty=\sup_x|K(x)|<\infty$.

The following theorem establishes the asymptotic normality of estimator $F_n^*(x)$.

Theorem 3. *Under the assumptions (A) and* $1/10 < \alpha < 1/5$ *,*

$$\sqrt{nh_1}\,(F_n^*(x) - \Phi(x)) \xrightarrow[n\to\infty]{d} N(0, \sigma^2(x)) .$$

Proof. Let $\overline{\varphi}(x) = E(\widetilde{\varphi}(x))$. Then we obviously have

$$\left|\frac{\widetilde{\varphi}(x)}{\widetilde{\varphi}(Y_j)} - \frac{\overline{\varphi}(x)}{\overline{\varphi}(Y_j)}\right| = \left|\frac{\widetilde{\varphi}(x)\overline{\varphi}(Y_j) - \overline{\varphi}(x)\widetilde{\varphi}(Y_j)}{\widetilde{\varphi}(Y_j)\overline{\varphi}(Y_j)}\right| =$$

$$= \left|\frac{\overline{\varphi}(Y_j)[\widetilde{\varphi}(x) - \overline{\varphi}(x)] + \overline{\varphi}(x)[\overline{\varphi}(Y_j) - \widetilde{\varphi}(Y_j)]}{\widetilde{\varphi}(Y_j)\overline{\varphi}(Y_j)}\right| \le$$

$$\le \frac{1}{\widetilde{\varphi}(Y_j)}\left|\widetilde{\varphi}(x) - \overline{\varphi}(x)\right| + \frac{\overline{\varphi}(x)}{\widetilde{\varphi}(Y_j)\overline{\varphi}(Y_j)}\left|\widetilde{\varphi}(Y_j) - \overline{\varphi}(Y_j)\right| .$$

From theorem 2 it now follows that,

$$\Delta_{n1} = \sup_x \left|\widetilde{\varphi}(x) - \overline{\varphi}(x)\right| = O_p\left(\sqrt{\frac{\ln n}{nh_0}}\right),$$

$$\Delta_{n2} = \sup_x \left|\widetilde{\varphi}(Y_j) - \overline{\varphi}(Y_j)\right| = O_p\left(\sqrt{\frac{\ln n}{nh_0}}\right).$$

The last bound obviously implies that $h_0 = C_2 n^{-\alpha}$ ($1/10 < \alpha < 1/5$) and $h_1 = C_3 n^{-1/5}$,

$$\sqrt{nh_1}\,\Delta_{n1} \xrightarrow[n\to\infty]{p} 0 , \qquad \sqrt{nh_1}\,\Delta_{n2} \xrightarrow[n\to\infty]{p} 0 .$$

Therefore we investigate the asymptotic behaviour of sum

$$S_{3n}(x) = \frac{1}{nh_1}\sum_{i=1}^{n} W_i K\left(\frac{x - u_i - \varepsilon_i}{h_1}\right)\frac{\overline{\varphi}(x)}{\overline{\varphi}(u_i + \varepsilon_i)} .$$

First we consider the expectation value of $S_{3n}(x)$.

$$ES_{3n}(x) = \frac{1}{nh_1}\sum_{i=1}^{n} E\{K\left(\frac{x - u_i - \varepsilon_i}{h_1}\right)\frac{\overline{\varphi}(x)}{\overline{\varphi}(u_i + \varepsilon_i)}\}E(I(X_i < u_i)) =$$

$$= \frac{1}{nh_1}\sum_{i=1}^{n} F(u_i)E\{K\left(\frac{x - u_i - \varepsilon_i}{h}\right)\frac{\overline{\varphi}(x)}{\overline{\varphi}(u_i + \varepsilon_i)}\} \sim \frac{1}{h_1}\int F(t)E\{K\left(\frac{x - t - \varepsilon_i}{h_1}\right)\frac{\overline{\varphi}(x)}{\overline{\varphi}(t + \varepsilon_i)}\}dt =$$

$$= \frac{\overline{\varphi}(x)}{h_1}\int F(t)dt\int q(y)K\left(\frac{x - t - y}{h_1}\right)\frac{1}{\overline{\varphi}(t + y)}dy =$$

$$= \int F(t)dt\int K(z)dz\{q(x - t) + q'(x - t)zh_1 + \frac{z^2 h_1^2}{2}q''(x - t) + o(h_1^2)\}\frac{\overline{\varphi}(x)}{\overline{\varphi}(x + zh_1)} .$$

Expanding $\dfrac{\overline{\varphi}(x)}{\overline{\varphi}(x + h_1 z)}$ and $\overline{\varphi}(x)$ in Taylor's serie by $h_0 t$ and calculated their product, we receive on condition (A)

$$\frac{\overline{\varphi}(x)}{\overline{\varphi}(x + h_1 z)} = \frac{\overline{\varphi}(x)}{\overline{\varphi}(x) + O(h_1 h_0^2)} \sim 1 + o(h_1 h_0^2) .$$

Therefore

$$ES_{3n}(x) = \Phi(x) + o(n^{-1/5-2\alpha}) \xrightarrow[n\to\infty]{} \Phi(x).$$

Remark 1. If $f'''(x)/F(x)$, $f(x)f''(x)/F^2(x)$, $f(x)f'(x)/F^2(x)$, $(f'(x))^2/F^2(x)$ are bounded on *R*, then we may demand $0 < \alpha < 1/5$.

Then

$$DS_{3n}(x) = \frac{1}{n^2 h_1^2}\sum_{i=1}^{n} D\{W_i K\left(\frac{x-u_i-\varepsilon_i}{h_1}\right)\frac{\overline{\varphi}(x)}{\overline{\varphi}(u_i+\varepsilon_i)}\}) =$$

$$= \frac{1}{n^2 h_1^2}\sum_{i=1}^{n} E\{W_i K^2\left(\frac{x-u_i-\varepsilon_i}{h_1}\right)\frac{\overline{\varphi}^2(x)}{\overline{\varphi}^2(u_i+\varepsilon_i)}\} - F^2(u_i)(E\{K\left(\frac{x-u_i-\varepsilon_i}{h_1}\right)\frac{\overline{\varphi}(x)}{\overline{\varphi}(u_i+\varepsilon_i)}\})^2 =$$

$$= \frac{1}{n^2 h_1^2}\sum_{i=1}^{n} F(u_i)\int q(y)K^2\left(\frac{x-u_i-y}{h_1}\right)\frac{\overline{\varphi}^2(x)}{\overline{\varphi}^2(u_i+y)}dy - F^2(u_i)(\int q(y)K\left(\frac{x-u_i-y}{h_1}\right)\frac{\overline{\varphi}(x)}{\overline{\varphi}(u_i+y)}dy)^2 =$$

$$= \frac{1}{n^2 h_1}\sum_{i=1}^{n} F(u_i)\int q(x-u_i+zh_1)K^2(z)\frac{\overline{\varphi}^2(x)}{\overline{\varphi}^2(x+zh_1)}dz -$$

$$- F^2(u_i)h_1(\int q(x-u_i+zh_1)K(z)\frac{\overline{\varphi}(x)}{\overline{\varphi}(x+zh_1)}dz)^2 \sim \frac{1}{n^2 h_1}\sum_{i=1}^{n} F(u_i)\int q(x-u_i)K^2(z)dz -$$

$- F^2(u_i)h_1(\int q(x-u_i)K(z)dz)^2$, as $\overline{\varphi}^2(x+h_1 t) \xrightarrow[n\to\infty]{} \overline{\varphi}^2(x)$ uniform convergence for $|t| \le t_0$.

Thus,

$$DS_{3n} \sim \frac{\|K\|^2}{nh_1}\int F(t)q(x-t)dt - \frac{1}{n}\int F^2(t)q^2(x-t)dt =$$

$$= \frac{\|K\|^2}{nh_1}\int F(t)q(x-t)dt + O\left(\frac{1}{n}\right) \sim \frac{\|K\|^2}{nh_1}\int F(t)q(x-t)dt.$$

The result of theorem 2 follows immediately using the theorem Lindeberg-Feller.

2. Integrated Square Error of the Nonparametric Distribution Function Estimation

Let $X_1, X_2, ... X_n$ be a random sample with a distribution function $F(x)$. We consider a sample $U^{(n)} = \{(W_i, Y_i), 1 \le i \le n\}$, where $W_i = I(X_i < u_i)$ is indicator of event $(X_i < u_i)$ and $Y_1 = u_i + \varepsilon_i$ so measured error is collided for u_i as an additive. We shall illustrate this case taking $Y_i = u_i + \varepsilon_i$ so measured error ε_i with a density $q(x) > 0$ is collided for u_i as an additive.

Let $F_n(x) = S_{2n}(x)$ be nonparametric estimator if the distribution function $F(x)$.

Integrated (weighted) square error of estimator $F_n(x)$ is given by

$$I_n = \int (F_n(x) - \Phi(x))^2 w(x)dx = \int (F_n(x) - EF_n(x))^2 w(x)dx + \int (EF_n(x) - \Phi(x))^2 w(x)dx +$$

$$+ 2\int (F_n(x) - EF_n(x))(EF_n(x) - \Phi(x))w(x)dx,$$

where $w(x)$ is a weight function.

Without loss of generality we may assume that $w(x) \equiv 1$. We consider every term of the sum I_n individually.

It is easy to see, that $\int (EF_n(x)-\Phi(x))^2 w(x)dx$ is purely deterministic in character.

(i) $I_{n1} = h^2 n^{-1} J_{n1}$, $J_{n1} = \frac{1}{n^{1/2}h}\sum_{i=1}^{n} Z_{n1i}$, where

$$Z_{n1i} = \frac{1}{h}\int \{W_i K\left(\frac{x-u_i-\varepsilon_i}{h}\right) - EW_i EK\left(\frac{x-u_i-\varepsilon_i}{h}\right)\}\{EF_n(x)-\Phi(x)\}dx .$$

Lemma 1. *Under the conditions (A),*

$$J_{n1} \xrightarrow[n\to\infty]{d} N(0,\sigma_1^2),$$

where $\sigma_1^2 = (1/4)\upsilon^4 \int \Phi(x)(1-\Phi(x))(\Phi''(x))^2 dx < \infty$.

Proof. It is easy to see, that $EZ_{n1i} = 0$. We consider the Lindeberg's condition,

$$\frac{1}{nh^4}\sum_{i=1}^{n} E\{Z_{n1i}^2 I(|Z_{n1i}| > \varepsilon n^{1/2}h^2)\} \le \frac{1}{nh^4(\varepsilon nh^4)}\sum_{i=1}^{n} E\{Z_{n1i}^4 I(|Z_{n1i}| > \varepsilon n^{1/2}h^2)\}$$

$$\le \frac{1}{nh^4(\varepsilon nh^4)}\sum_{i=1}^{n} E(Z_{n1i}^4) .$$

The next step is to derive the relations: $DJ_{n1} \underset{n\to\infty}{\to} \sigma_1^2$ and $EZ_{n1i}^4 = O(h^8)$.

Note that $Z_{n1i} = Y_{n1i} - EY_{n1i}$, where

$$]\, Y_{n1i} = \frac{1}{h}\int W_i K\left(\frac{x-u_i-\varepsilon_i}{h}\right)\{EF_n(x)-\Phi(x)\}dx .$$

Define $t_n^{(k)} = EY_{n1i}^k$ for positive integers *k*. Since $EF_n(x) - \Phi(x) = (1/2)h^2\upsilon^2\Phi''(x) + o(h^2)$ uniformly in x , then

$$t_n^{(k)} = \frac{1}{h^k} E\{W_i \int K\left(\frac{x-u_i-\varepsilon_i}{h}\right)[EF_n(x)-\Phi(x)]dx\}^k =$$

$$= \frac{1}{h^k} F(u_i)\int q(y)dy\{\int K\left(\frac{x-u_i-y}{h}\right)[\frac{\upsilon^2 h^2}{2}\Phi''(x) + o(h^2)]dx\}^k \sim$$

$$\sim \frac{\upsilon^{2k}h^{2k}}{2^k} F(u_i)\int q(y)dy\{\int K(z)\Phi''(y+u_i+zh)dx\}^k \sim \frac{\upsilon^{2k}h^{2k}}{2^k} F(u_i)\int q(y)\{\Phi''(y+u_i)\}^k dy .$$

The result $EZ_{n1i}^4 = O(h^8)$ follows immediately, on noting that

$\mathbf{E}Z_{n1i}^4 = t_n^{(4)} - 4t_n^{(3)}t_n^{(1)} + 6t_n^{(2)}(t_n^{(1)})^2 - 3(t_n^{(1)})^4$.

Further,

$$DJ_{n1} = \frac{1}{nh^2}\sum_{i=1}^{n} EZ_{n1i}^2 \sim \frac{\upsilon^4 h^4}{4}\{\int\int F(t)q(y)(\Phi''(y+t))^2 dtdy -$$

$$-\{\int\int F(t)q(y)\Phi''(y+t)dydt\}^2\} = \frac{\upsilon^4 h^4}{4}\{\int\int F(t)q(u-t)(\Phi''(u))^2 dtdu -$$

$$-\{\int\int F(t)q(u-t)\Phi''(u)dydt\}^2\} = \frac{\upsilon^4 h^4}{4}\{\int \Phi(u)(\Phi''(u))^2 du - \{\int \Phi(u)\Phi''(u)du\}^2\} =$$

$$=\frac{\upsilon^4 h^4}{4}\{\iint \Phi(u)\Phi''(u)\Phi''(v)dudv-\iint \Phi(v)\Phi(u)\Phi''(u)\Phi''(v)dudv\}=$$

$$=\frac{\upsilon^4 h^4}{4}\int \Phi(u)(1-\Phi(u))(\Phi''(u))^2 du\ .$$

Lindeberg-Feller theorem has given the result of Lemma 1.

Now

$$\int (F_n(x)-EF_n(x))^2 \mathrm{dx}=\frac{1}{n^2h^2}\sum_{i=1}^{n}\int \eta_i^2(x)dx+\frac{2}{n^2h^2}\sum_{1\le i<j\le n}\int \eta_i(x)\eta_j(x)dx\ ,$$

where

$$\eta_i(x)=W_iK\left(\frac{x-u_i-\varepsilon_i}{h}\right)-F(u_i)EK\left(\frac{x-u_i-\varepsilon_i}{h}\right).$$

(ii) $I_{n2}=n^{-1}h^{-1}J_{n2}\ ,\ \ J_{n2}\equiv\frac{1}{nh}\sum_{i=1}^{n}\int \eta_i^2(x)dx\ .$

Lemma 2. *Under the conditions (A),*

$$J_{n2}\xrightarrow[n\to\infty]{p}\sigma_2^2\ ,$$

where $\sigma_2^2=\|K\|^2\int \Phi(x)dx\ .$

Proof. Observed, that

$$EJ_{n2}=\frac{1}{nh}\sum_{i=1}^{n}\int E\eta_i^2(x)dx=\ ]$$

$$=\frac{1}{nh}\sum_{i=1}^{n}\int E\{W_iK\left(\frac{x-u_i-\varepsilon_i}{h}\right)\}^2-F^2(u_i)\{EK\left(\frac{x-u_i-\varepsilon_i}{h}\right)\}^2dx=$$

$$=\frac{1}{nh}\sum_{i=1}^{n}\int F(u_i)\{\int q(y)K^2\left(\frac{x-u_i-y}{h}\right)dy\}-F^2(u_i)\{\int q(y)K\left(\frac{x-u_i-\varepsilon_i}{h}\right)dy\}^2dx\sim$$

$$\sim\frac{1}{n}\sum_{i=1}^{n}\int \{F(u_i)\|K\|^2\,q(x-u_i)-hF^2(u_i)q^2(x-u_i)\}dx\sim$$

$$\sim\|K\|^2\int \Phi(x)dx-h\iint F^2(t)q^2(x-t)dxdt\ .$$

Let

$$T_i=\int W_iK\left(\frac{x-u_i-\varepsilon_i}{h}\right)dx\ ,$$

then

$$D(J_{n2})=D\left(\frac{1}{nh}\sum_{i=1}^{n}(T_i-ET_i)\right)^2=\frac{1}{n^2h^2}\sum_{i=1}^{n}D(T_i-ET_i)^2=$$

$$=\frac{1}{n^2h^2}\left(\sum_{i=1}^{n}E(T_i-ET_i)^4-\left(E(T_i-ET_i)^2\right)^2\right)\le\frac{1}{n^2h^2}\sum_{i=1}^{n}E(T_i-ET_i)^4\ .$$

Using the fact that $(a+b)^4 \le 8(a^4+b^4)$ we obtain for large n,

$$D(J_{n2}) \le \frac{8}{n^2h^2}\sum_{i=1}^{n}(ET_i^4+(ET_i)^4) \sim$$

$$\sim \frac{8}{n^2h}\sum_{i=1}^{n}\left(F(u_i)\frac{1}{h}\int K^4\left(\frac{x-u_i-y}{h}\right)q(y)dy\right) \sim \frac{8}{nh}\int K^4(z)dz \xrightarrow[n\to\infty]{} 0,$$

in view of condition (A1).

The result of this lemma follows from the Chebyshev's inequality.

(iii) $\quad I_{n3} = n^{-1}h^{-1/2}J_{n3}$, $J_{n3} = \dfrac{2}{nh^{3/2}}\sum_{1\le i<j\le n}\int \eta_i(x)\eta_j(x)dx$.

Lemma 3. *Under the conditions (A),*

$$J_{n3} \xrightarrow[n\to\infty]{d} N(0,\sigma_3^2),$$

where $\sigma_3^2 = \int\int F^2(x)q^2(y\text{-}x)dxdy\int dv\left(\int K(u)K(u+v)du\right)^2$.

Proof. Define

$$\alpha_{ni} = \eta_i(x)\sum_{j=1}^{i-1}\eta_j(x), \quad \xi_{ni} = \eta_i(x)\sum_{j=i+1}^{n}\eta_j(x).$$

Hence,

$$J_{n3} = \frac{2}{nh^{3/2}}\sum_{i=1}^{n-1}\xi_{ni} = \frac{2}{nh^{3/2}}\sum_{i=2}^{n}\alpha_{ni}.$$

Our aim is to prove a central limit theorem for J_{n3}.

Let $\mathscr{F}_k = \sigma(X_1, X_2, \dots X_k)$ is σ-field, generated by $X_1, X_2, \dots, X_k$. Then $\{\alpha_{nk}, \mathscr{F}_k\}_{1\le k\le n}$, $n \ge 1$, is martingale.

To prove a central limit theorem it suffices that (see [9])

(a) $\quad \dfrac{1}{n^2h^3}\sum_{i=1}^{n-1}E(\alpha_{ni}^2 \mid \mathscr{F}_{i-1}) \xrightarrow[n\to\infty]{p} \sigma_3^2$,

(b) $\quad \dfrac{1}{n^2h^3}\sum_{i=1}^{n-1}E(\alpha_{ni}^2 I(|\,\xi_{ni}| > \delta nh^{3/2})) \xrightarrow[n\to\infty]{p} 0\cdot$

We consider the first (a). Observed, that

$$E(\eta_i(x)\eta_i(y)) = F(u_i)E(K\left(\frac{x-u_i-\varepsilon_i}{h}\right)K\left(\frac{y-u_i-\varepsilon_i}{h}\right)) -$$

$$-F(u_i)E(W_iK\left(\frac{x-u_i-\varepsilon_i}{h}\right))EK\left(\frac{y-u_i-\varepsilon_i}{h}\right) - F(u_i)K\left(\frac{x-u_i-\varepsilon_i}{h}\right)E(W_iK\left(\frac{y-u_i-\varepsilon_i}{h}\right)) +$$

$$+F^2(u_i)EK\left(\frac{x-u_i-\varepsilon_i}{h}\right)EK\left(\frac{y-u_i-\varepsilon_i}{h}\right) \sim F(u_i)h\int q(x-u_i+zh)K(z)K\left(\frac{y-x}{h}-z\right)dz -$$

$$-F(u_i)h^2m(x-u_i)q(y-u_i) - F(u_i)h^2m(y-u_i)q(x-u_i) + F^2(u_i)h^2m(y-u_i)m(x-u_i) \sim$$

$$\sim F(u_i)hq(x-u_i)\int K(u)K\left(\frac{y-x}{h}-u\right)du + h^2F(u_i)(F(u_i)m(x-u_i)m(y-u_i) -$$

$$-m(x-u_i)q(y-u_i) - q(x-u_i)m(y-u_i)).$$

Further observe that

$E(\eta_i(x)\eta_j(y)) = 0$, as $i \neq j$ and $ns_n^2 = \sum_{i=1}^{n-1} E(\xi_{ni}^2) = \sum_{i=2}^{n} E(\alpha_{ni}^2)$.

But

$$E(\xi_{ni}^2) = E(\iint \eta_i(x)\eta_j(y) \sum_{j=i+1}^{n} \eta_j(x) \sum_{k=i+1}^{n} \eta_k(y) dxdy) =$$

$$= \iint E\{\eta_i(x)\eta_i(y) \sum_{j=i+1}^{n} \sum_{k=i+1}^{n} \eta_j(x)\eta_k(y)\} dxdy =$$

$$= \iint E\{\eta_i(x)\eta_i(y) \sum_{j=i+1}^{n} \eta_j(x)\eta_j(y)\} dxdy \sim$$

$$\sim h^2 \iint q(x-u_i)F(u_i) \sum_{j=i+1}^{n} F(u_j)q(x-u_j) \left(\int K(u)K\left(\frac{y-x}{h} - u\right) du \right)^2 dxdy =$$

$$= h^2 \int F(u_i)q(x-u_i) \sum_{j=i+1}^{n} F(u_j)q(x-u_j) dx \int dv \left(\int K(u)K(u+v) du \right)^2 .$$

Therefore as $n \to \infty$,

$$\frac{1}{n^2 h^3} \sum_{i=1}^{n} E(\xi_{ni}^2) = \frac{1}{n^2} \sum_{i=1}^{n-1} F(u_i) \int dv \left(\int K(u)K(u+v) du \right)^2 \sum_{j=i+1}^{n} F(u_j) \times$$

$$\times \int q(x-u_i)q(x-u_j) dx \xrightarrow[n\to\infty]{} \int F(t) \int_t^{+\infty} F(z) \int q(x-t)q(x-z) dxdzdt \int dv \left(\int K(u)K(u+v) du \right)^2 \sim$$

$$\sim \int\int F^2(x)q^2(y-x) dxdy \int dv \left(\int K(u)K(u+v) du \right)^2 = \sigma_3^2 .$$

Since

$$E(\eta_i(x)\eta_i(y)\eta_i(u)\eta_j(v)) \sim$$

$$\sim hF(u_i)q(x-u_i) \int dt (K(t)K\left(\frac{y-x}{h} - t\right)K\left(\frac{u-x}{h} - t\right)K\left(\frac{v-x}{h} - t\right)),$$

then

$E(\eta_i(x)\eta_j(y)\eta_k(u)\eta_l(v)) = 0$, as a set (i, j, k, l) has got if only two different elements.

Then

$$E(\alpha_{ni}^4) = \text{E}(\iint\iint \eta_i(x)\eta_i(y)\eta_i(u)\eta_i(v) \sum_{m=1}^{i-1} \eta_m(x) \sum_{j=1}^{i-1} \eta_j(y) \sum_{k=1}^{i-1} \eta_k(u) \sum_{l=1}^{i-1} \eta_l(v) dxdydudv) =$$

$$= \text{E}(\iint\iint \eta_i(x)\eta_i(y)\eta_i(u)\eta_i(v) \sum_{j=1}^{i-1} \eta_j(x)\eta_j(y)\eta_j(u)\eta_j(v) dxdydudv) \sim$$

$$\sim h^2 \iint\iint F(u_i)q(x-u_i) \sum_{j=1}^{i-1} F(u_j)q(x-u_j) \times$$

$$\times \left(\int K(t)K\left(\frac{y-x}{h} - t\right)K\left(\frac{u-x}{h} - t\right)K\left(\frac{v-x}{h} - t\right) dt \right)^2 dxdydudv \sim$$

$: \; Ah^5 \int F(u_i)q(x-u_i)\sum_{j=1}^{i-1} F(u_j)q(x-u_j)\,dx$,

where $A = \iiint dydudv\left(\int K(t)K(t+y)K(t+u)K(t+v)dt\right)^2$.

Consequently,

$$\frac{1}{n^4h^6}\sum_{i=2}^{n} E(\xi_{ni}^4) = \frac{A}{n^4h}\sum_{i=2}^{n}\int F(u_i)q(x-u_i)\sum_{j=1}^{i-1} F(u_j)q(x-u_j)dx \sim$$

$$\sim \frac{A}{n^2h}\int du \int_{-\infty}^{u} dv \int F(u)q(x-u)F(v)q(x-v)dx = \frac{2A}{n^2h}\int\int F^2(t)q^2(x-t)dxdt \xrightarrow[n\to\infty]{} 0 .$$

Let $V_n^2 = \sum_{i=2}^{n} v_{ni}$, where $v_{ni} = E(\alpha_{ni}^2 \mid \mathscr{F}_{i-1})$. Consequently as $n \to \infty$,

$$E(V_n^4) = 2\sum_{2\le i<j\le n} v_{ni}v_{nj} + \sum_{i=2}^{n} v_{ni}^2 \le C\cdot n^2 A \cdot \iint F^2(t)q^2(x-t)\,dxdt \cdot$$

If we apply the obtained result, we find that

$E(V_n^2 - s_n^2)^2 \le C\cdot nA\cdot \iint F^2(t)q^2(x-t)\,dxdt$,

$s_n^{-4}E(V_n^2 - s_n^2)^2 \to 0$ and $s_n^{-2}V_n^2 \to 1$ in probability as $n \to \infty$, (2)

which proves (a).

The relation (b) follows from inequalities

$$\frac{1}{n^2h^3}\sum_{i=2}^{n} E(\alpha_{ni}^2 I(|\alpha_{ni}| > \delta nh^{3/2})) \le \frac{1}{\delta^2 n^4h^6}\sum_{i=2}^{n} E(\alpha_{ni}^4) \le$$

$$\le \frac{A\cdot C}{\delta^2 n^4 h}\sum_{i=2}^{n}\int F(u_i)q(x-u_i)\sum_{j=1}^{i-1} F(u_j)q(x-u_j)\,dx \le$$

$$\frac{A\cdot C}{\delta^2 n^2 h}\int du \int_0^u dv \int F(u)q(x-u)F(v)q(x-v)\,dx = \frac{2A}{\delta^2 n^2 h}\int \Phi^2(x)\,dx \underset{n\to\infty}{\to} 0 \cdot$$

Finally from (a) and (b) and [13] it follows that J_{n3} is asymptotically normal $N(0,\sigma_3^2)$.

Remark 2. For the Epanechnikov kernel $K(x) = (3/4)(1-x^2)I(|x| \le 1)$ the convolution is $K_2(x) = (K*K)(x) = \begin{cases} (3/360)(32-40x^2+20x^3-x^5), & 0 \le x \le 2 \\ (3/360)(32-40x^2-20x^3+x^5), & -2 \le x < 0 \end{cases}$.

Then

$\int K_2^2(x)dx = \int dv\left(\int K(u)K(u+v)du\right)^2 = 167/387 \approx 0.434$.

(iv) $I_n - \int (EF_n(x) - \Phi(x))^2 dx = 2I_{n1} + I_{n2} + I_{n3}$.

Let

$\mu(n) = \int E(F_n(x) - \Phi(x))^2 dx = \int (EF_n(x) - \Phi(x))^2 dx + n^{-1}h^{-1}\sigma_2^2$.

This means that I_{n1} and I_{n2} are asymptotically independent. Combined the results from Lemmas 1-3, we get the following theorem.

Theorem 4. *Under the conditions (A), $h \to 0$ and $nh \to \infty$ as $n \to \infty$,*

(i) If $nh^5 \to \infty$, then $n^{1/2}h^{-2}(I_n - \mu(n)) \xrightarrow[n\leftarrow\infty]{d} N(0,\sigma_1^2)$.

(ii) If $nh^5 \to 0$*, then* $nh^{-1/2}(I_n - \mu(n)) \xrightarrow[n \leftarrow \infty]{d} N(0, 2\sigma_3^2)$.

(iii) If $nh^5 \to \lambda$, $\lambda \in (0, \infty)$*, then* $nh^{-1/2}(I_n - \mu(n)) \xrightarrow[n \leftarrow \infty]{d} N(0, \lambda^{4/5}\sigma_1^2 + 2\lambda^{-1/5}\sigma_3^2)$.

Thus, in this paper Nadaray-Watson estimators have been considered. Using them we have proved that the integrated square errors of these estimators are asymptotically normal. Also, we have been proved asymptotical normality of the offered asymptotically unbiased estimator (see [5]). Observe that, the result of Theorem 4 takes place also for this unbiased estimator. These results may be used for constructing goodness-of-fit test for this problem.

References

1. Krishtopenko, S. V., Tikhov, M. S. *Toxicometry of Effective Dozes*. Nizhniy Novgorod: University of Nizhniy Novgorod Publishers, 1997. 156 p. (In Russian)
2. Nadaraya, E. A. About Evaluation of Regression, *Probability Theory and Its Appliance,* Vol. 9, No 1, 1964, pp. 157-159. (In Russian)
3. Watson, G. S. Smooth regression analysis, *Sankhya*, Vol. 26, 1964, pp. 359-372.
4. Tikhov, M. S. Statistical Estimation on the Basis of Interval-Censored Data, *J. Math. Sciences*, Vol. 119**,** No 3, 2004, pp. 321-335.
5. Tikhov, M. S., Dolgikh, I. S. Asymptotically not displaced estimations of function of distribution in dependence a doze-effect. In: *Statistical Methods of Evaluation and Checks of Hypotheses: scientific papers*. Perm: University of Perm Publishers, 2005, pp. 60-72. (In Russian)
6. Loev, M. *Probability Theory*. M.: IL, 1962. 720 p. (In Russian)
7. Einmahl, U., Mason, D. M. Uniform in Bandwidth Consistency of Kernel-type Function Estimators, *Ann. Statist.*, Vol. 33**,** No 3, 2005, pp. 1380-1403.
8. Gnedenko, B. V. *Course of the Probability Theory.* M.: Editorial URSS, 2005. 448 p. (In Russian)
9. Liptser, R. Sh., Shiryayev, A.N. *Theory of Martingales*. M.: Nauka, 1986. 512 p. (In Russian)
10. Tikhov, M. S., Krishtopenko, D. S. Asymptotic distribution for integrated square error at the fixed plan of experiment. In: *9th Intern. Vilnius Conf. on Prob. Theor. and Math. Statist.: Abstract of Communications*. Vilnius: Inst. Math., 2006, pp. 312-314.
11. Tikhov, M. S., Krishtopenko, D. S. Evaluation of distributions in dependence a doze-effect at the fixed plan of experiment, *Statistical Methods of Evaluation and Checks of Hypotheses: scientific papers*. Perm: University of Perm Publishers, 2006, pp. 66-77. (In Russian)
12. Hall, P. Central limit theorem for integrated square error of multivariate nonparametric density estimators, *J. Multivar. Analysis,* Vol. 14, 1984, pp. 1-16.
13. Brown, B. M. Martingale central limit theorems, *Ann. Math. Statist.*, Vol. 42, No 1, 1971, pp. 59-66.

Computer Modelling and New Technologies, 2007, Vol.11, No.1, 57-61
Transport and Telecommunication Institute, Lomonosova 1, Riga, LV-1019, Latvia

ON NONPARAMETRIC INTERVAL ESTIMATION OF A REGRESSION FUNCTION BASED ON THE RESAMPLING

Alexander Andronov

Transport and Telecommunication Institute
Lomonosova Str. 1, Riga, LV-1019, Latvia
E-mail: lora@mailbox.riga.lv

A nonparametric regression model $E(Y) = m(x)$ is considered where Y is a dependent variable, x is d – a dimensional vector of independent variables (regressors) and m is an unknown function. A sequence of independent observations (Y_i, x_i) $i = 1, 2, \ldots, n$, is available. Our aim is to construct an upper confidence bound for $m(x)$ that corresponds to probability γ. The resampling approach is used. The suggested method allows calculating true cover probability.

Keywords: *nonparametric regression, interval estimation, resampling*

1. Introduction

We consider nonparametric regression

$$Y = m(x) + \varepsilon, \tag{1.1}$$

where Y is a dependent variable, $m(\circ)$ is an unknown regression function, x is a d-dimensional vector of independent variables (regressors), ε is a random term.

It is supposed that the random term has zero expectation ($E(\varepsilon) = 0$) and variance $Var(\varepsilon) = \sigma^2 w(x)$ where σ^2 is an unknown constant and $w(x)$ is a known weighted function. Furthermore we have a sequence of independent observations (Y_i, x_i), $i = 1, 2, .., n$. On that base we need to construct an upper confidence bound $\widetilde{m}(x)$ for $m(x)$ at the point x corresponding to probability γ:

$$P\{m(x) \le \widetilde{m}(x)\} \ge \gamma. \tag{1.2}$$

Usual way [DiCicco and Efron, 1996] consists of using a consistent and asymptotic normal distributed estimate $\widehat{m}(x)$ of $m(x)$. A final expression contains derivatives $m'(x)$, $m''(x)$ and variance σ^2 that are replaced by the corresponding estimators.

The resampling approach [Wu, 1986], [Andronov and Afanasyeva, 2004] gives an alternative way that can be described as follows. For the fixed point x we take k nearest neighbours $x_1^\bullet, x_2^\bullet, \ldots, x_k^\bullet$ of x among $x_1, x_2, \ldots, x_n$ (in some sense, for example using any kernel function $K_H(x - x_i^\bullet)$, Mahalanobis or other distance):

$$\{x_1^\bullet, x_2^\bullet, \ldots, x_k^\bullet\} = \{x_i : i \in I_c(x)\},$$

where

$$I_c(x) = \{i : x_i \text{ is one of the } k \text{ nearest neighbours of } x \text{ among } \{x_1, x_2, \ldots, x_n\}\}.$$

Now we have sample $(x_1^\bullet, Y_1^\bullet), (x_2^\bullet, Y_2^\bullet), \ldots, (x_k^\bullet, Y_k^\bullet)$ instead of $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$.

Then we derive sample without replacement $\{i_1, i_2, \ldots, i_r\}$ of size r ($r < k$) from set $\{1, 2, \ldots, k\}$, form resample $(x_1^\circ, Y_1^\circ), (x_2^\circ, Y_2^\circ), \ldots, (x_r^\circ, Y^\circ)$, where $x_j^\circ = x_{i_j}^\bullet$ and $Y_j^\circ = Y_{i_j}^\bullet$, and calculate estimate

$\hat{m}(x)$ of our function of interest $m(x)$. Then we return all selected elements into initial samples and we repeat this procedure R times. As a result the sequence of estimators $\hat{m}_1(x), \hat{m}_2(x), ..., \hat{m}_R(x)$ takes place. After ordering we have the sequence $\hat{m}^{(1)}(x), \hat{m}^{(2)}(x), ..., \hat{m}^{(R)}(x)$, where $\hat{m}^{(i)}(x) \le \hat{m}^{(i+1)}(x)$.

Let the number R is selected so that $R\gamma$ is an integer. Then we set $\tilde{m}(x) = \hat{m}^{(R\gamma)}(x)$.

In the presented paper averaging method of estimator $\tilde{m}(x)$ forming is considered. Our main aim is to elaborate a numerical method for cover probability calculation:

$$\Pr_\gamma(x) = P\{m(x) \le \tilde{m}(x)\}. \tag{1.3}$$

It means that we have to know a distribution of the $R\gamma$-th order statistics $\hat{m}^{(R\gamma)}(x)$. This is a main problem that is necessary to be solved.

2. Averaging Method

At first we consider the method of kernel regression estimation [Hardle *et al.*, 2004]. Let $K_H(\circ)$ be any kernel function (Epanechnikov, Quartic and so on). Then Nadaraya-Watson estimator $\hat{m}(x)$ is calculated by the following formula

$$\hat{m}(x) = \frac{1}{\sum_{i=1}^{r} K_H(x - x_i^\circ)} \sum_{i=1}^{r} K_H(x - x_i^\circ) Y_i^\circ, \tag{2.1}$$

where x_i° and Y_i° are a vector of independent variables and dependent variable for the *i*-th elements of the resample, $i = 1, 2, \ldots, r$.

The resampling procedure gives us sequence $\hat{m}_1(x), \hat{m}_2(x), ..., \hat{m}_R(x)$,

$$\hat{m}_j(x) = \frac{1}{\sum_{i=1}^{r} K_H(x - x_i^\circ(j))} \sum_{i=1}^{r} K_H(x - x_i^\circ(j)) Y_i^\circ(j) \tag{2.2}$$

where $x_i^\circ(j)$ and $Y_i^\circ(j)$ are a vector of independent variables and dependent variable for the *i*-th elements of the *j*-th resample, $i = 1, 2, \ldots, r, j = 1, 2, \ldots, R$.

With respect to (1.1) we have:

$$E(\hat{m}(x) | x^\circ(j)) = \frac{1}{\sum_{i=1}^{r} K_H(x - x_i^\circ(j))} \sum_{i=1}^{r} K_H(x - x_i^\circ(j)) m(x_i^\circ(j)),$$

$$Var(\hat{m}(x) | x^\circ(j)) = \frac{\sigma^2}{\left(\sum_{i=1}^{r} K_H(x - x_i^\circ(j))\right)^2} \sum_{i=1}^{r} \left(K_H(x - x_i^\circ(j))\right)^2 w(x_i^\circ(j)),$$

where $x^\circ(j) = (x_1^\circ(j), x_2^\circ(j), ..., x_r^\circ(j))$.

Then

$$E(\hat{m}(x)) = \frac{1}{\binom{k}{r}} \sum_{z \in \Omega} E(\hat{m}(x) | z) = \frac{1}{\binom{k}{r}} \sum_{z \in \Omega} \left(\frac{1}{\sum_{i=1}^{r} K_H(x - z_i)} \sum_{i=1}^{r} K_H(x - z_i) m(z_i) \right), \tag{2.3}$$

where the sums are taken on set Ω of all r-samples $z = (z_1, z_2, ..., z_r)$ without replacement from the set $\{x_1^\bullet, x_2^\bullet, .., x_k^\bullet\}$.

Analogous expression we can to write down for unconditional variance. At first let us calculate the second moment:

$$E(\hat{m}(x)^2) = \frac{1}{\binom{k}{r}} E\left(\left(\sum_{z\in\Omega} \hat{m}(x)\right)^2 |z\right) = \frac{1}{\binom{k}{r}} \sum_{z\in\Omega}\left(\frac{1}{\left(\sum_{i=1}^{r} K_H(x-z_i)\right)^2} E\left(\left(\sum_{i=1}^{r} K_H(x-z_i) Y_i^\circ(j)\right)^2 |z\right)\right) =$$

$$= \frac{1}{\binom{k}{r}} \sum_{z\in\Omega}\left(\frac{1}{\left(\sum_{i=1}^{r} K_H(x-z_i)\right)^2}\left(\sum_{i=1}^{r} K_H(x-z_i)^2\left(\sigma^2 w(z_i) + m(z_i)^2\right) + 2\sum_{i=1}^{r-1}\sum_{j=i+1}^{r} K_H(x-z_i)K_H(x-z_j)m(z_i)m(z_j)\right)\right).$$

Now the variance can be calculated by the following formula

$$Var(\hat{m}(x)) = E(\hat{m}(x))^2 - (E(\hat{m}(x)))^2. \qquad (2.4)$$

Now we need to calculate the covariance between two various estimates $\hat{m}_j(x)$ and $\hat{m}_{j'}(x)$. We have for $j \neq j'$:

$$Cov(\hat{m}_j(x), \hat{m}_{j'}(x)) \overset{d}{=} E\left((\hat{m}_j(x) - m(x))(\hat{m}_{j'}(x) - m(x))\right) = E\left(\hat{m}_j(x)\hat{m}_{j'}(x)\right) - \left(E(\hat{m}(x))\right)^2. \qquad (2.5)$$

Further

$$E\left(\hat{m}_j(x)\hat{m}_{j'}(x)\right) = \left(\binom{k}{r}\right)^{-2} \sum_{z\in\Omega}\sum_{v\in\Omega} E\left(\hat{m}_j(x)\hat{m}_{j'}(x) | z, v\right) =$$

$$= \left(E(\hat{m}_j(x))\right)^2 + \left(\binom{k}{r}\right)^{-2} \sum_{z\in\Omega} \frac{\sigma^2}{\sum_{i=1}^{r} K_H(x-z_i)}\left(\sum_{v\in\Omega}\frac{1}{\sum_{i=1}^{r} K_H(x-v_i)} \sum_{z_m\in z\wedge v}^{r} K_H(x-z_m)^2 w(z_m)\right). \qquad (2.6)$$

Therefore

$$Cov(\hat{m}(x)) = \left(\binom{k}{r}\right)^{-2} \sum_{z\in\Omega} \frac{\sigma^2}{\sum_{i=1}^{r} K_H(x-z_i)}\left(\sum_{v\in\Omega}\frac{1}{\sum_{i=1}^{r} K_H(x-v_i)} \sum_{z_m\in z\wedge v}^{r} K_H(x-z_m)^2 w(z_m)\right). \qquad (2.7)$$

To avoid the computational difficulties, it is possible to consider the following estimate instead of (2.1):

$$\hat{m}(x) = \frac{1}{r}\sum_{i=1}^{r} Y_i^\circ \qquad (2.8)$$

and the corresponding sequence $\hat{m}_1(x), \hat{m}_2(x), ..., \hat{m}_R(x)$.

Expectations, variances and covariance matrix for this sequence of random variables can be determined using the following lemmas.

Lemma 1.

Let $Z_1, Z_2, ..., Z_k$ be independent random variables with expectations $\mu_1, \mu_2, ..., \mu_k$ and variances $\sigma_1^2, \sigma_2^2, ..., \sigma_k^2$. Let $Z_1^\circ, Z_2^\circ, ..., Z_r^\circ$ be a random sample of size *r* from $Z_1, Z_2, ..., Z_k$ without replacement and *S* be their sum: $S = Z_1^\circ + Z_2^\circ + ... + Z_r^\circ$. Then

$$E(S) = \frac{r}{k}\left(\mu_1 + \mu_2 + ... + \mu_k\right), \tag{2.9}$$

$$Var(S) = \frac{r}{k}\sum_{j=1}^{k}\left(\sigma_j^2 + \mu_j^2\frac{k-r}{k}\right) - 2\frac{r(k-r)}{k^2(k-1)}\sum_{j=1}^{k-1}\sum_{i=j+1}^{k}\mu_i\mu_j. \tag{2.10}$$

Lemma 2.

For the conditions of the previous Lemma let the sample $Z_1^\circ, Z_2^\circ, ..., Z_r^\circ$ be returned into the set $\{Z_1, Z_2, ..., Z_k\}$ and the described procedure be repeated, so that we have new sample $Z_1^\bullet, Z_2^\bullet, ..., Z_r^\bullet$ and a corresponding sum $S^\bullet = Z_1^\bullet + Z_2^\bullet + ... + Z_r^\bullet$. Then the covariance between *S* and $S^\bullet$ is calculated by the formula

$$Cov\left(S, S^\bullet\right) = \left(\frac{r}{k}\right)^2\sum_{i=1}^{k}\sigma_i^2. \tag{2.11}$$

In our case $Y_i^\bullet$ and Y_i° play the part of Z_i and Z_i° correspondingly, $\widehat{m}(x)$ is equal to *S*/*r*. Furthermore $\mu_i = m\left(x_i^\bullet\right)$ and instead of σ_i^2 must be $\sigma^2 w\left(x_i^\bullet\right)$.

With respect to the given suppositions, random vector $\left(\widehat{m}_1(x), \widehat{m}_2(x), ..., \widehat{m}_R(x)\right)$ has multi-dimensional symmetric distribution with characteristics determined by (2.3), (2.4), (2.7) or (2.9)-(2.11). Therefore to calculate cover probability (1.3) means to calculate the probability that at last $R(1-\gamma)$ components of vector $\left(\widehat{m}_1(x), \widehat{m}_2(x), ..., \widehat{m}_R(x)\right)$ will be greater than *m*(*x*). For this it is possible to use normal approximation of the distribution. Unfortunately again we are faced with a hard computational problem. Usually for that solving crude Monte Carlo method is used.

APPENDIX

Proof of Lemma 1.

Let $\chi_j = 1$ if the random variable Z_j belongs to the sample $\{Z_1^\circ, Z_2^\circ, ..., Z_r^\circ\}$ and $\chi_j = 0$ otherwise. Of course $\chi_1, \chi_2, ..., \chi_k$ are dependent random variables because $\chi_1 + \chi_2 + ... + \chi_k = r$. We have: $P\{\chi_j = 1\} = r/k$, $P\{\chi_j = 0\} = 1 - r/k$, $E(\chi_j) = P\{\chi_j = 1\} = r/k$,
$Var(\chi_j) = (1 - r/k)\, r/k$, $E(\chi_i\,\chi_j) = P\{\chi_i = 1, \chi_j = 1\} = r(r-1)/(k(k-1))$ for $i \neq j$. Furthermore

$$S = \sum_{i=0}^{k}\chi_i Z_i.$$

Random variables χ_i and Z_i are independent therefore

$$E(S) = \sum_{i=1}^{k}E(\chi_i Z_i) = \sum_{i=1}^{k}E(\chi_i)E(Z_i) = \frac{r}{k}\sum_{i=1}^{k}\mu_i,$$

$$E\left((\chi_i Z_i)^2\right) = E(\chi_i^2)E(Z_i^2) = \frac{r}{k}(\mu_i^2 + \sigma_i^2),$$

$$Var(\chi_i Z_i) = E\big((\chi_i Z_i)^2\big) - \big(E(\chi_i Z_i)\big)^2 =$$

$$= \frac{r}{k}(\mu_i^2 + \sigma_i^2) - \left(\frac{r}{k}\mu_i\right)^2 = \frac{r}{k}\left(\sigma_i^2 + \mu_i^2\left(1 - \frac{r}{k}\right)\right). \qquad \text{(A.1)}$$

Random variables Z_i, Z_j and χ_i χ_j for $i \neq j$ are independent, too, therefore

$$E(\chi_i Z_i \chi_j Z_j) = E(\chi_i \chi_j) E(Z_i) E(Z_j) = \mu_i \mu_j \frac{r(r-1)}{k(k-1)},$$

$$Cov(\chi_i Z_i, \chi_j Z_j) = \mu_i \mu_j \frac{r(r-1)}{k(k-1)} - \mu_i \mu_j \left(\frac{r}{k}\right)^2 = -\mu_i \mu_j \frac{r(k-r)}{k^2(k-1)}. \qquad \text{(A.2)}$$

Formulas (A.1) and (A.2) give formula (2.10).

Proof of Lemma 2.
Let

$$S = \sum_{i=1}^{k} \chi_i Z_i, \; S^{\bullet} = \sum_{j=1}^{k} \chi_j^{\bullet} Z_j .$$

Then

$$Cov(S, S^{\bullet}) = Cov\left(\sum_{i=1}^{k} \chi_i Z_i, \sum_{j=1}^{k} \chi_j^{\bullet} Z_j\right) = \sum_{i=1}^{k}\sum_{j=1}^{k} Cov\big(\chi_i Z_i, \chi_j^{\bullet} Z_j\big) =$$

$$= \sum_{i=1}^{k} Cov\big(\chi_i Z_i, \chi_i^{\bullet} Z_i\big) + \sum_{i=1}^{k}\sum_{j \neq i} Cov\big(\chi_i Z_i, \chi_j^{\bullet} Z_j\big).$$

For $i \neq j$ random variables $\chi_i, \chi_j^{\bullet}, Z_i, Z_j$ are independent, therefore $Cov\big(\chi_i Z_i, \chi_j^{\bullet} Z_j\big) = 0$. Further

$$Cov\big(\chi_i Z_i, \chi_i^{\bullet} Z_i\big) = E\big(\chi_i \chi_i^{\bullet} Z_i^2\big) - E\big(\chi_i Z_i\big) E\big(\chi_i^{\bullet} Z_i\big) =$$

$$= E\big(\chi_i\big) E\big(\chi_i^{\bullet} Z_i\big) E\big(Z_i^2\big) - \left(\frac{r}{k}\mu_i\right)^2 = \left(\frac{r}{k}\sigma_i\right)^2.$$

Therefore

$$Cov\big(S, S^{\bullet}\big) = \left(\frac{r}{k}\right)^2 \sum_{i=1}^{k} \sigma_i^2 .$$

References

1. Andronov, A., Afanasyeva, H. Resampling-based nonparametric statistical inferences about the distributions of order statistics. In: *Transactions of XXIV International Seminar on Stability Problems for Stochastic Models*. Riga: Transport and Telecommunication Institute, 2004, pp. 300-307.
2. DiCicco, T. J., Efron, B. Bootstrap confidence intervals, *Statistical Sciences*, Vol. 11, No 3, 1996, pp. 189-228.
3. Hardle, W., Muller, M., Sperlich, S., Werwatz, A. *Nonparametric and Semiparametric Models*. Berlin: Springer, 2004.
4. Wu, C.F.J. Jackknife, Bootstrap and other resampling methods in regression analysis*, The Annals of Statistics*, Vol. 14, No 3, 1986, pp. 1261-1295.

Computer Modelling and New Technologies, 2007, Vol.11, No.1, 62-72
Transport and Telecommunication Institute, Lomonosova 1, Riga, LV-1019, Latvia

USE OF THE GENERALIZED LINEAR MODEL IN FORECASTING THE AIR PASSENGERS' CONVEYANCES FROM EU COUNTRIES

Catherine Zhukovskaya

Riga Technical University, Faculty of Transport and Mechanical Engineering
Kalku 1, LV-1658, Riga, Latvia
E-mail: *kat_zuk@hotmail.com*

Some regression models to forecast the air passengers' conveyances from EU countries are considered. Two different approaches for the above-mentioned task of forecasting are shown. The first one is the classical method of *linear regression* and the second one is its *generalized* approach. The considered regression models contain many explanatory factors and their combinations. The advantage of using the *generalized linear model* (GLM) in comparison with the classical *linear regression model* is shown.

Keywords: *air passengers' conveyances, forecasting, generalized linear model*

1. Introduction

Most the literature which is devoted to forecasting of transport flows contains only simple forecasting models on the base of the time series methods or linear regression methods with a small number of explanatory variables. Two approaches for the forecasting of air passengers' conveyances from EU countries are considered in this article: the classical method of *linear regression* and its *generalized* approach. The difference of linear regression models considered in this article comparing with the models presented in other papers [6] (autoregression integrated moving average models) and [3, 4, 8] (multiple regression models) consists in using the greater number of the explanatory factors and their combinations. Some models on the base of GLM are considered in the article as well. The aim of this article is to illustrate the advantage of using the GLM comparing with the linear regression models. The verification of the models and the evaluation of the unknown parameters are included in the research as well.

This article has the following structure. The second section contains the description of the informative base of the mentioned investigation. The used models for analyzing and forecasting of air passengers' conveyances are considered in the third section. The elaboration of linear regression models and generalized linear models are presented in the fourth and fifth sections. In the fifth section the advantage of using GLM in comparison with the classical linear regression is shown.

2. Informative Base

In this article the number of carried air passengers was our index of interest and we intend to forecast their volumes. We use the following factors influencing the volumes of air passengers' conveyances:

t_1 total population of the country (TP), millions of inhabitants;
t_2 area of the country (AREA), thousands of km^2;
t_3 density of the country population (PD), number of inhabitants per km^2;
t_4 monthly labour costs (MLC), thousands of euro;
t_5 gross domestic product (GDP) "per capita" in Purchasing Power Standards (PPS) (GDP_PPS);
t_6 gross domestic product (GDP), billions of euro;
t_7 comparative price level (CPL);
t_8 inflation rate (IR);
t_9 unemployment rate (UR);
t_{10} labour productivity per hour worked (LPHW).

The time interval of consideration was the period from 1996 to 2005. We consider the air passengers' conveyances from EU countries. By the moment of data collection there were 25 countries in the EU, such as Belgium, Czech Republic, Denmark, Germany, Estonia, Greece, Spain, France, Ireland, Italy, Cyprus, Latvia, Lithuania, Luxembourg, Hungary, Malta, Netherlands, Austria, Poland, Portugal, Slovenia, Slovakia, Finland, Sweden and the United Kingdom. All data for this investigation have been received from the electronic database "The Statistical Office of the European Communities" (EUROSTAT) [9].

Some of the considered above factors are to be commented:

a) GDP per capita (Latin: *for each head*) in PPS is the value of all final goods and services produced within a nation in a given year divided by the average population for the same year. This volume index of GDP is expressed in relation to the European Union (EU25 = 100).

b) Comparative price level is an index that used for cross-country comparison of price levels. If it is higher/lower than 100 (EU25 = 100), the country concerned is relatively expensive/cheap as compared with the EU average.

For each considered country and for each year we have the volumes of all ten basic factors mentioned above and the volumes of raw conveyances of the air passengers' carried. But during the data gathering we have collided with shortage of data on many countries, especially concerning the new members of EU; therefore the final number of the observation was 161.

The data for the period from 1996 to 2004 have been used for the estimation and forecasting, i.e. for finding of coefficients of the regressional models (140 observations). The data of the 2005 (21 observations) have been used to check out the quality of forecasting, the so-called the cross-validation (CV). Detailed description of CV approach is considered by Diana Santalova in the proceeding article [7].

3. The Used Models for Analyzing and Forecasting of the Air Passengers' Conveyances

The air passengers' conveyances from EU countries were the main *object* of the consideration in our investigation. The data about concrete country for the concrete year were taken as the *observation*. All the considered models were the *group models* [1]. It means that we have the identical regressional model for the various similar objects.

In our research the *linear regression models* and the *generalized regression models* have been used. In the simplest case *the linear regression model* can be expressed in the following form [5]

$$E\left(Y^{(k)}(x)\right) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}, \tag{1}$$

where $Y^{(k)}$ is a dependent variable for the k-th considered model (regressand), $\mathbf{x} = (x_1, x_2, ..., x_d)^T$ is d-dimensional vector of regressors or explanatory variables, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_d)^T$ is a coefficient vector that has to be estimated from observations for $Y^{(k)}$ and $\mathbf{x}$.

The great number of linear regressional models [3, 4, 8] offered in the literature contains small number of the explanatory variables. But just increasing their number does not lead to improving considerably the quality of the regressional models. So the *generalized linear model* can be applied [5]:

$$E\left(Y^{(k)}(x)\right) = G\left\{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\right\}, \tag{2}$$

where $G(\circ)$ is the *known link function* of one dimensional variable.

Firstly, we consider the linear regressional models. After that the generalized linear models are in focus of our research.

4. Elaboration of Linear Models

The big number of various models which differed with structure of factors and their combinations has been constructed and investigated. During the process of the models' construction the received results have been constantly analyzed and the new complementary factors have been added to them. All the considered models in this investigation are the group models.

As the basic criteria to choose the best model, the following ones were selected: *the multiple coefficient of determination* (R^2), *Fisher's criterion* (F), *the sum of the squares of the residuals* (SSRes) and *the sum of the squares of residuals for the cross-validation* (CV SSRes). In additional to these criteria the other ones have been considered as well, in particular the forms of bands of residuals have been analyzed. Let us note that for the checking of the statistical hypotheses we always used the statistical significance level $\alpha = 0.05$.

In the models 1-3 as the regressand $Y^{(i)} = y$ (where $i = 1, 2, 3$) the number of raw air passengers' carried was taken.

As the regressors in Model 1 all the considered above variables without their modification such as: $x_1 = t_1$, $x_2 = t_2$, $x_3 = t_3$, $x_4 = t_4$, $x_5 = t_5$, $x_6 = t_6$, $x_7 = t_7$, $x_8 = t_8$, $x_9 = t_9$ and $x_{10} = t_{10}$ were chosen.

Model 1 gives the following estimate for $E(Y)$:

$$\hat{E}\left(Y^{(1)}(x)\right) = 14 - 0.77x_1 + 0.16x_2 + 185.8x_3 - 2.44x_4 + 0.53x_5 + 0.07x_6 + 0.05x_7 + 0.32x_8 - $$
$$- 1.2x_9 - 1.03x_{10}.$$

Now we are going to consider some criteria for the used model. The value of the coefficient of determination for this model $R^2 = 0.831$ is high enough. The value of the Fisher criterion $F = 63.49$ is considerably higher than its critical value $F_k = 1.905$. This value was calculated with the significance level $\alpha = 0.05$ and with the degrees of freedom $df1 = m = 10$ and $df2 = (n - m - 1) = 129$, where m is a number of predicted values and n is a number of observations [2]. The critical level of model significance (*p-level*) p = 0.000000, so this model is adequate.

For each factor of model 1 table 1 contains factor estimates (b) and the results of the check of their significance: the calculated values of the Student *t*-criterion (t) and p-level. Some factors for this model are nonsignificant. The critical value of the 2-tailed Student criterion $t_k = 1.979$ which was calculated with the significance level $\alpha = 0.05$ and with the degrees of freedom $df = (n - m - 1) = 129$ [2]. The significant explanatory variables are the variables x_2, x_3, x_6 and x_{10}, so, the greatest influence on the air passengers' conveyances is provided by the aria of the country, the density of the country population, the value of the gross domestic product and the comparative price level. The positive and the negative signs for all regressors in this model correspond to physical sense of regressors. Such statistical procedure was used for all linear regression models considered below.

TABLE 1. The estimates of the coefficients and their significance level for Model 1

Variable	Factor	b	t(129)	p-level
	Intercept	14.00	0.84	0.405
x_1	TP	-0.77	-1.56	0.121
x_2	AREA	0.16	5.60	0.000
x_3	PD	185.80	4.67	0.000
x_4	MLC	-2.44	-0.44	0.660
x_5	GDP_PPS	0.53	1.68	0.096
x_6	GDP	0.07	3.81	0.000
x_7	CPL	0.05	0.37	0.710
x_8	IR	0.32	0.29	0.771
x_9	UR	-1.20	-1.59	0.114
x_{10}	LPHW	-1.03	-3.75	0.000

The analysis of the form of the band of residuals for Model 1 has shown the necessity of adding into the regression model the new explanatory factor t_{11} (ON). This factor takes 2 meanings: "0" if the considered country is the old member of EU, and "1" if the considered country is the new one. Additionally we remove some nonsignificant factors from Model 1.

Therefore the regressors in Model 2 are the following: $x_1 = t_2$, $x_2 = t_3$, $x_3 = t_6$ $x_4 = t_{10}$ and $x_5 = t_{11}$.

Model 2 gives the following estimate for $E(Y)$:

$$\hat{E}\left(Y^{(2)}(x)\right) = 13.56 + 0.09x_1 + 134.01x_2 + 0.05x_3 - 0.68x_4 + 29.36x_5.$$

The obtained results for Model 2 are shown in the Table 2. We see that our modification allows improving some characteristics of regressional Model 1. In this model $R^2 = 0.829$ but despite the decrease of it we see that the value of $F = 129.85$ has considerably increased comparing with the previous model.

TABLE 2. The estimates of the coefficients and their significance level for Model 2

Variable	Factor	b	t(134)	p-level
	Intercept	13.56	2.45	0.016
x_1	AREA	0.09	4.45	0.000
x_2	PD	134.01	4.32	0.000
x_3	GDP	0.05	10.34	0.000
x_4	LPHW	-0.68	-5.12	0.000
x_5	ON	29.36	4.21	0.000

The next step for improving the characteristics of the regressional models consists in adding to regressional models the modified basic factors and their different combinations, such as: $\sqrt{t_1}$, t_1^2, $\sqrt{t_2}$, t_2/t_1, $\sqrt{t_2}/t_1$, t_6/t_1, $t_6/(t_1 \cdot t_2)$, $t_6/(t_1 \cdot \sqrt{t_2})$. We begin with Model 3, where: $x_1 = t_3$, $x_2 = t_6$, $x_3 = t_{10}$, $x_4 = t_1^2$ and $x_5 = \sqrt{t_2}$.

This model gives the following estimate for $E(Y)$:

$$\hat{E}\left(Y^{(3)}(x)\right) = -6.34 + 113.26x_1 + 0.14x_2 - 0.52x_3 - 0.03x_4 + 3.03x_5 .$$

The analysis of the obtained results for Model 3 (Table 3) also shows the rightful appliance of this approach because it allows us to improve considerably the characteristics of the regressional model. Moreover, the comparison of the received results with the results which have been obtained for the models considered above shows that their input allows to improve two characteristics of the regression model at the same time: $R^2 = 0.867$ and $F = 174.078$.

TABLE 3. The estimates of the coefficients and their significance level for Model 3

Variable	Factor	b	t(134)	p-level
	Intercept	-6.34	-1.05	0.296
x_1	PD	113.26	4.00	0.000
x_2	GDP	0.14	10.66	0.000
x_3	LPHW	-0.52	-5.80	0.000
x_4	sq(TP)	-0.03	-7.56	0.000
x_5	sqrt(AREA)	3.03	5.74	0.000

The observed and predicted values of the air passengers' conveyences in Country-Year order for Model 3 are shown on Figure 1, the results of the cross-validation for this model are shown on Figure 2. The "Country-Year order" for all values which is shown on Figure 1 means that, firstly, they are sorted by the country name and for each country they are sorted by year.

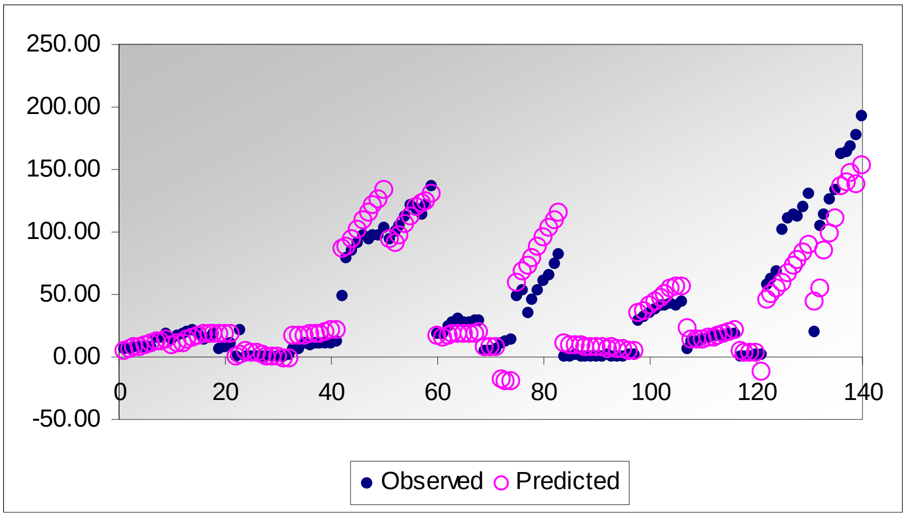


Figure 1. Plot of the observed and predicted values for Model 3

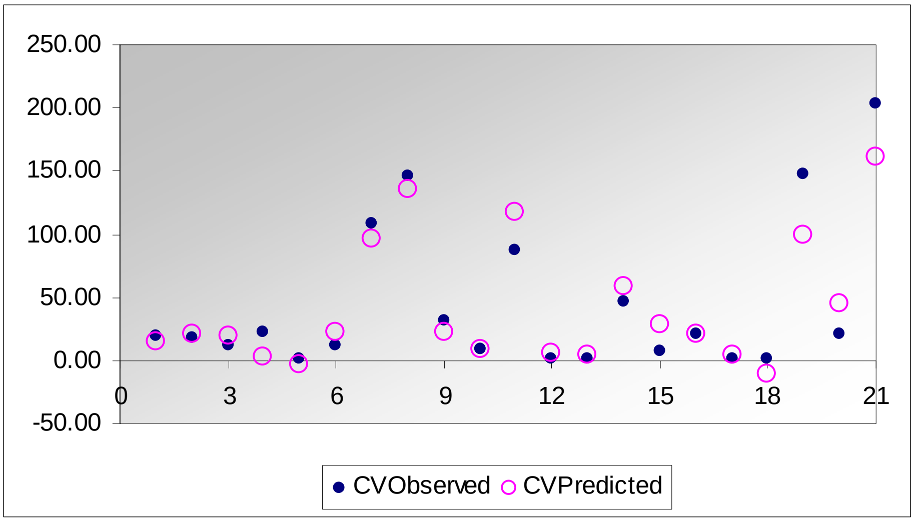


Figure 2. Plot of the observed and predicted values for the CV for Model 3

But in Figures 1 and 2 we can see considerable inconvenience of the 3 Model, which consists in the fact that some predicted values for this model, lies in the negative area. Models 1 and 2 have the same disadvantage. Therefore as the next step for the improving of the regressional model was the transfer to the new forecasting variables.

So in the Models 4-5 as the regressand we considered the ratio between the total number of air passenger carried and the number of inhabitants of the country $Y^{(i)} = y/t_1$, $i = 4, 5$.

As the regressors in Model 4 we used the following variables: $x_1 = t_2$, $x_2 = t_3$, $x_3 = t_4$, $x_4 = t_6$, $x_5 = t_{11}$, $x_6 = \sqrt{t_1}$, $x_7 = \sqrt{t_2}$, $x_8 = t_2/t_1$, $x_8 = \sqrt{t_2}/t_1$ and $x_9 = t_6/t_1$.

Model 4 gives the following estimate for $E(Y)$:

$$\hat{E}\left(Y^{(4)}(x)\right) = 0.56 + 2.33x_1 - 1.04x_2 - 0.02x_3 + 0.001x_4 + 1.76x_5 - 0.0004x_6 + 0.04x_7 + 0.17x_8.$$

The obtained results for Model 4 are shown in the Table 4 (R^2= 0.760, F = 45.81).

TABLE 4. The estimates of the coefficients and their significance level for Model 4

Variable	Factor	b	t(131)	p-level
	Intercept	-5.67	-6.25	0.000
x_1	AREA	-0.02	-6.73	0.000
x_2	PD	10.37	6.19	0.000
x_3	MLC	-0.73	-4.19	0.000
x_4	ON	0.83	8.30	0.000
x_5	sqrt(TP)	-1.02	-7.32	0.000
x_6	sqrt(AREA)	1.06	7.10	0.000
x_7	AREA/TP	-0.12	-6.98	0.000
x_8	sqrt(AREA)/TP	0.94	5.84	0.000
Variable	Factor	b	t(131)	p-level

The analysis of the obtained results doesn't show the improvement of the characteristics of this model. Therefore we decide to enter one more variable t_{12} (HL), which expresses the relative value of the conveyances. It takes 2 meanings: 0 if the value of y/t_1 for the considered country is small (less than 2) and is equal 1 if this value is larger than 2

As the regressors in Model 5 we used the following variables: $x_1 = t_4$, $x_2 = t_5$, $x_3 = t_8$, $x_4 = t_9$, $x_5 = t_{10}$, $x_6 = t_{11}$, $x_7 = t_{12}$ and $x_8 = t_6/t_1$.

Model 5 gives the following estimate for $E(Y)$:

$$\hat{E}\left(Y^{(5)}(x)\right) = 0.99 - 0.46x_1 - 0.02x_2 - 0.02x_3 - 0.02x_4 + 0.01x_5 + 1.27x_6 + 1.15x_7 + 0.07x_8 .$$

The obtained results for Model 5 are shown in the Table 5 ($R^2 = 0.864$, $F = 104.174$).

TABLE 5. The estimates of the coefficients and their significance level for Model 5

Variable	Factor	b	t(131)	p-level
	Intercept	0.99	3.93	0.000
x_1	MLC	-0.46	-3.41	0.001
x_2	GDP_PPS	-0.02	-3.81	0.000
x_3	IR	-0.02	-1.33	0.187
x_4	UR	-0.02	-1.90	0.056
x_5	LPHW	0.01	3.72	0.000
x_6	ON	1.27	9.21	0.000
x_7	HL	1.15	15.30	0.000
x_8	GDP/TP	0.07	3.41	0.001

The data for all considered models and for the four mentioned above criteria have been brought in the Table 6. For each model and for each criterion the rank R_i (where $i = 1, 2, 3, 4$) has been calculated. Here, the rank of a model is its i-th criterion position number among all values of this criterion. The sum of the ranks (Sum R) for all four criteria and the total rank (Total R) has allowed us to define the best model. In order to compare the results obtained for the Models 4-5 with the previous ones (Models 1-3), for the Models 4-5 the recalculated data for SSRes and CV SSRes have been used. These data were multiplied by the value of the country population. So according to the sum of ranks for all considered models and taking into account the inconvenience of the first three models we can conclude that the best model is Model 5.

TABLE 6. Pivot results for the first three models

Model	R^2	R_1	F	R_2	SSRes	R_3	CV SSRes	R_4	Sum R	Total R
Model 1	0.8311	3	63.49	4	52651.33	5	17232.75	5	17	5
Model 2	0.8289	4	129.85	2	53343.53	5	16458.41	4	15	3
Model 3	0.8666	1	174.1	1	41598.60	2	7417.482	1	5	1
Model 4	0.7603	5	45.81	5	35064.04	3	8596.43	3	16	4
Model 5	*0.8642*	*2*	*104.2*	*3*	*12774.59*	*1*	*7717.23*	*2*	*8*	*2*

Figure 3 shows the recalculated observed and predicted values for the air passengers' conveyances for Model 5 in the order Country-Year. The result of the cross-validation for this model is shown in Figure 4.

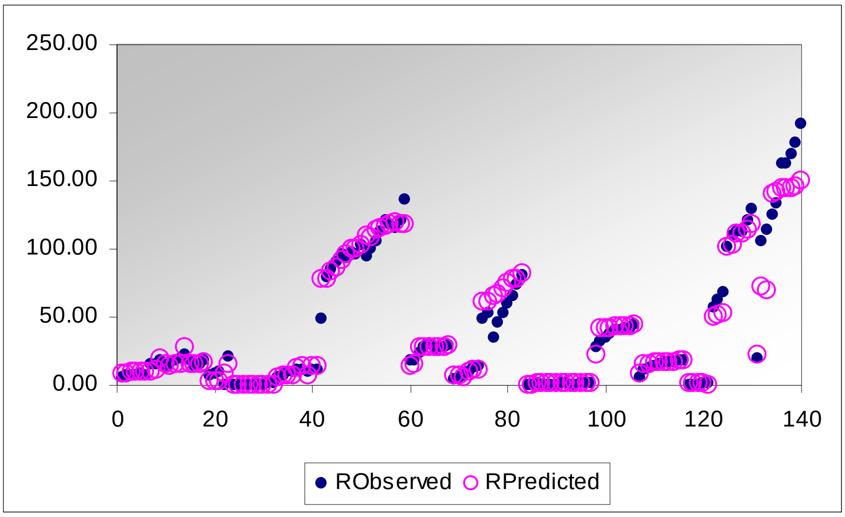


Figure 3. Plot of the recalculated observed and predicted values for Model 5

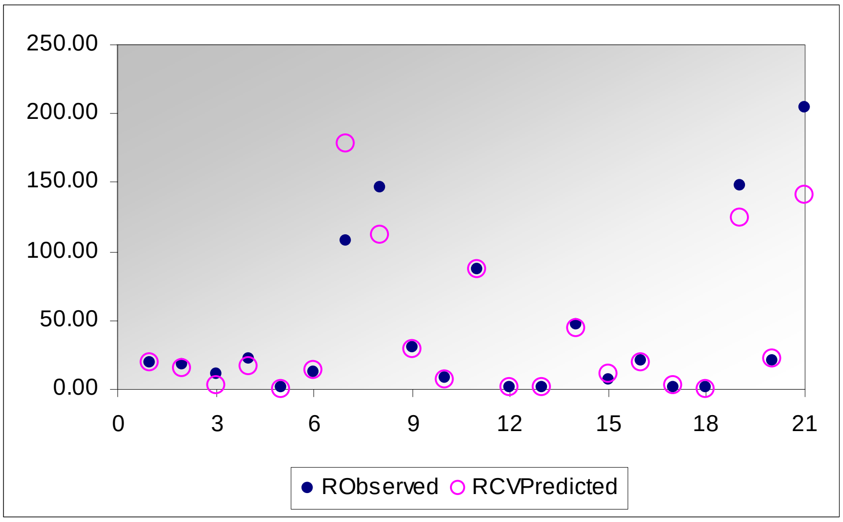


Figure 4. Plot of the recalculated observed and predicted values for the CV for Model 5

5. Elaboration of Generalized Linear Models

Tempting to improve the characteristics of the received linear regression models (1) the generalized linear model (2) has been used.

For the further investigation the best linear regression model (Model 5) has been chosen. Two different GLM are considered. In both of them the value of the regressand $Y^{(GLM)} = Y^{(5)}/t_1$ and the collection of the regressors are the same as for Model 5.

The first GLM (GLM1) is *the modification of the logit model* [5] which can be written in the following form:

$$E\left(Y^{(GLM1)}(x_i)\right) = h_i \frac{\exp\left(\sum_j \beta_j x_{i,j}\right)}{1 + \exp\left(\sum_j \beta_j x_{i,j}\right)}, \tag{3}$$

where h_i is the total population number, x_i is vector-columns of the independent variables, i is the observation number, $i = 1, 2, ..., n$.

The second of the investigated GLM (GLM2) has the following form:

$$E\left(Y^{(GLM2)}(x_i)\right) = h_i \frac{1}{a + \exp\left(\sum_j \beta_j x_{i,j}\right)}, \tag{4}$$

where a is the additional parameter (constant).

The unknown parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_d)^T$ for both GLM is estimated by the use of the least squares (LS) criterion:

$$R_0(\beta) = \sum_{i=1}^{n} \left(Y_i - \tilde{Y}_i\right)^2 \to \min_{\beta}, \tag{5}$$

where Y_i and $\tilde{Y}_i$ are observed and calculated values of Y.

Note, that *linearization* of the logistics models is the traditional way for the estimation of their unknown parameters. Let us show that it gives bad results.

After GLM linearization their liniarizated forms (LM) LM1 for the model (3) and LM2 for the model (4) correspondingly were obtained:

for the model (3) the LM1

$$\ln \frac{Y^*}{1-Y^*} = \sum_j \beta_j x_{i,j} \,, \tag{6}$$

and for the model (4) the LM2

$$\ln\left(\frac{1}{Y^*} - a\right) = \sum_j \beta_j x_{i,j} \,, \tag{7}$$

where $Y^* = Y/h$.

Here the dependent variables are situated in the left side of the formulas, and their parameters' β estimation is not difficult to do.

The models LM1 and LM2 gives the following estimate for $E(Y)$:

$$\hat{E}\left(Y^{(LM1)}(x)\right) = h\frac{e^{-13.78+0.001x_1-6.68x_2-0.02x_3+0.7x_4+48.8x_5-0.44x_6+0.29x_7+7.81x_8-0.64x_9}}{1+e^{-13.78+0.001x_1-6.68x_2-0.02x_3+0.7x_4+48.8x_5-0.44x_6+0.29x_7+7.81x_8-0.64x_9}},$$

$$\hat{E}\left(Y^{(LM2)}(x)\right) = h\frac{1}{0.3+e^{11.65+1.63x_1-1.7x_2+0.04x_3-0.81x_4-17.96x_5-1.67x_6+0.2x_7+0.41x_8-0.11x_9}}.$$

The values of SSRes and CV SSRes for the Models LM1 and LM2 comparing with Model 5 are calculated and shown in the Table 7.

TABLE 7. The value of SSRes and CV SSRes for the Models 5, LM1 and LM2

	SSRes			CV SSRes		
	Model 5	LM1	LM2	Model 5	LM1	LM2
R_0/n	12 775	27 447	21 834	7 717	676 576	229 554

We can see that linearization gives bad results. Making attempts to improve the obtained results a two-stage estimation procedure is developed. The first stage corresponds to the considered linearization. As the second step we use the procedure of *calibration* when we precise the gotten estimates by using the well-known gradient method.

The gradients with the respect to the unknown parameter vector $\boldsymbol{\beta}$ for the GLM1 and GLM2 can be written in the following forms:
for the model (3)

$$\nabla R(\beta) = -2\sum_{i=1}^{n-1}\left(Y_i - h_i\frac{\exp\left(\sum_j \beta_j x_{i,j}\right)}{1+\exp\left(\sum_j \beta_j x_{i,j}\right)}\right)\cdot h_i \cdot \frac{\exp\left(\sum_j \beta_j x_{i,j}\right)}{\left(1+\exp\left(\sum_j \beta_j x_{i,j}\right)\right)^2}\cdot x_i \,, \tag{8}$$

and for the model (4)

$$\nabla R(\beta) = 2\sum_{i=1}^{n-1}\left(Y_i - h_i\frac{1}{a+\exp\left(\sum_j \beta_j x_{i,j}\right)}\right)\cdot h_i \cdot \frac{\exp\left(\sum_j \beta_j x_{i,j}\right)}{\left(a+\exp\left(\sum_j \beta_j x_{i,j}\right)\right)^2}\cdot x_i \,. \tag{9}$$

For the GLM2 we found the optimum value of R_0 not only from the values β but from the parameter a also. Firstly, we fix the value of a and search the optimum values of β according to (5). After that we fix the founded values of β and search the optimum values of R_0 by changing the value of parameter a. This procedure has been repeated many times until R_0 reaches its minimum.
The GLM1 and GLM2 have the following estimates for $E(Y)$:

$$\hat{E}\left(Y^{(GLM1)}(x)\right)=h\frac{e^{-7.05-1.05x_1+1.22x_2-0.02x_3+0.76x_4+5.77x_5+1.26x_6-0.11x_7-0.68x_8+0.15x_9}}{1+e^{-7.05-1.05x_1+1.22x_2-0.02x_3+0.76x_4+5.77x_5+1.26x_6-0.11x_7-0.68x_8+0.15x_9}},$$

$$\hat{E}\left(Y^{(GLM2)}(x)\right)=h\frac{1}{6.3+e^{7.26+1.09x_1-0.78x_2+0.02x_3-0.82x_4-7.81x_5-1.12x_6+0.1x_7+0.13x_8-0.06x_9}}.$$

Figure 5 shows the observed values of air passengers' conveyances and predicted values obtained by using of the generalized linear regression models GLM1 and GLM2 in the order to Country-Year. The results of the cross-validation for the models GLM1 and GLM2 are shown in Figure 6.

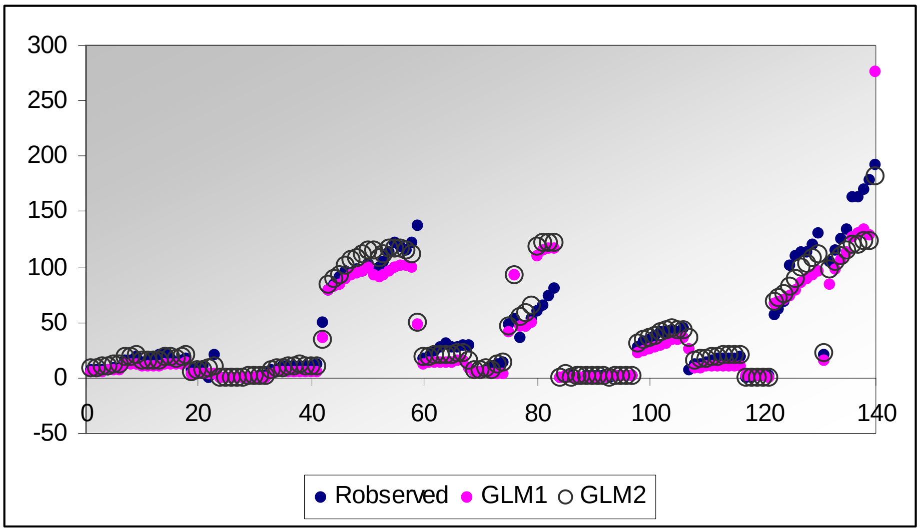


Figure 5. Plot of the observed and predicted values for the GLM1 and GLM2

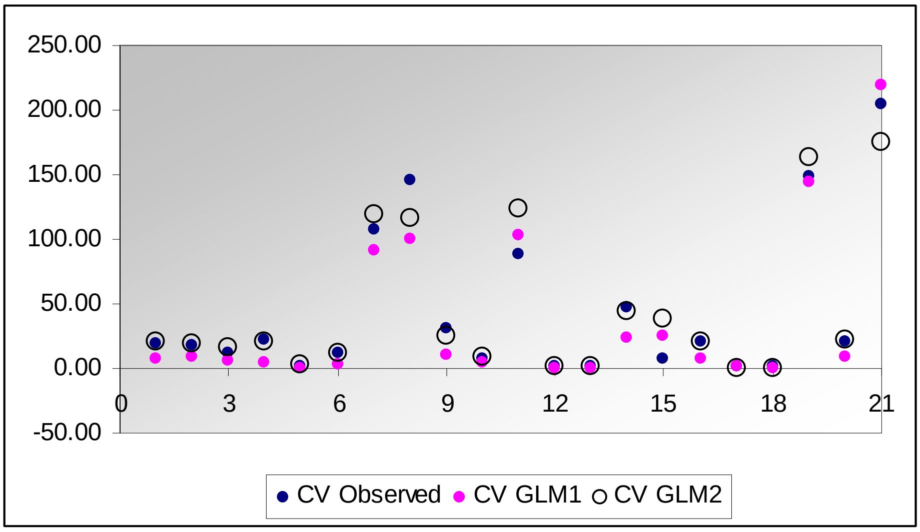


Figure 6. Plot of the observed and predicted values for the CV for the GLM1 and GLM2

The SSRes for the CV for the Models 5, GLM1 and GLM2 are shown in the Table 8.

TABLE 8. The value of SSRes for the CV for the Models 5, GLM1 and GLM2

	CV SSRes		
	Model 5	GLM1	GLM2
R_0/n	7 717	7 171	5 185

The comparison of the data from Tables 7 and 8 allows stating the following: generalized linear models give better results for the forecasting of air passengers' conveyances in comparison with traditional linear regression models; simple linearization gives considerably worse results for the forecasting and needs in its optimization. For this purpose the two-stage estimation procedure which is shown in this article can be used.

Besides this for the GLM2 the dependence of values SSRes and CV SSRes on the value of parameter α is investigated. The obtained results are shown in Figure 7. The optimal value for analysis of SSRes is obtained, then $\alpha = 2$, but the best result for the analysis of CV SSRes is obtained, then $\alpha = 6$.

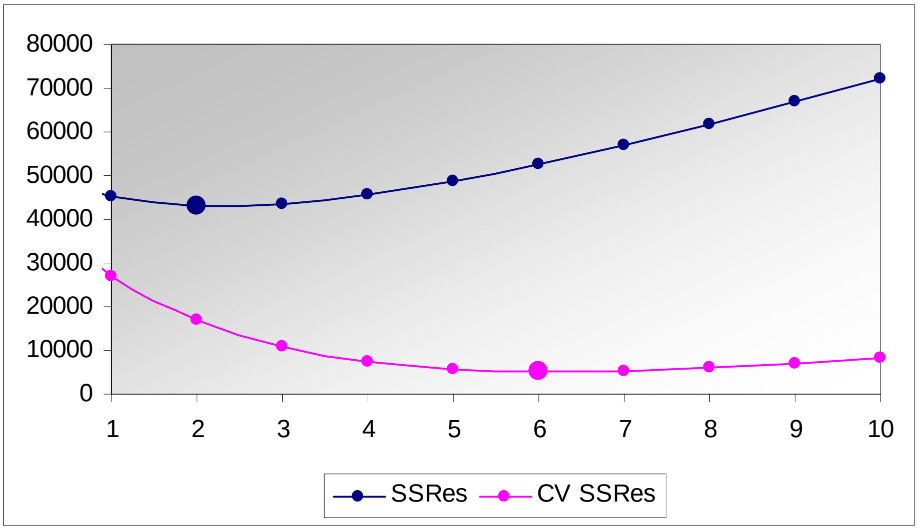


Figure 7. The values of SSRes and CV SSRes as a function of parameter α for GLM 2

Conclusion

The linear and generalized linear regressional models for the forecasting of air passengers'' conveyances from EU countries are considered. These models contain a big number of explanatory factors and their combinations. For the estimation of the unknown parameters of the linear regressional models we use the standard procedures. For the estimation of unknown parameters of GLM the special two-stage procedure has been elaborated. The cross-validation approach has been taken as the main procedure for the check out the adequacy of all considered models and choosing the best model for the forecasting. The advantage of GLM application has been shown.

References

1. Andronov, A. M. etc. *Forecasting of air passengers' conveyances on the transport.* Moscow: Transport, 1983. (In Russian)
2. Andronov, A. M, Kopytov, E. A and Gringlaz, L. J. *The probability theory and the mathematical statistics.* StPetersburg: Piter, 2004. (In Russian)
3. Butkevicius, J., Mazura, M., Ivankovas, V., Mazura, S. Analysis and forecast of the dynamics of passenger transportation by public land transport, *Transport*, Vol. XIX, No 1, 2004, pp. 3-8.
4. Butkevicius, J., Vyskupaitis, A. Development of passenger transportation by Lithuanian sea transport. In: *Proceedings of International Conference RelStat'04, Transport and Telecommunication,* Vol. 6, No 2, 2005.

5. Hardle, W., Muller, M., Sperlich, S., Werwatz, A. *Nonparametric and Semiparametric Models*. Berlin: Springer, 2004.
6. Hunt, U. Forecasting of railway freight volume: approach of Estonian railway to arise efficiency, *Transport*, Vol. XXVIII, No 6, 2003, pp.255-258.
7. Santalova, D. Forecasting of Rail Freight Conveyances in EU Countries on the Base of the Single Index Model. (In printing)
8. Sliupas, T. Annual average daily traffic forecasting using different techniques, *Transport*, Vol. XXI, No 1, 2006, pp.38-43.
9. *EUROSTAT YEARBOOK 2005. The statistical guide to Europe. Data 1993-2004. EU, EuroSTAT, 2005. URL –* http://epp.eurostat.ec.europa.eu

Computer Modelling and New Technologies, 2007, Vol.11, No.1, 73-83
Transport and Telecommunication Institute, Lomonosova 1, Riga, LV-1019, Latvia

FORECASTING OF RAIL FREIGHT CONVEYANCES IN EU COUNTRIES ON THE BASE OF THE SINGLE INDEX MODEL

Diana Santalova

Riga Technical University, Faculty of Transport and Mechanical Engineering
Kalku Str. 1, Riga, LV-1658, Latvia
E-mail: Diana.Santalova@rtu.lv

There are the regression models which describe rail freight conveyances of the member countries of the European Union considered in the investigation. The models contain such factors for each country as: total length of railways, gross domestic product per capita in Purchasing Power Standards and so on. All calculations are performed on the basis of the statistical data taken from EUROSTAT YEARBOOK 2005. Two estimation approaches are compared: the classical linear regression model and the single index model. Various tests for hypothesis of explanatory variables insignificance and model correctness have been carried out, and the cross-validation approach has been applied as well. The analysis has shown obvious advantage of the single index model.

Keywords: *freight conveyances, forecasting, single index model*

1. Introduction

In this paper we consider the problem of forecasting of rail freight conveyances from the member countries of the European Union on the basis of EUROSTAT YEARBOOK 2005 data [5]. For that the *linear regression model* [4] and the *single index model* (SIM) [3] are used. The *object* of consideration is rail freight conveyance expressed in million tonne-kilometres. We call *observation* the data about an object for a concrete year from 1996 till 2000. The following countries were considered: Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden and the United Kingdom. The main difficulty is to choose the set of convenient factors influencing the rail freight conveyances. The task of research is to construct various regression models, i.e. models with different combinations of explanatory factors, and then to choose from them the ones, that give the best forecasts of conveyances. We use the following well known criteria for comparing the elaborating models: the coefficient of multiple determination R^2, Fisher's and Student's criteria and the residual sum of squares R_0 [1, 4]. The described below cross-validation approach is used as well. Especially for the single index model the series of experiments is carried out with the aim to determine the optimal value of bandwidth h. In the present paper a lot of attention is paid to this problem.

The paper is organized in a following way. First of all the used regression models are considered from theoretical point of view, then the used data are described. After that we consider the suggested group models for the forecasting of conveyances. The results of the carried out estimation and the comparative analysis of these models are presented as well.

2. Structure of the Used Models

In this research all investigated models are group models [1]. The main object of consideration is named an *object.* It is a freight conveyance from some EU country. The data about an object for a definite period of time is called *observation*. We talk about the *individual model* if one object corresponds to another object for various observations, and about the *group model* if one corresponds to various objects. In other words we are able to forecast rail freight conveyances for all considered countries using one and the same model.

With respect to used mathematical model we consider *linear regression models* and *semiparametric regression models*.

In general the regression model can be described as

$$Y_i = m(x_i) + \varepsilon_i, \tag{1}$$

where Y_i is a dependent variable in the *i*-th observation, $m(\circ)$ is an unknown regression function, x_i is a d-dimensional vector of independent variables, ε_i is a random term.

It is supposed that the random term has zero expectation ($E(\varepsilon) = 0$) and the variance $Var(\varepsilon) = \sigma^2 \psi(x)$, where σ^2 is an unknown constant and $\psi(x)$ is a known weighted function. Furthermore we have a sequence of independent observations (Y_i, x_i), $x_i = (x_{i,1}, x_{i,2}, ..., x_{i,d})$, $i = 1, 2, ..., n$. On that base we need to estimate the unknown function $m(x)$.

In the simplest case *the linear regression model* is used:

$$m(x_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_d x_{i,d} = \beta^T x_i , \quad (2)$$

where $\beta^T = (\beta_0 \quad \beta_1 \quad \dots \quad \beta_d)$ is vector of unknown coefficients, $x_i = (1 \quad x_{i,1} \quad \dots \quad x_{i,d})^T$ is a vector of independent variables in *i*-th observation.

As it is known the forecasts obtained using the linear regression models are not very good. So, for rail conveyances forecasting we use the *single index regression model* [3] as well:

$$m(x_i) = g(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_d x_{i,d}) = g(\beta^T x_i), \quad (3)$$

where $g(\circ)$ is an *unknown link function* of one dimensional variable and $\tau_i = \beta^T x_i$ is called an *index*.

3. Informative Base

For experiments we will use the below-described statistical data. All necessary data have been received from "The Statistical Office of the European Communities" electronic database (EUROSTAT) [5]. First of all, the variable of interest is the rail freight conveyance, expressed in million tonne-kilometres. Let us denote it by t_0.

The following factors have been selected as explanatory variables:

t_1 – country area, in thousands of km^2;

t_2 – Gross Domestic Product per capita in Purchasing Power Standards;

t_3 – comparative price level;

t_4 – total length of railways, in thousands of km;

t_5 – number of locomotives, in thousands;

t_6 – number of goods wagons, in thousands.

Let us comment on some of the described factors.

Gross Domestic Product is a measure for the economic activity. It is defined as the value of all goods and services produced less the value of any goods or services used in their creation. The volume index of GDP per capita in Purchasing Power Standards (PPS) for each country is expressed in relation to the European Union (EU-25) average set to equal 100.

Comparative price level is the ratio between Purchasing Power Parities (PPPs) and market exchange rate for each country.

4. Considered Models

Now let us describe four investigated regression models. Two of them are linear regression models and other two ones are SIM.

The first model is a simple linear regression model (2). The dependent variable $Y^{(L1)} = t_0$ is conveyance of rail freight transport in millions tonne-kilometres. Note, that superscript by *Y* is introduced just for identification of models. Explanatory variables are $x_1 = t_2$, $x_2 = t_3$, $x_3 = \frac{t_2}{t_3}$, $x_4 = t_4$, $x_5 = t_5$, $x_6 = t_6$. The ratio $\frac{t_2}{t_3}$ enables us to see how these two factors in aggregate influence conveyances.

The second model is modification of the previous one. The dependent variable $Y^{(L2)} = \frac{t_0}{\sqrt{t_1}}$ is the ratio between the conveyance and the square root of the country area. Explanatory factors are $x_1 = t_2, x_2 = t_3, x_3 = \frac{t_2}{t_3}, x_4 = t_4, x_5 = t_5, x_6 = t_6$. In addition we introduce here the factor t_7, which is the index of the country area, by which we are able to consider gradation of the countries' areas. It is equal to 1 for relatively small countries (with areas less than 40 000 km^2 or equal to 40 000 km^2), and it is equal to 0 for countries with areas larger than 40 000 km^2. For example, this index is equal to 1 for Belgium, Luxembourg and Austria, because the areas of these countries are smaller than 40 000 km^2.

Finally we consider two variants of the Single Index Model (3). In the first variant the value of the dependent variable $Y^{(SIM1)} = \frac{t_0}{t_1}$ is the ratio between the conveyance and the country area for a concrete year. In the second variant the dependent variable $Y^{(SIM2)} = \frac{t_0}{\sqrt{t_1}}$ coincides with a dependent variable from the second linear Model L2.

The sets of explanatory variables for the models *SIM*1 and *SIM*2 coincide with the set for the first linear Model L1.

Thus, we have four regression models. Our task is to estimate the unknown coefficients β for the models, to compare the suggested models and to choose the best ones taking in account their significance. All calculations are performed using Statistica 6.0 and MathCad 12 packages.

5. Estimation of the Linear Models

Firstly, we analyse all the suggested models in case of data smoothing. It means we estimate the unknown coefficients β by all the observations. Thus, we are able to evaluate, how the considered models can only smooth the known conveyances and what variables have the greatest influence upon the conveyances.

Let us describe the obtained results.

The estimated Model L1 has the following form:

$$\hat{E}\left(Y^{(L1)}(x)\right) = -3\,713 + 118x_1 + 26x_2 - 11\,769x_3 + 879x_4 + 549x_5 + 158x_6 .$$

The estimates of the coefficients and calculated values of the Student's criterion for the Model L1 are presented in Table 1. Here $\widetilde{\beta}_i$ is an estimate of β_i, *t*(68) is the calculated value of Student's criteria for 68 degrees of freedom, *p*-level is the error of the second kind (or level of insignificance of variable). The theoretical value of Student's criterion for 68 degrees of freedom and level of significance (or error of the first kind) α = 5% is equal to 1.67. Taking into account the fact that the hypothesis of *insignificance* of explanatory variable is tested, we can see that calculated value of Student's criterion exceeds its theoretical value for two variables only, i.e. these two variables cannot be recognized as insignificant. Thus, the most significant explanatory variables are x_4 and x_6, so, the greatest influence on conveyances is rendered by the total length of railways and the number of wagons. The positive sign for these variables corresponds to the physical sense of the regressors. The coefficient R^2 for this model is equal to 0.985 and the calculated value of Fisher's criterion is 383.69. The theoretical value of Fisher's criterion for 6 and 68 degrees of freedom and level of significance α = 5% is equal to 2.23. Comparing the theoretical and calculated values of Fisher's criterion we can conclude that the estimated Model L1 cannot be recognized as insignificant. So, Model L1 is adequate.

TABLE 1. Estimates of coefficients of Model L1 and their insignificance levels

Coefficients	$\widetilde{\beta}_i$	***t*(68)**	***p*-level**
β_0	-3 713	0.149195	0.881842
β_1	118	0.480762	0.632229
β_2	26	0.109604	0.913046
β_3	-11 769	-0.462115	0.645474
β_4	879	6.866741	0.000000
β_5	549	0.799173	0.426973
β_6	158	8.375650	0.000000

The estimated Model L2 is as follows:

$$\hat{E}\left(Y^{(L2)}(x)\right) = -120.4 - 1.2x_1 + 1.4x_2 + 110.2x_3 + 0.2x_4 + 5.9x_5 + 0.3x_6 + 29.2x_7 .$$

The results of the analysis of Model L2 are presented in the Table 2. As we can see, almost all explanatory variables are recognized to be significant by Student's criterion. Only total length of railways does not influence the dependent variable. We obtain the positive signs for all significant variables with the exception of GDP; that means the positive correlation between these explanatory variables and the dependent variable. The coefficient R^2 for this model is equal to 0.985 and the calculated value of Fisher's criterion is 313.78. The theoretical value of Fisher's criterion for 7 and 67 degrees of freedom and level of significance α = 5% is 2.15, so, this regression model is significant as well.

TABLE 2. Estimates of coefficients of Model L2 and their insignificance levels

Coefficients	$\tilde{\beta}_i$	***t*(68)**	***p*-level**
β_0	-120.4	-3.00514	0.003732
β_1	-1.2	-3.11117	0.002738
β_2	1.4	3.55818	0.000692
β_3	110.2	2.68390	0.009160
β_4	0.2	1.03172	0.305913
β_5	5.9	5.42836	0.000001
β_6	0.3	9.33665	0.000000
β_7	29.2	12.79621	0.000000

Figures 1 and 2 demonstrate how the investigated models smooth the observed true data. The observations are arranged in "country-year" order: every five points correspond to conveyances of some country during the analysed period from 1996 till 2000, i.e. for five years. Moreover, countries are sorted in alphabetical order. Horizontal axis reflects the number of observations, arranged in the above-mentioned order. Vertical axis reflects the corresponding conveyances, expressed in thousands. It is obvious that both linear models show the similar smoothing.

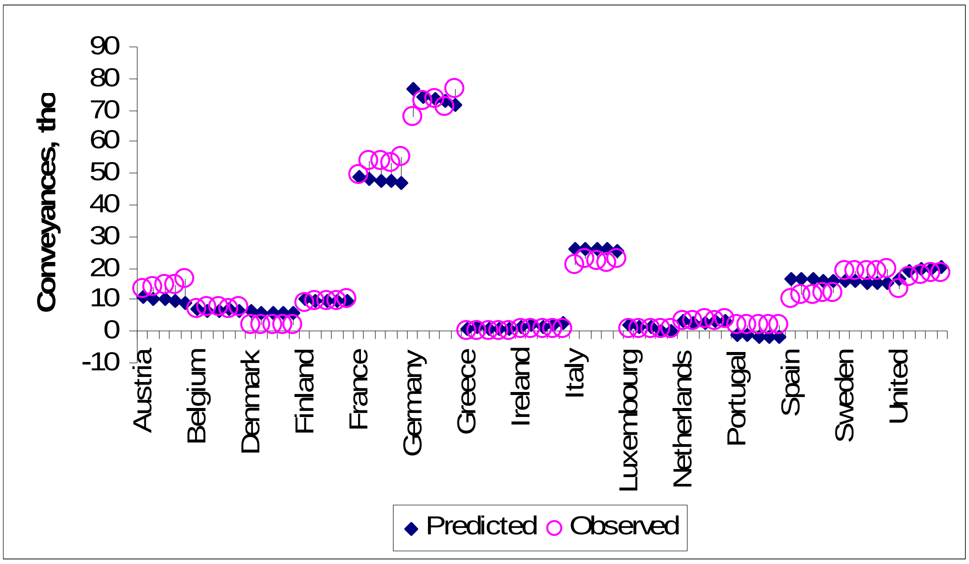


Figure 1. Smoothing by Model L1

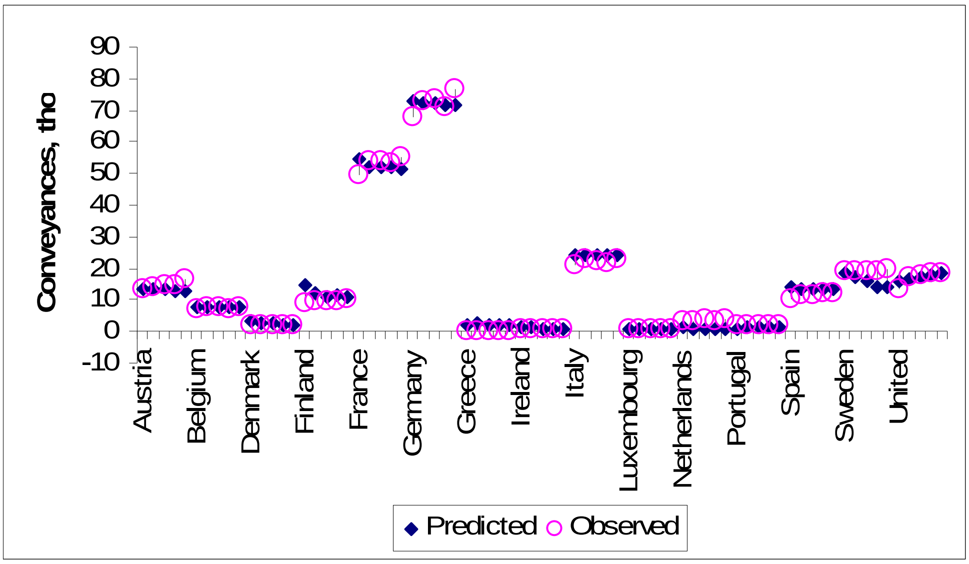


Figure 2. Smoothing by Model L2

6. Estimation of the Single Index Models

Now we will consider the suggested single index Models *SIM*1 and *SIM*2.

The estimation of these models consists of two steps: we have to estimate the unknown coefficients vector β and the link function g. For the latter the Nadaraya-Watson kernel estimator can be applied [3]:

$$\tilde{g}(x) = \frac{1}{\sum_{i=1}^{n} K_h(\tau_i)} \sum_{i=1}^{n} K_h(\tau_i) Y_i \,, \tag{4}$$

where $\tau_i = (x - x_i)^T \beta$ is a value of index for the i-th observation, Y_i is a value of the dependent variable for i-th observation and $K_h(\circ)$ is the so-called *kernel function*.

We use the Gaussian function as $K_h(\circ)$:

$$K_h(\tau) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\tau}{h}\right)^2 \right), \qquad -\infty < \tau < \infty, \tag{5}$$

where h is a *bandwidth*.

The unknown parameter vector β is estimated using the least squares criterion:

$$R(\beta) = \sum_{i=1}^{n} (Y_i - \tilde{g}(x_i))^2 \to \min_{\beta} \,. \tag{6}$$

For that we use the gradient method. The corresponding gradient is the following:

$$\nabla R(\beta) = -2\sum_{i=1}^{n} \left(Y_i - \frac{\sum_{i=1}^{n} K_h(\tau_i) Y_i}{\sum_{i=1}^{n} K_h(\tau_i)} \right) \cdot \left(\sum_{i=1}^{n} K_h(\tau_i) \right)^{-2} \cdot \left(\frac{1}{h} \sum_{i=1}^{n} Y_i \frac{\partial}{\partial \tau_i} K_h(\tau_i) \cdot (Y_i - \tilde{Y}_i) \cdot x_i \right), \tag{7}$$

where

$$\widetilde{Y}_i = \sum_{j=1}^{n} K_h(\tau_j) Y_j \tag{8}$$

and

$$\frac{\partial}{\partial \tau_i} K_h(\tau_i) = -\frac{\tau_i}{h^2\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\tau_i}{h}\right)^2\right) \tag{9}$$

is the derivative of the Gaussian kernel.

We are able to compare single index models by the residual sum of squares R_0 only. We calculate the residual sum of squares as follows:

$$R_0 = \frac{1}{n-d}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2, \tag{10}$$

where n is a number of observations, d is a number of estimated coefficients, Y_i is an observed value and $\hat{Y}_i = \widetilde{g}(x_i)$ is an estimated value.

The estimates of coefficients β , i.e. the values of coefficients β optimizing the object function (6), for both single index models have been obtained from the same starting point $\beta^{(0)}$ and with bandwidth $h = 7$ for $SIM1$ and $h = 6$ for $SIM2$. Note that these values of bandwidth are optimal and have been obtained as a result of the series of experiments using our own program written in MathCad12 package.

The estimated Model $SIM1$ has the following form:

$$\hat{E}\left(Y^{(SIM1)}(x)\right) = \frac{\sum_{i=1}^{n} Y_i K_h\left(710(x-x_{1,i}) - 1\times10^3(x-x_{2,i}) + 18\times10^{-5}(x-x_{3,i}) + 758(x-x_{4,i}) + 155(x-x_{5,i}) - 2\times10^3(x-x_{6,i})\right)}{\sum_{i=1}^{n} K_h\left(710(x-x_{1,i}) - 1\times10^3(x-x_{2,i}) + 18\times10^{-5}(x-x_{3,i}) + 758(x-x_{4,i}) + 155(x-x_{5,i}) - 2\times10^3(x-x_{6,i})\right)}.$$

The estimated Model $SIM2$ can be written in the following way:

$$\hat{E}\left(Y^{(SIM2)}(x)\right) = \frac{\sum_{i=1}^{n} Y_i K_h\left(716(x-x_{1,i}) - 1\times10^3(x-x_{2,i}) - 4(x-x_{3,i}) + 853(x-x_{4,i}) + 62(x-x_{5,i}) - 871(x-x_{6,i})\right)}{\sum_{i=1}^{n} K_h\left(716(x-x_{1,i}) - 1\times10^3(x-x_{2,i}) - 4(x-x_{3,i}) + 853(x-x_{4,i}) + 62(x-x_{5,i}) - 871(x-x_{6,i})\right)}.$$

Figures 3 and 4 represent smoothing by these models. Obviously, the estimates of conveyances almost in all observations coincide with the true conveyances.

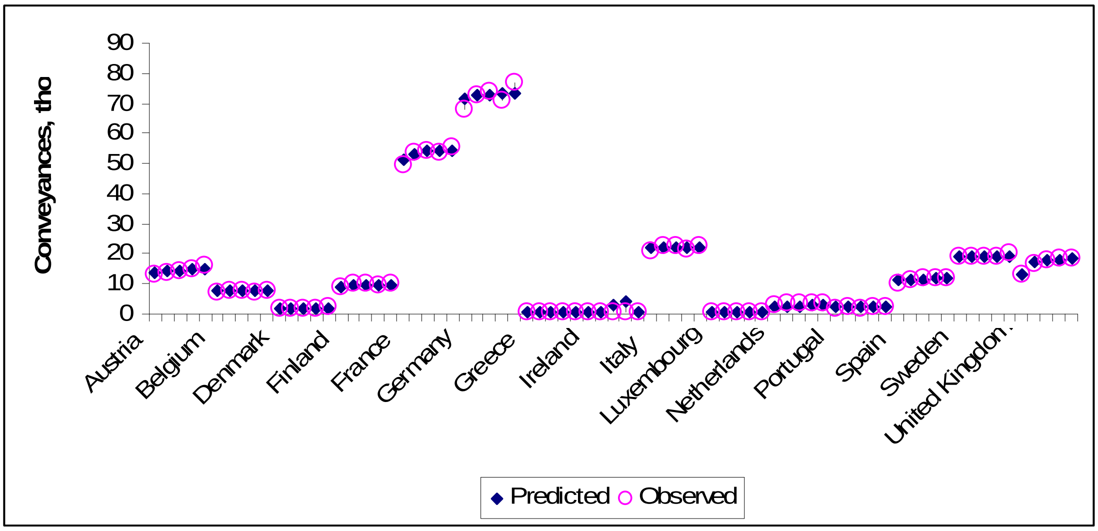


Figure 3. Smoothing by Model SIM1

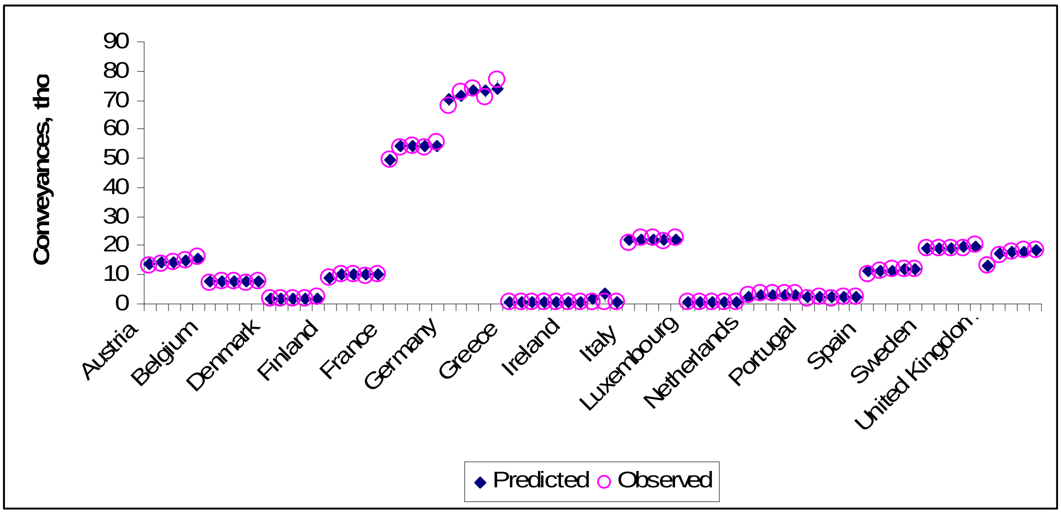


Figure 4. Smoothing by Model SIM2

As both linear models and both single index models give approximately similar results in data smoothing, we have to consider how precise the forecasts are which are given by analysing models. For this purpose we use the residual sum of squares R_0 (10). Table 3 involves the values of the residual sums of squares for all the models.

TABLE 3. Values of R_0 in case of smoothing

Model	***L*1**	***L*2**	***SIM*1**	***SIM*2**
R_0	11 543 065	4 830 576	894 265	565 407

So we can conclude that the linear Model L2 and the single index Model *SIM*2 have the minimum value of R_0 that means greater significance of these models in comparison with two others. As it was supposed, in general SIM gives the most precise estimates.

7. Cross-Validation Analysis

Now we will consider the suggested models from the other point of view. We use the *cross-validation approach*. That means we estimate the unknown coefficients β for the models on the basis of a part of the data. Then using the obtained estimates of β we forecast the conveyances for a remained part of the data and compare these forecasted conveyances with the real ones, i.e. we calculate R_0 for each model. Also the optimum value of bandwidth *h* is found for both single index models.

We estimate the coefficients β on the basis of the period from 1996 till 1999 and perform the forecast for the year 2000. Table 4 contains the estimates of β for the considered linear regression models. The signs of estimates correspond to physical sense of explanatory factors.

TABLE 4. Estimates of coefficients for the linear models

Coefficients	***L*1**	***L*2**
β_0	4 154.4	-119.7
β_1	197.7	28.7
β_2	45.4	-1.2
β_3	-20 555.8	1.4
β_4	898.5	110.0
β_5	531.6	0.2
β_6	148.1	6.0
β_7	–	0.3

The residual sum of squares R_0 for Model L1 is 18 509 464 and for Model L2 is 8 941 875. Obviously, forecasts of rail freight conveyances obtained by the second linear model have to be much better than those obtained by the first one. Moreover, the first linear model gives negative forecasts of some small conveyances. The true observed values of conveyances and the corresponding forecasts are displayed in Figures 5 and 6. We can see that Model L2 is more sensitive to the small conveyances which belong to the countries with small areas. Obviously this effect is achieved by using the above-mentioned additional gradation factor.

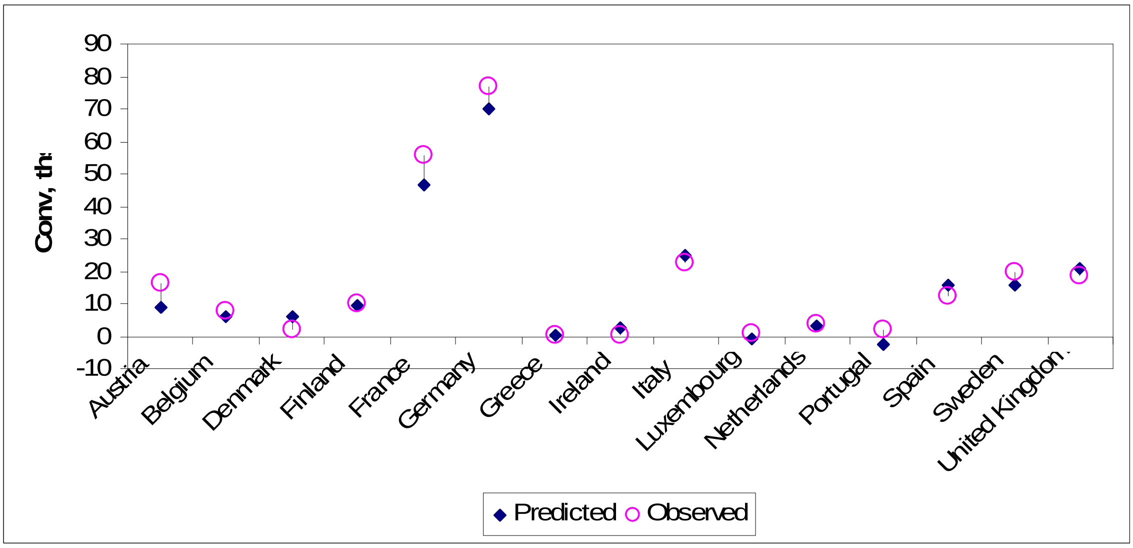


Figure 5. Forecasting by Model L1

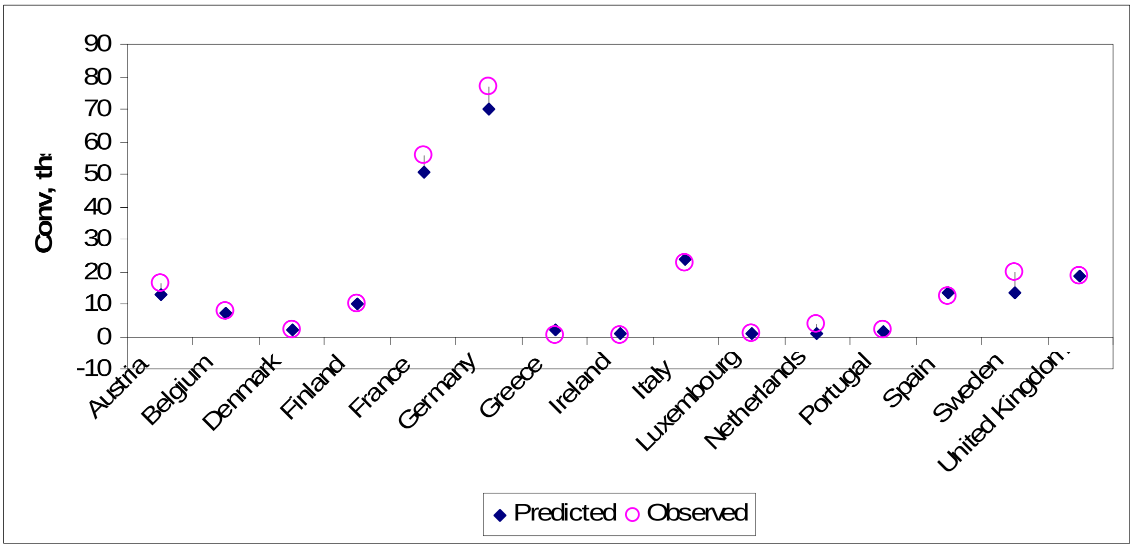


Figure 6. Forecasting by Model L2

Now we will analyze SIM in detail. We begin with a choice of the bandwidth size. Our task is to find the optimal value of bandwidth h_0 that gives a minimal value of R_0 (see [3]). The series of experiments was performed and the different estimates of β and values of R_0 depending on various h were obtained as well. The corresponding results for Models *SIM*1 and *SIM*2 are shown in Tables 5 and 6 respectively. We can see that all β estimates differ from each other depending on h in spite of the fact that they were obtained from the same initial value β_0. The values of R_0 (expressed in millions) corresponding to various h for both SIM are represented in Table 7. Thus, the best result for R_0 is achieved for h_0 = 7 and h_0 = 8 for *SIM*1 and for h_0 = 6 for *SIM*2. As it was supposed the sum of squared residuals increases if h is bigger and smaller than the optimal value. The forecasted conveyances by *SIM*1 with h_0 = 7 and by SIM_2 with h_0 = 6 and observed conveyances are shown on the Figures 7 and 8, respectively.

TABLE 5. The estimates of β for *SIM*1

Coefficients	Bandwidth ***h***								
	1	5	6	7	8	9	10	15	20
β_1	22.8	566.4	248.6	33 160.0	344.5	721.2	3 530.0	1 327.0	-901.7
β_2	26.6	299.9	216.7	-22 420.0	-95.8	-24.1	-512.5	358.2	1 304.0
β_3	-0.04	1.9	-0.03	565.5	4.4	7.2	38.4	8.6	-19.7
β_4	0.13	257.9	88.9	19 870.0	120.5	207.1	1 310.0	550.0	1 011.0
β_5	$4x10^{-5}$	62.9	17.9	3 996.0	25.3	37.7	174.4	119.2	198.3
β_6	$1x10^{-5}$	885.7	252.2	56 310.0	356.9	572.7	3 832.0	3 058.0	724.6

TABLE 6. The estimates of β for *SIM*2

Coefficients	Bandwidth ***h***								
	1	5	6	7	8	9	10	15	20
β_1	29.3	859.9	962.7	$1.7x10^3$	$1.2x10^3$	697.5	618.7	757.1	$-4.5x10^3$
β_2	18.3	462.3	$1.2x10^3$	$1.0x10^3$	654.8	578.5	276.8	791.3	$3.4x10^3$
β_3	0.1	3.4	-2.4	7.5	6.0	1.9	3.3	-0.2	-73.6
β_4	-0.2	440.7	604.3	$1.2x10^3$	636.6	664.1	525.1	737.1	4.0E+3
β_5	$4x10^{-5}$	103.7	49.0	97.1	61.0	24.1	12.1	17.3	916.8
β_6	$1x10^{-5}$	$1.5x10^3$	690.0	$1.4x10^3$	859.7	851.7	672.9	$1.2x10^3$	$1.9x10^4$

TABLE 7. The values of R_0 for *SIMs*

	Bandwidth ***h***								
	1	5	6	7	8	9	10	15	20
*SIM*1	676.9	24.3	2.0	1.9	1.9	2.5	24.3	24.3	44.5
*SIM*2	676.9	24.2	1.9	1.9	1.9	2.4	3.0	7.5	7.3

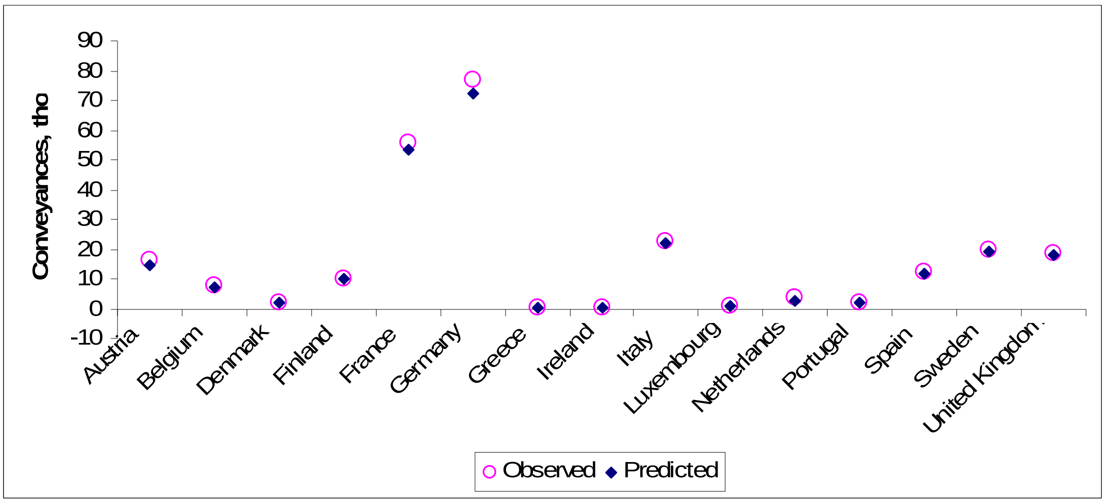


Figure 7. Forecasting by SIM1

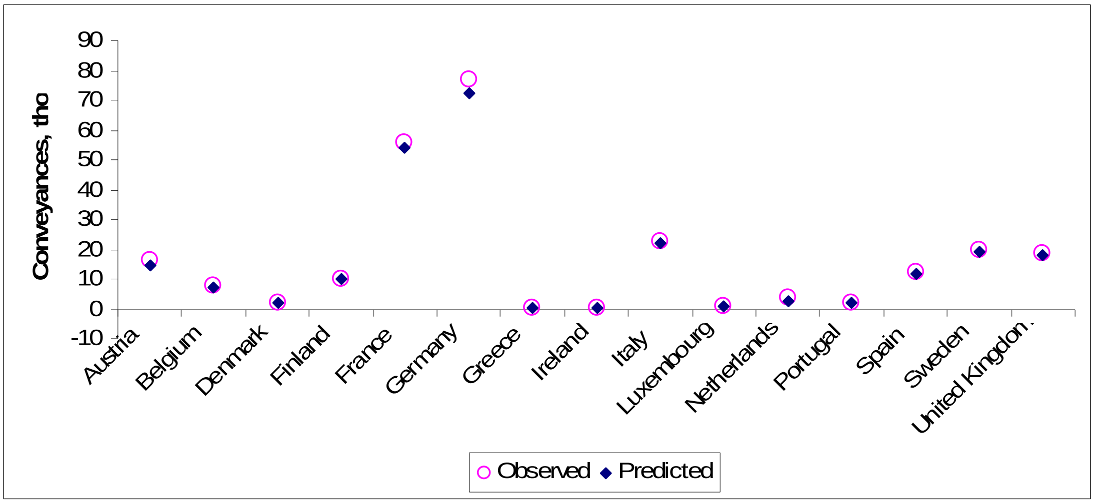


Figure 8. Forecasting by SIM2

Obviously, the forecasted values are very close to the observed values almost in all the observations.

Table 8 contains the values of R_0 for four investigated models. As we can see, values of R_0 for single index models are in a number of orders less than for linear models. This fact gives evidence of greater accuracy of SIM models.

TABLE 8. The values of R_0 in case of forecasting

Model	***L*1**	***L*2**	***SIM*1**	***SIM*2**
R_0	18 509 464	8 941 875	1 894 237	1 896 287

From Figure 9 we can also visually evaluate behaviour of R_0 with respect to bandwidth h for both single index models.

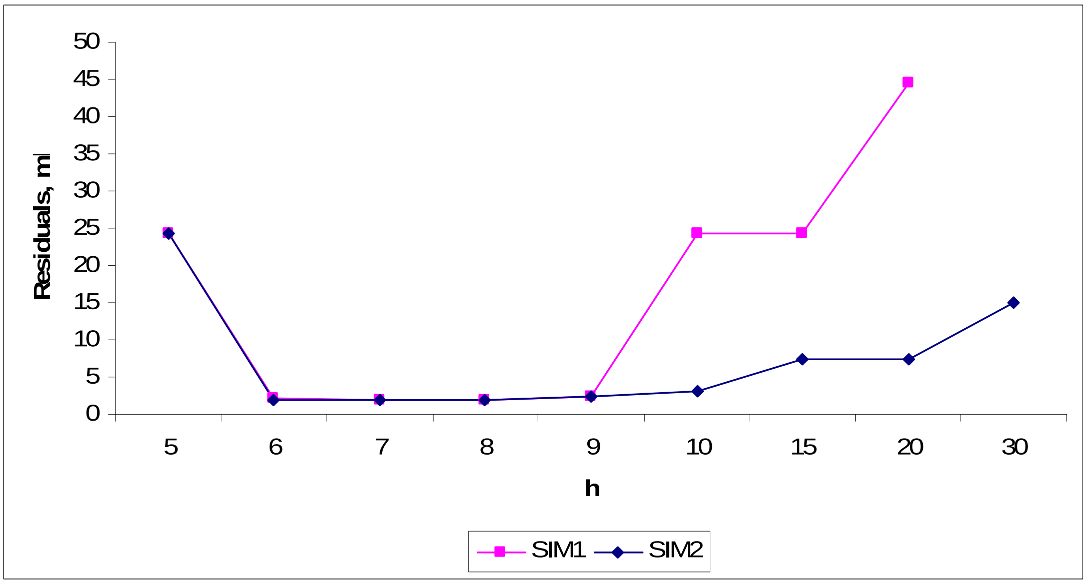


Figure 9. R_0 depending on bandwidth for SIMs

Conclusions

In the presented paper two kinds of models for forecasting of inland rail freight conveyances are considered: linear regression model and single index regression model. Four different regression models were constructed and tested, two of them are linear regression models and two others are single index models. For the estimation of unknown coefficients in case of SIM the Nadaraya-Watson estimator and Gaussian kernel function were used. The efficiency of these models was investigated through the consideration of conveyances for the 15 member countries of the European Union. All the considered models include a great number of explanatory factors. The performed investigations show that the single index regression model gives more precise forecasts than classical methods of linear regression. For this purpose all the models have been estimated and compared by the criterion of the residual sum of squares in case of data smoothing and in case of forecasting as well, that required the cross-validation approach. Moreover, the optimal values of smoothing parameter h for the considered single index models have been obtained experimentally.

References

1. Andronov, A. M. et al. *Forecasting of the passenger conveyances on the air transport.* Moscow: Transport, 1983. (In Russian)
2. Andronov, A., Zhukovskaya, C., Santalova, D. On Mathematical Models for Analysis and Forecasting of the Europe Union Countries Conveyances. In: *Scientific Papers of the Riga Technical University, Information Technology and Management Science*. Riga: RTU 2006. (In printing)
3. Hardle, W., Muller, M., Sperlich, S., Werwatz, A. *Nonparametric and Semiparametric Models*. Berlin: Springer-Verlag, 2004.
4. Srivastava, M. S. *Methods of Multivariate Statistics.* New York: John Wiley & Sons Inc, 2002.
5. *EUROSTAT YEARBOOK 2005. The statistical guide to Europe. Data 1993–2004. EU, EuroSTAT, 2005* – http://epp.eurostat.ec.europa.eu

Computer Modelling & New Technologies, 2007, Volume 11, No. 1
Transport and Telecommunication Institute, Lomonosova 1, Riga, LV-1019, Latvia

Authors' index

Igor V. Kabashkin (born in Riga, August 6, 1954)

- Vice-rector for Research and Development Affairs of Transport and Telecommunication Institute, Professor, Director of Telematics and Logistics Institute,
- PhD in Aviation (1981, Moscow Institute of Civil Aviation Engineering), Dr. habil.sc. in Aviation (1992, Riga Aviation University), the Member of the International Telecommunication Academy, the Member of IEEE, the Corresponding Member of Latvian Academy of Sciences (1998)
- **Publications:** 330 scientific papers and 67 patents.
- **Research activities:** information technology applications, operations research, electronics and telecommunication, analysis and modelling of complex systems, transport telematics and logistics

Alexander M. Andronov (born in Moscow, September 20, 1937)

- Professor of Riga Technical University, a Head of Department "Mathematical Support of Transport System's Control"
- **University study**: Moscow Aviation Technological Institute (1955-1960); PhD (Technical Cybernetic), Central Institute of Economics & Mathematics of USSR Academy of Sciences, Moscow, USSR (1967); Dr.habil.sc. (Operation of Aviation Transport), Civil Aviation Institute, Kiev, USSR (1973)
- **Publications:** 10 books, more than 200 scientific papers
- **Scientific activities**: Probability Theory and Mathematical Statistics, Stochastic Processes and Its Applications, Discrete Mathematics
- **Honour and Memberships:** the Honorary Scientist of Latvia (1990), the Winner of USSR Cabinet Prize (1982), the Outstanding Aeroflot Worker (1991), the Member of Editorial Board of the Journal "Automatic Control and Computer Sciences" (Riga-New York), the Member of American Statistical Association, the Member of the Presidium of Latvian Simulation Society, the Member of Latvian Society of Higher School Professors

Eugene A. Kopytov (born in Lignica, Poland, December 5, 1947)

- Rector of Transport and Telecommunication Institute, Professor, Dr.habil.sc.ing,, the Member of International Telecommunication Academy
- *Director of Programme* "Master of Natural Sciences in Computer Science" (Transport and Telecommunication Institute)
- **University study**: Riga Civil Aviation Engineering Institute (1966-1971)
- Candidate of Technical Science Degree (1984), Kiev Civil Aviation Engineering Institute); Dr.sc.ing. (1992) and Dr.habil.sc.ing. (1997), Riga Aviation University
- **Publications:** 220 publications, 1 certificate of inventions
- **Scientific activities**: mathematical and computer modelling, system analysis, statistical recognition and classification, modern database technologies

Leonid Y. Greenglaz (born in Bobruisk, December 29, 1940)

- Professor of the Department of Mathematics and Informatics in Riga International School of Economics and Business Administration, Dr.math., the Member of Latvian Mathematical Society
- **University study**: Ural State University, Sverdlovsk (1957-1963)
- Dr.math. (1969), the University of Latvia
- **Publications:** more than 60 scientific papers and teaching books, including 2 monographs
- **Scientific activities**: algebraic structures in automata; application of mathematical methods in business, especially application of statistical methods in business; computer modelling

Aivar Muravyov (born in Riga, May 3, 1984)

- Postgraduate student of Programme “Master of Natural Sciences in Computer Science” (Transport and Telecommunication Institute)
- *System administrator of IT Department* (Transport and Telecommunication Institute)
- **Bachelor study**: Bachelor of Natural Sciences in Computer Science, Transport and Telecommunication Institute (2002-2006).
- **Publications:** 3 publications
- **Scientific activities**: computer modelling

Edvin Puzinkevich (born in Daugavpils, August 8, 1977)

- Lecturer (Transport and Telecommunication Institute)
- Computer Designer (Daugavpils University)
- **University study**: Daugavpils University (1996-2002)
- Master degree (Computer Science)
- **Publications:** 5 publications
- **Scientific activities**: logistics, inventory control, mathematical and computer modelling

Andrey Kashurin (born in Riga, July 31, 1980)

- *Network Administrator*, Department of Mathematical Support of Transport System Management, Riga Technical University
- *IT specialist*, Ministry of Education and Science of the Republic of Latvia, the Sports Administration
- **University study**: Riga Technical University, Faculty of Computer Science and Information Technology (1998-2004)
- Mg.sc.ing. (Information technologies) 2004, Riga Technical University, Faculty of Computer Science and Information Technology

Andrey Svirchenkov (born in Uzbekistan, December 28, 1962)

Chairman of the Board of JSC „NORVIK BANKA”

Michael S. Tikhov (born in Makovo, Astrakhan region, Russia, January 26, 1947)

- Professor of the Chair of the Theory of Statistical Decisions, Dr. phys.&math., the State University of Nizhegorodsk
- The Member of the Dissertation Council (Д 212.166.13)
- **University study**: Faculty of Mechanics and Mathematics at the State University of Gorky
- The Candidate of Phys.&Math., 1983, the Institute of Mathematics named by V.I. Steklov, Academy of Sciences of the USSR, Leningrad; Dr.phys.&math., 1994, the University of St. Petersburg
- **Publications:** about 180 publications, 1 patent, 3 monographs
- **Scientific activities**: mathematical statistics, statistical evaluation on censored samples

Dmitriy S. Krishtopenko (born in Gorky, Russia, November 16, 1984)

- **University study**: Bachelor, 2nd year Master studies, the Chair of Applied Theory of Probabilities, the State University of Nizhegorodsk named by N.I. Lobachevskiy
- **Publications**: about 20 publications
- **Scientific activities:** mathematical statistics, checking of statistical hypotheses, nonparametric evaluation of distributions

Marina V. Yaroschuk (born in Gorky, Russia, June 26, 1977)

- The Assistant at the Chair of the Theory of Statistical Decisions**,** the State University of Nizhegorodsk named by N.I. Lobachevskiy
- **University study:** the State University of Nizhegorodsk, Faculty of Mechanics and Mathematics (1994-1999), the post-graduate student of the speciality 01.01.05 – probability theory and mathematical statistics at the Chair of the Theory of Statistical Decisions
- **Publications: 21 publications**
- **Scientific activities:** mathematical statistics, nonparametric evaluation of distributions

Catherine V. Zhukovskaya (born in Riga, March 24, 1970)

- Docent of the BIA (Baltic International Academy), Mg.sc.ing.
- **University study**: RAU (Riga Institute of Civil Aviation Engineers up to 1992) (1978-1992) – radio-engineer
- Mg.sc.ing. RAU (1994-1996)
- **Doctoral study**: from 2004 up till now at Riga Technical University
- **Publications:** 7 publications, 4 books
- **Scientific activities**: multivariate statistics, discrete choice models, optimisation, nonparametric statistics, applied statistics

CUMULATIVE INDEX

COMPUTER MODELLING and NEW TECHNOLOGIES, volume 11, No. 1, 2007 (Abstracts)

Yury Paramonov, Janis Andersons. Analysis of Fiber Strength Dependence on Length Using an Extended Weakest Link Distribution Family, *Computer Modelling and New Technologies*, vol. 11, No 1, 2007, pp. 8-20.

An extended family of the weakest-link models based on the assumption of a two-stage failure process of a fiber specimen was developed in [1, 2]. A generalization of this family is presented in this paper. As in [1, 2] we consider the specimen as a chain of n elements (links). The fracture process is modelled as follows: in the first stage initiation of defects (before loading or during loading), and in the second stage a specimen fracture takes place. As distinct from our previous publications, the strength of items without defects is taken into account and two types of the influence of defect number on the specimen strength are considered. The comparison of the models and the choice of the best one are made using cross validation method. The offered models sometimes describe more adequately the experimentally observed fibre strength scatter and the strength dependence on fibre length than the traditional models do.

Keywords: distribution function, composite, static strength

Eugene Kopytov, Leonid Greenglaz, Aivar Muravyov, Edvin Puzinkevich. Modelling of Two Strategies in Inventory Control System with Random Lead Time and Demand, *Computer Modelling and New Technologies*, vol. 11, No 1, 2007, pp. 21-30.

The paper considers two multiple period single-product inventory control models with random parameters. These models are of interest because they illustrate real situations of the business. The first model is a model with fixed reorder point and fixed order quantity. The second model is the model with fixed period of time between the moments of placing neighbouring orders. Order quantity is determined as difference between the fixed stock level and quantity of goods in the moment of ordering. The considered models are realized using analytical and simulation approaches. The numerical examples of problem solving are presented.

Keywords: inventory control, demand, lead time, order quantity, reorder point, analytical model, simulation

Alexander Andronov, Andrey Kashurin. On a Problem of Spatial Arrangement of Service Stations, *Computer Modelling and New Technologies*, vol. 11, No 1, 2007, pp. 31-37.

A problem of service station arrangement on spatial space is considered. A density function of serviced object location and a function that describes the corresponding loss are known. As criteria of the arrangement is an average total loss. For optimisation the gradient method is used. Numerical examples illustrate the suggested approach to setting problem solution.

Keywords: spatial arrangement, service stations, gradient method

Andrey Svirchenkov. Practical Method of Ruin Probability Calculation for Finite Time Interval, *Computer Modelling and New Technologies*, vol. 11, No 1, 2007, pp. 38-43.

A modification of the classical ruin problem is considered. Novelty consists in a consideration of nonhomogeneous Poisson flow of claims, arbitrary distribution of claim costs and existence of lower level of necessary capital for any time moment t. The problem is to calculate a probability that this lower level is not to be passed.. A numerical method has been elaborated for the probability evaluation. The considered method is based on Markov chain theory and Edgeworth expansion for the probabilistic density function.

Keywords: ruin problem, Edgeworth expansion, Markov chain, numerical method

Mikhail S. Tikhov, Dmitriy S. Krishtopenko and Marina V. Yarochuk. Asymptotic Normality of the Integrated Square Error at the Fixed Plan of Experiment for Indirect Observations, *Computer Modelling and New Technologies*, vol. 11, No 1, 2007, pp. 46-56.

The goal of this paper is to establish the asymptotic normality of the L_2-deviation of the kernel distribution function estimator $F_n(x)$ defined by $I_n = \int (F_n(x) - F(x))^2 \omega(x) dx$, where $F(x)$ is the unknown distribution function of a random variable X,. $\omega(x)$ is the weight function in dose-response dependence on the sample $U^{(n)} = \{(W_i, Y_i), 1 \le i \le n\}$, $W_i = I(X_i < u_i)$ is the indicator of even 5. $(X_i < u_i)$ and Y is a random variable, u_i is fixed values. This result is useful for constructing the test goodness-of-fit for the distribution function $F(x)$.

Keywords: dose-response dependence, indirect observation, integrated square error

Alexander Andronov. On Nonparametric Interval Estimation of a Regression Function Based on the Resampling, *Computer Modelling and New Technologies*, vol. 11, No 1, 2007, pp. 57-61.

A nonparametric regression model $E(Y) = m(x)$ is considered where Y is a dependent variable, x is a d-dimensional vector of independent variables (regressors) and m is an unknown function. A sequence of independent observations $(Y_i, x_i,)$ $i = 1, 2, \ldots, n$, is available. Our aim is to construct an upper confidence bound for $m(x)$ that corresponds to probability γ. The resampling approach is used. The suggested methods allow calculating true cover probability.

Keywords: nonparametric regression, interval estimation, resampling

Catherine Zhukovskaya. Use of the Generalized Linear Model in Forecasting the Air Passengers' Conveyances from EU Countries, *Computer Modelling and New Technologies*, vol. 11, No 1, 2007, pp. 62-72.

Some regression models to forecast the air passengers' conveyances from EU countries are considered. Two different approaches for the above-mentioned task of forecasting are shown. The first one is the classical method of *linear regression* and the second one is its *generalized* approach. The considered regression models contain many explanatory factors and their combinations. The advantage of using the *generalized linear model* (GLM) in comparison with the classical *linear regression model* is shown.

Keywords: air passengers' conveyances, forecasting, generalized linear model

Diana Santalova. Forecasting of Rail Freight Conveyances in EU Countries on the Base of the Single Index Model, *Computer Modelling and New Technologies*, vol. 11, No 1, 2007, pp. 73-83.

There are the regression models which describe rail freight conveyances of the member countries of the European Union considered in the investigation. The models contain such factors for each country as: total length of railways, gross domestic product per capita in Purchasing Power Standards and so on. All calculations were performed on the basis of the statistical data taken from EUROSTAT YEARBOOK 2005. Two estimation approaches were compared: the classical linear regression model and the single index model. Various tests for hypothesis of explanatory variables insignificance and model correctness have been carried out, and the cross-validation approach has been applied as well. The analysis has shown obvious advantage of the single index model.

Keywords: freight conveyances, forecasting, single index model

COMPUTER MODELLING and NEW TECHNOLOGIES, 11.sējums, Nr.1, 2007 (Anotācijas)

Jurijs Paramonovs, Janis Andersons. Šķiedras stiprības atkarības no šķiedras garuma modelēšana, izmantojot paplašināto vājākā ķēdes posma sadalījumu kopu, *Computer Modelling and New Technologies,* 11.sēj., Nr.1, 2007, 8.-20. lpp.

Iepriekšējos rakstos ir aprakstīta paplašināta vājākā ķēdes posma sadalījumu kopa, kas balstās uz pieņēmumu par divu stadiju sabrukšanas procesu šķiedrām. Šajā rakstā izklāstīts minētās kopas vispārinājums. Autori paraugu aplūko kā *n* elementu (posmu) ķēdi. Sabrukšana tiek modelēta sekojoši: pirmajā stadijā notiek defektu veidošanās (pirms slogošanas vai tās laikā) un otrajā stadijā – parauga sabrukšana. Atšķirībā no mūsu iepriekšējām publikācijām, tiek ņemta vērā arī no defektiem brīvo elementu stiprība un tiek aplūkoti divi mehānismi defektu skaita ietekmei uz parauga stiprību. Modeļu salīdzinājums un labākā modeļa izvēle balstās uz „krosvalidācijas" metodi. Piedāvātie modeļi dažkārt apraksta eksperimentāli novēroto šķiedru stiprības izkliedi un stiprības atkarību no šķiedras garuma, precīzāk nekā tradicionālie modeļi.

Atslēgvārdi: sadalījuma funkcija, kompozīts, statiskā stiprība

Jevgeņijs Kopitovs, Leonīds Gringlazs, Aivars Muravjovs, Edvīns Puzinkevičs. Divu stratēģiju modelēšana uzskaites kontroles sistēma ar nejaušu piegāzu laiku un pieprasījumu**,** *Computer Modelling and New Technologies,* 11.sēj., Nr.1, 2007, 21.-30. lpp.

Rakstā tiek apskatīti divi daudzkārtīga perioda viena produkta uzskaites kontroles modeļi ar nejaušiem parametriem. Šie modeļi izraisa interesi tādēļ, ka tie parāda biznesa reālas situācijas. Pirmais modelis ir modelis ar fiksētu atkārtotu pasūtījumu punktu un fiksētu pasūtījumu kvantitāti. Otrais modelis ir modelis ar fiksētiem laika periodiem starp momentiem, kad tiek izvietoti blakus pasūtījumi. Pasūtījumu kvantitāte tiek noteikta kā diference starp fiksēto krājuma līmeni un preču kvantitāti pasūtījuma momentā. Apskatītie modeļi tiek īstenoti, lietojot analītisko un modelēšanas pieeju. Problēmas risināšanā tiek piedāvāti arī skaitliskie piemēri.

Atslēgvārdi: uzskaites kontrole, pieprasījums, pasūtījuma kvantitāte, atkārtotu pasūtījumu punkts, analītiskais modelis, modelēšana

Aleksandrs Andronovs, Andrejs Kašurins. Par apkalpošanas staciju novietošanas problēmu, *Computer Modelling and New Technologies,* 11.sēj., Nr.1, 2007, 31.-37. lpp.

Ir apskatīta problēma par apkalpošanas staciju novietošanu. Apkalpojamo objektu dislokācijas blīvuma funkcija un izmaksas funkcija ir zināmas. Par novietošanas kritēriju ir pieņemtas vidējas kopīgas izmaksas. Optimizācijas gaitā ir izmantota gradienta metode. Skaitliskais piemērs ilustrē piedāvājumu pieeju apskatītās problēmas atrisinājumam.

Atslēgvārdi: apkalpošanas staciju novietošana, blīvuma funkcija, gradienta metode

Andrejs Svirčenkovs. Praktiskā metode izputēšanas varbūtības aprēķinam galīgā laika intervāla gadījumā, *Computer Modelling and New Technologies,* 11.sēj., Nr.1, 2007, 38.-43. lpp.

Rakstā ir apskatīta klasiskās izputēšanas problēmas modifikācija. Novitātes ir šādas: apskatīta nestacionāra Puasona plūsma, patvaļīgs sadalījums pieprasījuma maksai un zemākā līmeņa eksistence nepieciešamam kapitālam katram laika momentam. Uzdevums ir aprēķināt varbūtību, lai tās līmeni nepārsniegtu. Skaitliskā metode bija izstrādāta šādas varbūtības aprēķinam. Metode ir balstīta uz Markova ķēžu teoriju un *Edgeworth* izvirzījumu varbūtiskai blīvuma funkcijai.

Atslēgvārdi: izputēšanas problēma, *Edgeworth* izvirzījums, Markova ķēde, skaitliskā metode

Mihails S. Tihovs, Dmitrijs S. Krištopenko, Marina V. Jaročuka. Integrētās kvadrāta kļūdas asimptotiskā normalitāte eksperimenta fiksētā projektā netiešajiem novērojumiem, *Computer Modelling and New Technologies,* 11.sēj., Nr.1, 2007, 46.-56. lpp.

Šī darba mērķis ir noteikt sadales funkcijas kodola novērtētāja $F_n(x)$ L_2 -novirzes asimptotisko normalitāti, kas ir noteikta ar $I_n = \int (F_n(x) - F(x))^2 \omega(x)\,dx$, kur $F_n(x)$ nezināma sadales funkcija nejaušā mainīgā *X*, $\omega(x)$ ir svara funkcija dozas-atbildes atkarībā no modeļa $U^{(n)} = \{(W_i, Y_i), 1 \le i \le n\}$, $W_i = I(X_i < u_i)$ ir vienmērīgā $(X_i < u_i)$ rādītājs un *Y* ir nejaušais mainīgais, u_i ir fiksētās vērtības. Šis rezultāts ir noderīgs, lai veidotu izpētes-labuma-derīguma sadales funkciju *F*(*x*).

Atslēgvārdi: dozas-atbildes atkarība, netiešā novērošana, integrētā kvadrāta kļūda

Aleksandrs Andronovs. Par regresijas funkcijas, pamatotas uz *resampling*, neparametriskā intervāla novērtējumu, *Computer Modelling and New Technologies,* 11.sēj., Nr.1, 2007, 57.-61. lpp.

Tiek apskatīts neparametriskais regresijas modelis *E*(*Y*) = *m*(*x*), kur *Y* ir atkarīgais mainīgais, *x* ir *d* -dimensionālo neatkarīgo mainīgo (regresoru) vektors un *m* ir nezināma funkcija. Neatkarīgo novērojumu secība (Y_i, x_i,) *i* = 1, 2, …, *n*, ir pieejama. Mūsu mērķis ir uzbūvēt augšējās pārliecības saikni *m*(*x*), kas atbilst varbūtībai *y.* Tiek pielietota *resampling* pieeja. Piedāvātās metodes ļauj aprēķināt patieso virsmas varbūtību.

Atslēgvārdi: neparametriskā regresija, intervāla novērtējums, *resampling*

Jekaterina Žukovska. Vispārināto lineāro modeļu pielietošana pasažieru aviopārvadājumu prognozēšanai ES valstīs, *Computer Modelling and New Technologies,* 11.sēj., Nr.1, 2007, 62.-72. lpp.

Tika apskatīti dažādi regresijas modeļi pasažieru aviopārvadājumu prognozēšanai ES valstīs. Tika parādīti divi atsevišķi paņēmieni iepriekš minētam prognozēšanas uzdevumam. Pirmais paņēmiens ir klasiskā lineārā regresijas metode un otrais paņēmiens ir vispārinātā modeļa izmantošana. Apskatītie regresijas modeļi satur vairākus ietekmējošus faktorus un to kombinācijas. Tika parādītas vispārinātā regresijas modeļa priekšrocības salīdzinājumā ar klasisko regresijas modeli.

Atslēgvārdi: pasažieru aviopārvadājumi, prognozēšana, vispārinātais lineārais modelis

Diana Santalova. Dzelzceļu kravu pārvadājumu prognozēšana ES valstīs uz vienindeksa modeļa bāzes, *Computer Modelling and New Technologies,* 11.sēj., Nr.1, 2007, 73.-83. lpp.

Šajā rakstā tiek apskatīti regresijas modeļi, kas apraksta dzelzceļa kravu pārvadājumus Eiropas Savienības dalībvalstīs. Modeļi satur tādus noteicošus faktorus katrai valstij, kā: dzelzceļu līniju garums, iekšzemes kopprodukts uz vienu iedzīvotāju cenu standartos utt. Visi aprēķini tika veikti, pamatojoties uz statistiskiem datiem, kas ņemti no EUROSTAT YEARBOOK 2005 gadagrāmatas. Tika salīdzinātas divas novērtēšanas pieejas: klasiskais lineārais regresijas modelis un vienindeksa modelis. Tika pārbaudītas hipotēzes par pavadmainīgo nenozīmību un regresijas modeļa korektumu un tika pielietota krustveida-pārbaudes pieeja. Izdarītā analīze parādīja vienindeksa modeļa neapšaubāmas priekšrocības.

Atslēgvārdi: kravu pārvadājumi, prognozēšana, vienindeksa modelis

COMPUTER MODELLING & NEW TECHNOLOGIES

ISSN 1407-5806 & ISSN 1407-5814 (on-line)

THE JOURNAL IS DESIGNED FOR PUBLISHING PAPERS CONCERNING THE FOLLOWING FIELDS OF RESEARCH:

- mathematical and computer modelling
- mathematical methods in natural and engineering sciences
- physical and technical sciences
- computer sciences and technologies
- semiconductor electronics and semiconductor technologies
- aviation and aerospace technologies
- electronics and telecommunication
- navigation and radar systems
- telematics and information technologies
- transport and logistics
- economics and management
- social sciences

Articles can be presented in journal in English. All articles are reviewed.

EDITORIAL CORRESPONDENCE

Transporta un sakaru institūts (Transport and Telecommunication Institute)
Lomonosova 1, Riga, LV-1019, Latvia. Phone: (+371)7100593. Fax: (+371)7100535.
E-mail: journal@tsi.lv, http:// www.tsi.lv

COMPUTER MODELLING AND NEW TECHNOLOGIES, 2007, Vol. 11, No.1

Scientific and research journal of Transport and Telecommunication Institute (Riga, Latvia)
The journal is being published since 1996.

PREPARATION OF CAMERA-READY TYPESCRIPT: COMPUTER MODELLING AND NEW TECHNOLOGIES

1. In order to format your manuscript correctly, see the Page Layout Guideline for A4 (21 cm x 29,7 cm) paper size. Page Layout should be as follows: Top – 3 cm, Bottom – 3 cm, Left – 3 cm, Right – 3 cm.
2. Maximum length for the article is **10 pages**.
3. **Using of other Styles with the exception of Normal is not to be allowed!**
4. ***Articles*** should be Times New Roman typeface, single-spaced.
5. The article should include:
   - title;
   - author's name(s) and information (institution, city, country, the present address, phones, and e-mail addresses);
   - abstract (100-150 words);
   - keywords (max. 6);
   - introduction – clear explanation of the essence of the problem, previous work, purpose of the research and contribution;
   - description of the research;
   - conclusion section (this is mandatory) – should clearly indicate on the advantages, limitations and possible applications;
   - references.

   Attention! First name, last name, the title of the article, abstract and keywords must be submitted in the English and Latvian languages (in Latvian it is only for Latvian authors) as well as in the language of the original (when an article is written in different language).
6. The text should be in clear, concise English (or other declared language). Please be consistent in punctuation, abbreviations, spelling (***British English***), headings and the style of referencing.
7. ***The title of the article*** – 14 point, UPPERCASE, style Bold and centred.
8. ***Author's names*** – centred, type size 12 point, Upper and lower case, style Bold Italic.
9. ***Author's information*** – 10 point, Upper and lower case, style Italic, centred.
10. ***Abstract and keywords*** – 8 point size, style Normal, alignment Justify.
11. ***The first level Headings*** – 11 point, Upper and lower case, style Bold, alignment Left. Use one line space before the first level Heading and one line space after the first level Heading.
12. ***The second level Headings*** – 10 point, Upper and lower case, style Bold, alignment Left. One line space should be used before the second level Heading and 1/2 line space after the second level Heading.
13. ***The third level Headings*** – 10 point, Upper and lower case, style Italic, alignment Left. One line space should be used before the second level Heading and 1/2 line space after the third level Heading.
14. ***Text*** of the article – 10 point, single-spaced, alignment Justify.
15. The set of ***formulas*** on application of fonts, signs and a way of design should be uniform throughout the text. The set of formulas is carried out with use of editors of formulas MS Equation 3.0 or MathType. The formula with a number – the formula itself should be located on the left edge of the text, but a number – on the right one. Font sizes for equations are the following: 11pt – full, 7pt – subscripts/superscripts, 5pt – sub- subscripts/superscripts, 16pt – symbols, 11pt – subsymbols.
16. All ***Figures*** – must be centred. Figure number and caption always appear below the Figure, type size 8 point.

Figure 1. This is figure caption

Diagrams, Figures and Photographs – must be of high quality, in format *.TIFF, *.JPG, *.BMP with resolution not less than 300 dpi. Also formats *.CDR, *.PSD are possible. Combination of Figures in format, for instance, *.TIFF with elements of the in-built Figure Editor in MS Word is prohibited.

17. ***Table Number and Title*** – always appear above the Table. Alignment Left. Type size 8 point. Use one line space before the Table Title, one line space after the Table Title and 1/2 line space after the Table.

Table 1. This is an example of a Table

Heading	**Heading**	**Heading**
Text	Text	Text
Text	Text	Text

18. ***References*** in the text should be indicated by a number in square brackets, e.g. [1].
References should be numbered in the order cited in the manuscript. The correct format for references is the following:

Article: author, title, journal (in italics), volume and issue number, year, inclusive pages
Example: 1. Amrahamsson M., Wandel S. A Model of Tearing in Third – Party Logistics with a Service Parts Distribution Case Study, *Transport Logistics*, Vol. 1, No 3, 1998, pp. 181-194.

Book: author, title (in Italics), location of publishers, publishers, year, whole pages
Example: 2. Kayston M. and Fried W. R. *Avionic Navigation Systems.* New York: John Wiley and Sons Inc, 1969. 356 p.

Conference Proceedings: author; title of an article; proceedings (in italics); title of a conference, date and place of a conference; publishing house, year, concrete pages
Example: 3. Canales Romero J. A First Step to Consolidate the European Association of Aerospace Students in Latvia (Presented by the Munich Local Group). In: *Research and Technology – Step into the Future: Programme and Abstracts. Research and Academic Conference, Riga, Latvia, April 7–11, 2003, Transport and Telecommunication Institute.* Riga: TTI, 2003, p. 20.

19. ***Authors Index***
Editors form the author's index of a whole Volume. Thus, all contributors are expected to present personal colour photos with the short information on the education, scientific titles and activities.
20. ***Acknowledgements***
Acknowledgements (if present) mention some specialists, grants and foundations connected with the presented paper. The first page of the contribution should start on page 1 (right-hand, upper, without computer page numbering). Please paginate the contributions, in the order they are to be published. Use simple pencil only.
21. **Articles poorly produced or incorrectly formatted may not be included in the proceedings**.

The K. Kordonsky Charitable Foundation

The 7th International Conference

RELIABILITY and STATISTICS in TRANSPORTATION and COMMUNICATION (RelStat'07)

24-27 October 2007. Riga, Latvia

PURPOSE

The purpose of the conference is to bring together academics and professionals from all over the world to discuss the themes of the conference:

- Theory and Applications of Reliability and Statistics
- Reliability and Safety of Transport Systems
- Rare Events and Risk Management
- Modelling and Simulation
- Intelligent Transport Systems
- Transport Logistics
- Education Programmes and Academic Research in Reliability and Statistics

DEDICATION

The Conference is devoted to the memory of Prof. Kh.Kordonsky.

OFFICIAL LANGUAGES

English and Russian will be the official languages of the Conference.

SUPPORTED BY:

Transport and Telecommunication Institute (Latvia) and
The K. Kordonsky Charitable Foundation (USA) in co-operation with:
Latvian Transport Development and Education Association (Latvia)
Telecommunication Association of Latvia (Latvia)
Latvian Academy of Science (Latvia)
Baltic Operations Research Society

SPONSORED BY

Transport and Telecommunication Institute (Latvia)
The K. Kordonsky Charitable Foundation (USA)
Latvian Operations Research Society
PAREX bank (Latvia)

HOSTED BY

Transport and Telecommunication Institute (Latvia)

SECRETARIAT

Prof. Igor Kabashkin, Latvia – Chairman
Ms. Inna Kordonsky-Frankel, USA – Co-Chairman
Prof. Irina Yatskiv, Latvia – Co-Chairman
Ms. Elena Rutkovska, Latvia – Secretary

DEADLINES AND REQUIREMENTS

Submission of abstracts:	**15 May**	**2007**
Acceptance of abstracts:	**29 May**	**2007**
Submission of final papers:	**3 July**	**2007**
Acceptance of final papers:	**4 September**	**2007**

Abstracts (about 600 words in length) and papers submitted for review should be in English and, should present a clear and concise view of the motivation of the subject, give an outline, and include information on all authors (the full name, affiliation, address, telephone number, fax number, and e-mail address of the corresponding author).

Submitted abstracts and papers will be reviewed. Accepted and invited papers will be published in the proceedings of the conference and in the journal "Transport and Telecommunication" (ISSN 1407-6160).

Instruction for papers preparing can be found on the conference WWW page: http://RelStat.tsi.lv.

INVITED SESSIONS (workshops)

Proposals for invited sessions (workshops) within the technical scope of the conference are accepted. Each proposal should describe the theme and scope of the proposed session. The proposal must contain the title and theme of the session and a list of paper titles, names and email addresses of the corresponding authors. Session proposals and paper must be submitted by **21 May 2007**.

REGISTRATION FEE

The registration fees will be **Euro 100** before 10 September 2006, and **Euro 150** after this date. This fee will cover the participation in the sessions, coffee breaks, daily launch, hard copy of the conference proceedings.

VENUE

Riga is the capital of the Republic of Latvia. Thanks to its geographical location, Riga has wonderful trade, cultural and tourist facilities. Whilst able to offer all the benefits of a modern city, Riga has preserved its historical charm. It's especially famous for its medieval part – Old Riga.

Old Riga still preserves many mute witnesses of bygone times. Its old narrow streets, historical monuments, organ music at one of the oldest organ halls in Europe attract guests of our city. In 1998 Old Riga was included into the UNESCO list of world cultural heritage.

ACCOMMODATION

A wide range of hotels will be at the disposal of participants of the conference and accompanying persons (http://eng.meeting.lv/hotels/latvia_hotels.php).

FURTHER INFORMATION

Contact:

Elena Rutkovska, Secretary, RelStat'07
Transport and Telecommunication Institute
Lomonosova 1, Riga, LV-1019, Latvia
Telephone: +(371)-7100665
Fax: +(371)-7100535
E-mail: RelStat@tsi.lv
http:// RelStat.tsi.lv