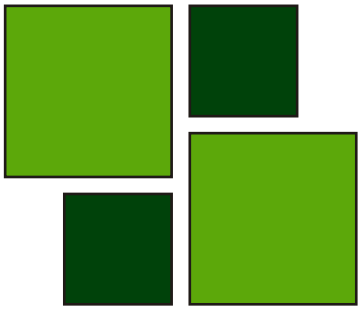


WWW.JAQM.RO

JOURNAL OF APPLIED QUANTITATIVE METHODS

International Symposium on Stochastic Models in Reliability Engineering, Life Sciences and Operations Management (SMRLO'10)

Vol. 5
No. 3
Fall
2010

ISSN 1842–4562

JAQM Editorial Board

JAQM Advisory Board

JOURNAL OF APPLIED QUANTITATIVE METHODS

Contents

Page

International Symposium on Stochastic Models in Reliability Engineering, Life Sciences and Operations Management (SMRLO'10)

JOURNAL OF APPLIED QUANTITATIVE METHODS

Contents

Prof. Zohar Laslo, Dean of Industrial Engineering and Management Departments, SCE - Shamoon College of Engineering, Israel, together with Prof. N. Balakrishnan from McMaster University in Canada, and Prof. Alan Hutson from the State University of New-York at Buffalo, USA, has been member of the Conference Council at the recent International Symposium on Stochastic Models in Reliability Engineering, Life Sciences and Operations Management (SMRLO'10) which was held in Beer-Sheva Campus of SCE on February 8-11, 2010.

SPECIAL ISSUE EDITORIAL

A stochastic process, being in essence a random one, is a process which development is dependent on irregular factors, and the system behaving under such process is subject to transfer from an initial fixed position to a variety of final destinations, each one with its own probability to occur. This is of course in contrary to deterministic processes, when a certain initial position of the system determines one-and-for-all its final conclusive destination. When a system comprises activities and processes with a high amount of uncertainty and logistic ties between the reason and the consequence are complex and obscure, and especially when internal details of the system's functioning have not been studied in depth, there is no alternative to describing such a system but implementing stochastic models. In fact, stochastic processes serve nowadays as models for systems in a vast variety of applications. Examples are overwhelming in engineering, in life sciences and specifically in logistics management (including economics, demography and many more areas).

This Special Issue hosted by the Journal of Applied Quantitative Methods (JAQM) includes a selected representative sample of some of the research papers delivered at the International Symposium on Stochastic Models in Reliability Engineering, Life Sciences and Operations Management (SMRLO'10) that was held in Beer-Sheva Campus of SCE – Shamoon College of Engineering on February 8-11, 2010. SMRLO'10 is already the second international symposium on stochastic modeling taking place at SCE (the first one, named SMRSSL'05, has taken place on February, 2005). SMRLO'10 accommodated more than 120 guest participants from Ukraine, Italy, Ireland, Great Britain, USA, Bulgaria, Germany, South Africa, India, Netherlands, Taiwan, Turkey, Greece, Latvia, Norway, China, Singapore, Slovakia, Spain, Serbia, Poland, Portugal, Czech Republic, France, Canada, Cyprus, Romania, Russia, Sweden, Switzerland, and 150 participants from Israel. The Scientific Program Committee received over 200 abstracts from potential participants willing to share their research results with colleagues at the Symposium. Within the four days of SMRLO'10, over 185 presentations have been made at open plenary sittings and at parallel sessions, 17 from which were delivered by SCE academic staff and students.

SCE – Shamoon College of Engineering is nowadays a leading academic institution in providing engineering education to vast sectors of Israeli population. With over 4,000 students studying 6 different engineering disciplines, it has become actually the key player

not only in the south of the country where it is geographically located, but on a national and even international level as well. This special issue acknowledges long-years fruitful scientific ties between SCE and JAQM. On behalf of the SMRLO'10 Conference Council, joined by Prof. Ilya Gertsbakh the Chair of the Scientific Program Committee and Dr. Ilia Frenkel the Chair of the Organization Committee, I am thankful to JAQM Editorial Board which enabled this Special Issue.

Zohar Laslo
Guest Editor

JOURNAL OF APPLIED QUANTITATIVE METHODS

International Symposium on Stochastic Models in Reliability Engineering, Life Sciences and Operations Management (SMRLO'10)

UPON SCHEDULING AND CONTROLLING LARGE-SCALE STOCHASTIC NETWORK PROJECTS

Doron GREENBERG

PhD, Department of Economics and Business Administration, Faculty of Social Science, Ariel University Center (AUC) of Samaria, Ariel, Israel

E-mail: dorongreen2@gmail.com

Dimitri GOLENKO-GINZBURG

Prof, Department of Industrial Engineering and Management (Emeritus), Ben-Gurion University of the Negev, Beer-Sheva, Israel
& Department of Industrial Engineering and Management, Ariel University Center (AUC) of Samaria, Ariel, Israel

E-mail: dimitri@bgu.ac.il

Abstract: *The problem of controlling large-size stochastic network projects of PERT type is considered. A conclusion is drawn that the need of proper control models for PERT projects is very important. The authors suggest aggregating the initial model in order to modify the latter to an equivalent one, but of medium or small-size.*
For those network models effective on-line control algorithms are already developed. After observing the project's output at a routine control point and introducing proper control actions the aggregated network is transformed to the initial one, and the project's realization proceeds.
The developed control techniques are especially effective for those R&D projects, when an on-line control has to be undertaken under a chance constraint. The suggested control model can be regarded as an additional tool to help the project manager to realize the project in time.

Key words: *project management; on-line control; scheduling; network project; generalized network models*

1. Introduction

In recent years the problems associated with controlling projects by means of network analysis have not been discussed extensively in the literature. Scanty publications refer mostly to network modelling and to the calculation of activity network parameters. However, little investigation has been undertaken in the area of decision-making and determining control actions while controlling stochastic network projects. The main questions: "How a PERT project should be controlled?" and "What are the main stages of controlling PERT projects?" have not previously received satisfactory answers, especially for

highly complicated long-term projects under random disturbances. It can be well-recognized from studying the literature on planning and control techniques in project management that an overwhelming majority of modern management systems use only PERT techniques to plan and control projects with uncertainty [6]. This occurs because PERT is simpler than other more complicated techniques [2-4] with a high level of uncertainty. However, the PERT conception deals with random disturbances since a PERT network comprises activities of random duration.

It can be clearly recognized that in the last two decades, various control problems in project management, especially for PERT projects, have been the subject of lengthy debate and very sharp criticism [1, 5-7]. In our opinion, the main reason that, in practice, those projects are all usually completed late and remain uncontrolled is that PERT projects are carried out under random disturbances (new estimates of a random nature without any previous experience, random activities' durations, periodical revisions of networks over time due to random emergency situations, etc.). However, project managers usually [6] avoid probabilistic terms since they arc not sufficiently trained. They are trying to control highly complicated projects with uncertainly by using deterministic techniques. This leads to biased estimates that usually underestimate the actual time needed to accomplish the project. Therefore the project's due date can rarely be met. *Thus the need of proper control models for PERT projects is very important*.

In our opinion, there is another important reason for numerous failures of PERT techniques in project management. This is because the models are too complicated to be effective. They are not flexible. Usually, they incorporate both scheduling and control techniques. But since it is practically impossible to develop *a proper deterministic schedule for a project under random disturbances*, such models are not adequate to the real life. Therefore the control procedures are also non-effective.

We suggest using a control model only at several control (inspection) points in order to determine the next routine inspection point and the project's speed to proceed with until that next control point. Such control techniques can be applied only to a network model with a medium amount of activities (up to 50-100 activities). Thus, the problem is to modify the initial network model (which for some projects may comprise a tremendous amount of activities) to an equivalent one, but of medium or small size.

For such a model an activity is equivalent to a subnetwork (a fragment) of the initial network. Such aggregated, small-size networks for construction projects of deterministic type have been developed in [9].

In the next section we will describe the general idea of an aggregation for PERT type projects with activities of random durations.

2. Developing Enlarged Aggregated Networks with Random Activity Durations

According to the project's Work Breakdown Structure (WBS) [10] an initial network is presented in the form of a group of lists of initial activities. The name of the activity is taken from the WBS.

We will henceforth call a fragment a list of activities together with all the links both entering and leaving that fragment. The step-by-step procedure of developing an aggregated network is as follows:

Given:

- activities (i, j) entering the PERT initial network $G(N, A)$;
- random activity durations t_{ij} with pregiven density distribution.

Step 1. Simulate random durations t_{ij}, $(i, j) \subset G(N, A)$.

Step 2. On the basis of simulated values t_{ij} calculate for each $i \in N$ the earliest moment of the event's realization, $T^{\xi}(i)$, where ξ denotes the index of the simulation run.

Step 3. Repeat Steps 1-2 M times in order to obtain representative statistics.

Step 4. Calculate

$$T_{ear}(i) = \min_{1 \le \xi \le M} T^{\xi}(i);$$

$$T_{lat}(i) = \max_{1 \le \xi \le M} T^{\xi}(i).$$

Step 5. By using decomposition methods [9, 10] subdivide the initial set into enlarged fragments. Each fragment comprises a list of detailed activities together with all links connecting activities entering the list ("internal" links) as well as all "external" links connecting the fragment with other fragments.

Steps 6-11 have to be realized for each fragment $F \subset G(N, A)$ separately.

Step 6. Determine two events i_{st}^{F} and i_{fin}^{F} which we will henceforth call the start and the finish events of fragment F:

$i_{st}^{F} \in F$ delivers the minimum to $\underset{i \in F}{Min}\{T_{ear}(i)\}$ and

$i_{fin}^{F} \in F$ delivers the maximum to $\underset{i \in F}{Max}\{T_{lat}(i)\}$, where $T_{ear}(i)$ and $T_{lat}(i)$ have been calculated on Step 4.

Step 7. For both events i_{st}^{F} and i_{fin}^{F} calculate the earliest and the latest time moments (see Step 4): $T_{ear}(i_{st}^{F})$, $T_{lat}(i_{st}^{F})$, $T_{ear}(i_{fin}^{F})$, $T_{lat}(i_{fin}^{F})$.

Step 8. Calculate the minimal fragment's duration

$$\tau_{F}^{min} = T_{ear}(i_{fin}^{F}) - T_{lat}(i_{st}^{F}).$$

Step 9. Calculate the maximal fragment's duration

$$\tau_{F}^{max} = T_{lat}(i_{fin}^{F}) - T_{ear}(i_{st}^{F}).$$

Step 10. Assume that the fragment's duration τ_F is a random variable with a β-distribution density function

$$p_F(x) = \frac{12}{\left(\tau_F^{max} - \tau_F^{min}\right)^4}\left(x - \tau_F^{min}\right)\left(\tau_F^{max} - x\right)^2$$

with the mathematical expectation

$$\tilde{\tau}_F = \left(3\,\tau_F^{min} + 2\,\tau_F^{max}\right)\cdot 0.2 \,.$$

Such a distribution has been successfully used over a long time in various network projects [2].

Step 11. External links (arrows) entering and leaving fragment F are determined [9, 10]. For each external arrow the corresponding receiver (emitter) is calculated in percentage of the fragment's duration.

After realizing Steps 6-11 for each fragment $F \subset G(N, A)$ the enlarged aggregate network with random fragments' durations is determined. As outlined above the aggregated network must be of a small or a medium size. The model enables applying on-line control techniques to introduce proper control actions.

3. On-Line Control Problems

For most network projects under random disturbances the progress of the project cannot be inspected and measured continuously, but only at preset inspection points. An on-line control determines both inspection points and control actions to be introduced at those points in order to alter the progress of the project in the desired direction. Such control actions may be as follows:

a) to redistribute the budget among the project activities in order to enhance the project's speed,
b) to introduce additional shifts, etc., to change the speed of the progress of the project without using additional resources, etc.

Such control actions [3, 4] usually have the tendency to minimize either the number of inspection points, or the average project's speed subject to a chance constraint to meet the project's due date on time. The corresponding control algorithms are described in [3]. They have been applied to medium size construction projects [4].

After realizing the control actions the modified aggregated network is transformed to the initial network [8].

Consider a medium-size PERT type network model with a due date D. A desirable probability p^* that in practice enables completion of the project on time is pregiven. At each control moment t_g the project management may introduce several possible alternative speeds v_{t_g} to proceed with until the next control point. Let V_t be the project's output (project volume) observed at control point $t > 0$ and let the project's target (goal) be V^*. Denote $\Pr\left(t_g, v_{t_g}\right)$ the confidence probability to accomplish the project on time after introducing speed v_{t_g} and control point t_g.

The main control problem [3] is to determine both control (inspection) points t_g $(g = 1,\ldots,N)$ and speeds to proceed with from that point on until the next adjacent control point t_{g+1}, in order to minimize the number N of inspection points:

$$\underset{\{t_g, v_{t_g}\}}{Min}\ N \tag{1}$$

subject to

$$Pr\left\{t_g, v_{t_g}\right\} \geq p^*, \tag{2}$$

$$t_0 = 0, \tag{3}$$

$$t_N = D, \tag{4}$$

$$t_{g+1} - t_g \geq \Delta. \tag{5}$$

Pregiven value Δ is usually introduced to force convergence.

Note that if introducing a control action results in determining the project's speed v_{t_g} to proceed with until the next control point t_{g+1} and if several alternative speeds can be chosen, then the optimal control action enables using the minimal speed to develop the project honouring chance constraint (2) [3, 4].

Control model (1-5) is a stochastic optimization problem with a non-linear chance constraint and a random number of optimized variables. The problem is too difficult to solve in the general case. Thus, heuristic control algorithms have been developed [4] to determine the next inspection point t_{g+1}. Two algorithms are considered:

A. Using sequential statistical analysis to maximize the time span $\Delta t_g = t_{g+1} - t_g$.

B. Using the idea of a risk averse decision-maker.

Algorithm A [3, 4] solves the on-line control problem as follows: to maximize the objective $(t_{g+1} - t_g)$ subject to (3-5) and

$$Pr\left\{V_t \geq V_t^*\left(t_g\right)\right\} \geq p^*, \quad \forall t : t_g \leq t \leq t_{g+1}, \tag{6}$$

where $V_t^*\left(t_g\right)$ is a trajectory control curve connecting two points $\left(t_g, V_{t_g}\right)$ and $\left(D, V^*\right)$. This problem can be solved by determining the maximal value T^* satisfying

$$T^* = \underset{t_g < t \leq D}{Max}\left\{t : \Psi(q_t) \geq p^*\right\}. \tag{7}$$

Here

$$\Psi(x) = \frac{1}{\sqrt{2\pi}}\int_x^{\infty} e^{-\frac{u^2}{2}}\,du, \quad q_t = \frac{\overline{H}}{S^2(H_t)}, \quad H_t = V_t - V_t^*\left(t_g\right), \tag{8}$$

and $\overline{H}_t$ and $S(H_t)$ are the mean value and the standard deviation of random value H_t, correspondingly. In practice, T^* can be calculated via simulation with a constant step of

length Δ. The procedure of increasing t step-by-step is followed until (7) ceases to hold. It can be well-recognized that $t_g + T^* = t_{g+1}$ holds.

Algorithm B is based on the concept of a risk averse decision-maker [3, 8]. Given a routine inspection point t_g, the project's output observed at that moment V_{t_g} and the speed v_{t_g} to be introduced at the moment t_g up to the next inspection point, the problem is to determine that next point t_{g+1}. Just as for Algorithm A, the objective is to maximize the time span $t_{g+1} - t_g$. Value t_{g+1} is determined so that even if the project develops most unfavourably in the interval $[t_g, t_{g+1}]$, i.e., with the minimal rate v'_{t_g}, then introducing the highest speed $v_{\max}$ at moment t_{g+1} enables the project to meet its target on time, subject to the chance constraints.

Value t_{g+1} is determined via a "risk averse" heuristic [3, 8]

$$V_{t_g} + v'_{t_g}\left(t_{g+1} - t_g\right) + \bar{v}_{max}\left(D - t_{g+1}\right) = V^* \quad (9)$$

Usually Algorithm B is more efficient than Algorithm A. Both algorithms can be applied to those projects when the output can be measured at inspection points in quantitative attitudes, e.g. in percentages of the whole target (goal). This often happens in various construction projects. Another fruitful application area is PERT-COST projects when the assigned budget defines the project's target while the remaining budget actually defines the remaining project's volume.

4. Application

The outlined above methodology together with on-line techniques have been used successfully for various construction projects [4]. Note that simulating on-line control procedures for a medium-size project (40-50 activities) takes about 4 hours on PC-486 using model (6-8). A risk averse method (9) has a higher speed. Thus, controlling a project comprising several hundred activities offers quite a lot of computational time. Such projects of large-size need decomposition in order to be controlled.

5. Conclusions

After introducing the control actions outlined above the modified medium-size aggregated network is transformed to the initial network [9, 10] and the project's realization proceeds.

All other procedures at the project's level, e.g. scheduling procedures, are carried out for the initial network between two adjacent control points. Although those procedures usually comprise biased estimates and errors, they are periodically corrected by introducing proper control actions. *That is why those procedures in combination with control actions are more effective than without controlling the project in inspection points.*

In conclusion, the on-line control model has to be used as *an additional tool* in order to help the project manager to realize the project on time. Implementing the model does not result in undertaking any revisions in traditional PM procedures.

References

1. Bobrowski, P.M. **Project management control problems: An information systems focus**, Project Management Journal, 20(2), 1989, pp. 11-16
2. Golenko-Ginzburg, D. **Statistical Models in Network Planning and Control**, Nauka, Moscow, 1966 (in Russian)
3. Golenko-Ginzburg, D., Gonik A. **Project planning and controlling by stochastic network models**, in book: "Managing and Modelling Complex Projects" (T. Williams ed.), NATO ASI Series, Kluwer Academic Publication, 1997, pp. 21-43
4. Golenko-Ginzburg, D. and Gonik, A. **On-line control models for network construction projects**, J. Opl. Res. Soc., 47, 1996, pp. 266-283
5. Hughes, M.W. **Why projects fail: An effect of ignoring the obvious**, Ind. Engnr., 18(4), 1986, pp. 14-18
6. Pearson, A.W. **Planning and control in research and development**, Omega, 18(6), 1990, pp. 573-581
7. Shoenberg, R.I. **Why projects are always late: A rationale based on manual simulation of a PERT/CPM network**, Interfaces, 11(5), 1981, pp. 66-70
8. Aronov, I., Liubkin, S. and Golenko-Ginzburg, D. **Controlling large-size stochastic network projects**, Proceedings of the International Symposium "Project Management in Transfer Economics", Moscow, Russia, June 4-6, 1997
9. Voropaev, V.I. and Liubkin, S.M. **Aggregation of Generalized Networks for Construction Projects**, SIBNIIGIM, Krasnoyarsk, 1989 (in Russian)
10. Voropaev, V.I. and Liubkin, S.M. **Managing complex projects by active hierarchical systems**, in book: "Managing and Modelling Complex Projects" (T. Williams ed.), NATO ASI Series, Kluwer Academic Publication, 1997, pp. 221-236

JOURNAL OF APPLIED QUANTITATIVE METHODS

International Symposium on Stochastic Models in Reliability Engineering, Life Sciences and Operations Management (SMRLO'10)

SOLVING NONLINEAR OPTIMIZATION PROBLEMS BY MEANS OF THE NETWORK PROGRAMMING METHOD

Vladimir N. BURKOV

Prof., Institute of Control Sciences of V.A. Trapeznikov,
Russian Academy of Sciences,
Moscow, Russia

E-mail: vlab17@bk.ru

Irina V. BURKOVA

PhD Candidate, Institute of Control Sciences of V.A. Trapeznikov,
Russian Academy of Sciences,
Moscow, Russia

E-mail: irbur27@gmail.com

Abstract: *We suggest a new approach to solve discrete optimization problems, based on the possibility of presenting a function as a superposition of simpler functions. Such a superposition can be easily represented in the form of a network for which the inputs correspond to variables, intermediate nodes – to functions entering the superposition, and in the final node the function is calculated. Due to such representation the method has been called the method of network programming (in particular, dichotomic). The network programming method is applied for solving nonlinear optimization problems. The concept of a dual problem is implemented. It is proved that the dual problem is a convex programming problem. Necessary and sufficient optimality conditions for a dual problem of integer linear programming are developed.*

Key words: *network programming; nonlinear optimization; dual problem; integer linear programming*

1. Introduction

Problems of nonlinear optimization (in particular, discrete optimization) refer to the class of so-called NP-difficult problems for which no effective methods of exact solution do exist. Some general approaches are available, among others the branch and bounds method and the method of dynamic programming [1]. Unfortunately, the dynamic programming method is applicable only to a narrow class of problems. The efficiency of the branch and bounds method depends essentially on accuracy of the upper and lower estimates (bounds).

To assess those estimates the method of multipliers of Lagrange [1] is developed. These methods are known from the past 60-s, and since then more than they have not been improved significantly.

In 2004 V.N.Burkov and I.V.Burkova suggested a new approach to solve discrete optimization problems, based on the possibility of presenting a function as a superposition of simpler functions. Such a superposition can be easily represented in the form of a network for which the inputs correspond to variables, intermediate nodes – to functions entering the superposition, and in the final node the function is calculated. Due to such representation the method has been called the method of network programming [2] (in particular, dichotomic). This method is applicable to cases when the goal function and restriction functions obtain identical network structure. For such cases network node optimization problems, simpler than the pregiven ones, are solved. The problems' solution for the final node presents the upper (or lower) estimates for the given problem. For the case when the network structure is a tree, the solution becomes an exact one. The Bellman's dynamic programming method for which the network structure comprises tree branches, becomes thus a particular case of the more generalized proposed approach. A variety of problems for which the dynamic programming method is inapplicable, have been solved by the network programming method.

In the present paper the network programming method [2] is applied to nonlinear programming problems. The concept of a dual problem, for which one of the feasible (but usually non-optimal) solutions is obtained, is suggested by means of multipliers of Lagrange. It is proved that the dual problem is a convex programming problem. Necessary and sufficient optimality conditions for a dual problem of integer linear programming are developed.

2. The Network Form of a Nonlinear Programming Problem

Let's consider a problem of nonlinear programming - to determine $x = \{x_i,\ i = \overline{1,\ n}\}$, satisfying

$$f(x) \to \max \tag{1}$$

subject to

$$\varphi_j(x) \le b_j,\quad j = \overline{1,m}, \tag{2}$$

$$x \in X_{m+1}. \tag{3}$$

On Figure 1 the network representation of restrictions (2-3) is given. Here X_i denotes the j -th restriction (2), $j = \overline{1,m}$.

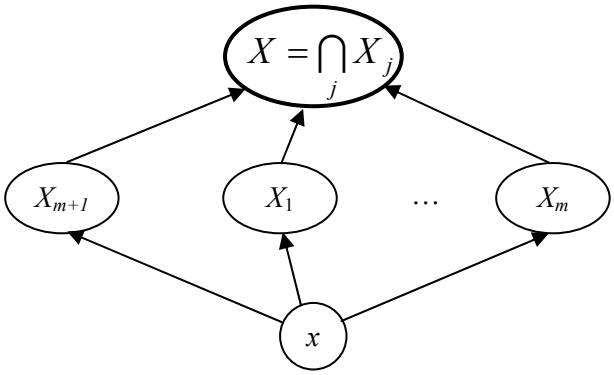


Figure 1. Network representation of restrictions

In order to apply the network programming method we have to represent the goal function with the same network structure. For this purpose we will present *f(x)* in the form

$$f(x) = \sum_{j=1}^{m} h_j(x) + h_{m+1}(x), \tag{4}$$

where $h_j(x)$ stands for functions which deliver solutions for the below problems (5-6).

In each vertex of the network structure several optimization sub-problems with one restriction are solved. The first *m* sub-problems are as follows:

$$\begin{aligned} &\max h_j(x), \\ &\varphi_j(x) \le b_j. \end{aligned} \tag{5}$$

while the *(m+1)* -th sub-problem looks as follows:

$$\max_{x \in X} h_{m+1}(x) = \max_{x \in X_{m+1}} \left[f(x) - \sum_{j=1}^{m} h_i(x) \right]. \tag{6}$$

Denote $F_j(h)$ the value of the goal function for the optimal solution of the *j* -th sub-problem.

Theorem 1. Linear model

$$F(h) = \sum_{j=1}^{m} F_j(h_j) + F_{m+1}(h) \tag{7}$$

delivers the upper estimate for a pre-given problem.

Proof. All feasible solutions (1-3) are feasible for all sub-problems (5-6), and any feasible solution x satisfies

$$\sum_{j=1}^{m+1} h_j(x) = f(x)$$

.

Therefore $F(h) \ge f(x)$ for any feasible x.

3. The Dual Problem

It is obvious to suggest the problem of determining functions $h_j(x)$, $j = \overline{1, m}$, which minimize the upper estimate (7). This problem is, in essence, a generalized dual problem for the initial problem of nonlinear programming. The reasons for this are as follows. First, as shown below (see Example 1), one of the feasible solutions of the generalized dual problem

is a minimax of function of Lagrange. Note that determining the minimax Lagrange function is often called the dual one for the problem of nonlinear programming. Second, for a problem of linear programming without an integer solution the generalized dual problem is a usual dual problem of linear programming (see Section 4).

Theorem 2. Function *F(h)* is a convex one.

Proof. Let $h_1(x)$ and $h_2(x)$ be two solutions of a dual problem. Consider the solution

$$h(x)=\alpha h^1+(1-\alpha)h^2,\quad 0\le\alpha\le 1.$$

We obtain

$$F(h)=F\left[\alpha h^1+(1-\alpha)h^2\right]=\max_{x\in X}\left[\alpha\left(f(x)-\sum_{j=1}^{m}h_j^1(x)\right)+(1-\alpha)\left(f(x)-\sum_{j=1}^{m}h_j^2(x)\right)\right]+$$

$$+\sum_{j=1}^{m}\max_{\varphi_j(x)\le b_j}\left[\alpha h_j^1(x)+(1-\alpha)h_j^2(x)\right]\le\alpha\left[\max_{x\in X}\left(f(x)-\sum_{j=1}^{m}h_j^1(x)\right)+\sum_{j=1}^{m}\max_{\varphi_j(x)\le b_j}h_j^1(x)\right]+$$

$$+(1-\alpha)\left(\max_{x\in X}\left(f(x)-\sum_{j=1}^{m}h^2(x)\right)+\sum_{j=1}^{m}\max_{\varphi_j(x)\le b_j}h_j^2(x)\right)=\alpha F(h^1)+(1-\alpha)F(h^2).$$

The inequality stems from the evident reason that the maximum of the sum is less or equal to the sum of maxima.

Thus, the dual problem is a convex programming problem.

Example 1. Consider one of the feasible solutions of the dual problem, namely, $h_j(x)=\lambda_j\varphi_j(x),\ j=\overline{1,m}$. The first *m* sub-problems are as follows:

$$\lambda_j\varphi_j(x)\to\max$$

subject to

$$\varphi_j(x)\le b_j.$$

Evidently $F_j(h_j)\le\lambda_j b_j,\quad j=\overline{1,m}$. By means of this assertion together with (7), we finally obtain

$$F(\lambda)\le\max_{x\in X_{m+1}}\left(f(x)-\sum_{j=1}^{m}\lambda_j(\varphi_j(x)-b_j)\right)=\max_{x\in X_{m+1}}L(\lambda,x).\qquad(8)$$

Maximizing the right part (8) on λ is nothing but the method of multipliers of Lagrange. Thus, the method of multipliers of Lagrange provides a feasible solution of the dual problem (which, generally speaking, may be not an optimal one).

4. Upon One Integer Linear Programming Problem

Consider an integer linear programming problem as follows: determine an integer nonnegative vector x, to maximize

$$C(x)=\sum_{i=1}^{n}c_i x_i\qquad(9)$$

subject to

$$\sum_{i=1}^{n}a_{ij}x_i\le b_j,\quad j=\overline{1,m+1}.\qquad(10)$$

Take the last restriction in (10) as the set X_{m+1}. Sub-divide each value $c_i, i=\overline{1,m}$, on m partial values s_{ij} as follows:

$$s_{i;m+1} = c_i - \sum_{j=1}^{m} s_{ij}, \quad i=\overline{1,n}. \tag{11}$$

Solve $(m+1)$ sub-problems as follows: determine an integer nonnegative vector x, to maximize

$$S_j(x) = \sum_i s_{ij} x_i . \tag{12}$$

subject to

$$\sum_{i=1}^{n} a_{ij} x_i \le b_j . \tag{13}$$

Denote by $F_j(s)$ the value $S_j(x)$ providing the optimal solution for the j -th subproblem. According to Theorem 1

$$F(s) = \sum_{j=1}^{m} F_j(s_j) + F_{m+1}(s) \tag{14}$$

is an upper estimate for $C(x)$:

$$F(s) \ge C(x).$$

The dual problem: determine $\{s_{ij},\ i=\overline{1,n}, j=\overline{1,m}\}$, minimizing (14). Note that cancelling the requirement of integrality results in transforming problem (14) to a common dual linear programming problem [3].

To prove this accession consider problem (9-10) without the integrality requirement. In this case the estimation problems are easily solved, namely

$$F_j(s_j) = b_j \max_i \frac{s_{ij}}{a_{ij}}.$$

Denote

$$y_j = \max_i \frac{s_{ij}}{a_{ij}}, \quad j=\overline{1,m+1}.$$

Thus, the upper estimate for the objective of the initial problem looks as follows:

$$\Phi(y) = \sum_j y_j b_j . \tag{15}$$

Since $a_{ij} y_j \ge s_{ij}$, relation (11) transfers to

$$\sum_j a_{ij} y_j \ge c_i, \quad i=\overline{1,n}. \tag{16}$$

The dual problem is to minimize (15) subject to (16). This is a common dual linear programming problem.

Set $s_{ij} = \lambda_j \alpha_{ij}$, $i=\overline{1,n}, j=\overline{1,m}$. As outlined above, the problem boils down to the method of multipliers of Lagrange as follows: determine vector λ, minimizing

$$\max_{x \in X_{m+1}} \left(\sum_i c_i x_i - \sum_{j=1}^{m} \lambda_j \left(\sum_i a_{ij} - b_j \right) \right). \tag{17}$$

Consider necessary and sufficient conditions to obtain the optimal solution of the dual problem. Let *s* be a feasible solution. Denote $P_j(s_j)$ the set of optimal solutions for (*m*+1) sub-problems (12-13), $j = \overline{1, m+1}$.

Theorem 3. The necessary and sufficient condition to obtain the optimal solution *s* is the inability to solve inequality

$$\sum_j \max_{x \in P_j(s_j)} \sum_i y_{ij} x_i < 0 \qquad (18)$$

subject to

$$\sum_{j=1}^{m+1} y_{ij} = 0, \quad i = \overline{1, n}. \qquad (19)$$

Proof. Denote by y_{ij} small increments of s_{ij}. We will prove that relations (19) stem from (11). Indeed, it boils down from (11) that

$$\sum_{j=1}^{m+1} \left(y_{ij} + s_{ij} \right) = c_i \quad and \quad \sum_{i=1}^{m+1} s_{ij} = c_i$$

hold. The latter provides (19). The increment of value $F_j(s_j)$ is, obviously, equal

$$\Delta F_j = \max_{x \in P_j(s_j)} \sum_i y_{ij} x_i,$$

while the total increment satisfies

$$\Delta F = \sum_j \Delta F_j.$$

Since *s* is the optimal solution, Δ*F* cannot be negative. ■

Numerical Example 2. $x_i = 0, 1; i = \overline{1,4}$.

$$10x_1 + 8x_2 + 6x_3 + 7x_4 \to \max, \qquad (20)$$

$$6x_1 + 3x_2 + 2x_3 + 5x_4 \le 11, \qquad (21)$$

$$3x_1 + 5x_2 + 6x_3 + 3x_4 \le 11. \qquad (22)$$

Apply the method of multipliers of Lagrange, i.e. determine the minimum of λ functions

$$11\lambda + \max_{x \in X_2} \left[(10 - 6\lambda)x_1 + (8 - 3\lambda)x_2 + (6 - 2\lambda)x_3 + (7 - 5\lambda)x_4 \right],$$

where X_2 is determined by (22). With pre-set λ this is a one-dimensional knapsack problem. In case when the dependence of the right part of b_2 (see restriction (22)) from *n* is unknown, this problem turns to be NP-difficult [4]. However in practice, b_2 either does not depend on *n*, or is a linear function of *n*. In such cases, for integer parameters, the problem is efficiently solved by means of either dynamic or dichotomic programming. The determined optimal value $\lambda_0 = 1\frac{2}{9}$, with the upper estimate $F_0 = 21\frac{1}{3}$. This level λ_0 corresponds to the following values s_{ij}, $i = \overline{1,4}, j = 1,2$:

$$\begin{array}{llll} s_{11} = \lambda_0 a_{11} = 7\frac{1}{3}; & s_{21} = \lambda_0 a_{21} = 3\frac{2}{3}; & s_{31} = \lambda_0 a_{31} = 2\frac{4}{9}; & s_{41} = \lambda_0 a_{41} = 6\frac{1}{9}; \\ s_{12} = c_1 - s_{11} = 2\frac{2}{3}; & s_{22} = 4\frac{1}{3}; & s_{32} = 3\frac{5}{9}; & s_{42} = \frac{8}{9}. \end{array} \qquad (23)$$

Let's apply the network programming method.

Step 1. Determine necessary optimality conditions for solution (23). Consider:

$P_1(s_1) = \{(1,1,1,0);\ \ (1,0,0,1)\};$
$P_2(s_2) = \{(1,1,0,1);\ \ (0,1,1,0)\}.$

Since $y_{i1} + y_{i2} = 0$, denote $y_i = y_{i1} = -y_{i2}$. In this case relations (18-19) can be represented as

$$\max(y_1 + y_2 + y_3;\ y_1 + y_4) < \min(y_1 + y_2 + y_4;\ y_2 + y_3).$$

One of the solutions for those relations is as follows:

$$y_1 = -\varepsilon;\quad y_2 = \varepsilon;\quad y_3 = -\varepsilon;\quad y_4 = 0;\quad \varepsilon > 0.$$

Set $\varepsilon = 5/6$; since this value results in a new solution of the second sub-problem. We obtain:

$s_{11} = 6\frac{1}{2};\quad s_{21} = 4\frac{1}{2};\quad s_{31} = 1\frac{11}{18};\quad s_{41} = 6\frac{1}{9};$
$s_{12} = 3\frac{1}{2};\quad s_{22} = 3\frac{1}{2};\quad s_{32} = 4\frac{7}{18};\quad s_{42} = \frac{8}{9};$
$P_1(s_1) = \{(1,0,0,1);\ (1,1,1,0)\};\quad F_1 = 12\frac{11}{18};$
$P_2(s_2) = \{(1,1,0,1);\ (0,1,1,0);\ (1,0,1,0)\};\quad F_2 = 7\frac{8}{9};$
$F = 20\frac{1}{2}.$

Step 2. Consider optimality conditions

$$\max(y_1 + y_2 + y_3;\ y_1 + y_4) < \min(y_1 + y_2 + y_4;\ y_2 + y_3;\ y_1 + y_3).$$

It can be well-recognized that this inequality has no feasible solutions. Indeed, condition y1 + y2 + y3 < y1 + y2 + y4 results in $y_3 < y_4$, while condition $y_2 + y_4 < y_2 + y_3$ leads to a contradictive $y_4 < y_3$. Hence, the optimal solution of the dual problem is obtained. The determined upper estimate may be used in the branch and bounds method. Start branching with variable x_1. If $x_1 = 1$ then the solution of the corresponding dual problem results in the same estimate $F(x_1=1) = 20^1/_2$. In case $x_1 = 0$ the obtained estimate $F(x_1=0) = 14$. Choose value $x_1 = 1$ and undertake branching for variable x_2. $x_2 = 1$ results in a feasible estimate $F(x_1=1, x_2=1) = 18$. $x_2 = 0$ results in another feasible estimate $F(x_1=1, x_2=0) = 17$. Thus, the optimal solution is $x_1 = 1, x_2 = 1, x_3 = 0, x_4 = 0, C_{max} = 18$.

5. Conclusions

The suggested approach provides a generalized method to determine estimates for a broad class of nonlinear programming problems. This approach enables using new algorithms to solve a variety of problems, with the computing complexity being less, than that when using classical algorithms (the knapsack problem [3], the maximal flow problem [5], the "stones" problem [3], etc.). Further research has to be undertaken to estimate the computing complexity of the network programming method for various problems of nonlinear programming.

References

1. Burkov, V.N., Zalozhnev, A.J. and Novikov, D.A. **Graph Theory in Managing Organizational Systems,** Sinteg, Moscow, 2001 (in Russian)
2. Burkov, V.N. and Burkova, I.V. **Network programming method,** Management Problems, 3, 2005, pp. 23-29 (in Russian)
3. Burkov, V.N. and Burkova, I.V. **Method of dichotomizing programming,** Institute of Control Sciences, the Russian Academy of Sciences, 2004, pp. 57-75 (in Russian)

4. Gary, M. and Johnson, D. **Computers and Difficultly-Solved Problems,** Mir, Moscow, 1982 (in Russian)
5. Burkov, V.N. (Ed.), Mathematical backgrounds of project management, Manual, Visshaya Shkola, Moscow, 2005, pp. 312-336 (in Russian)

JOURNAL OF APPLIED QUANTITATIVE METHODS

International Symposium on Stochastic Models in Reliability Engineering, Life Sciences and Operations Management (SMRLO'10)

STRUCTURE DECISION MAKING BASED ON UNIVERSAL GENERATING FUNCTIONS FOR REFRIGERATION SYSTEM

Ilia FRENKEL
PhD, Industrial Engineering and Management Department,
SCE - Shamoon College of Engineering,
Beer-Sheva, Israel

E-mail: iliaf@sce.ac.il

Lev KHVATSKIN
PhD, Industrial Engineering and Management Department,
SCE - Shamoon College of Engineering,
Beer-Sheva, Israel

E-mail: khvat@sce.ac.il

Anatoly LISNIANSKI
PhD, Reliability Department,
The Israel Electric Corporation Ltd.,
Haifa, Israel

E-mail: anatoly-l@iec.co.il

Abstract: *This paper presents a method for calculation of reliability measures for supermarket refrigeration system. The system and its components can have different performance levels ranging from perfect functioning to complete failure and, so it can be interpreted as a multi-state system. Calculated reliability measures are used for decision making of system structure. The suggested approach based on combined Universal Generating Functions and stochastic processes method for computation of availability, output performance and performance deficiency for multi-state system. Corresponding procedures are suggested. A numerical example is presented in order to illustrate the approach.*

Key words: *reliability measures; multi-state system; Universal Generating Functions; availability; output performance; performance deficiency*

1. Introduction

Supermarkets suffer serious financial losses owing to problems with their refrigeration systems. A typical supermarket may contain more than one hundred individual refrigerated cabinets, cold store rooms and items of plant machinery which interact as part of a complex integrated refrigeration system within the store. Things very often go wrong with individual units (icing up of components, electrical or mechanical failure, and so

forth…) or with components which serve a network of units (coolant tanks, pumps, compressors, and so on).

The most commonly used refrigeration system for supermarkets today is the multiplex direct expansion system (Baxter (2002), IEA Annex 26 (2003)). All display cases and cold store rooms use direct expansion air-refrigerant coils that are connected to the system compressors in a remote machine room located in the back or on the roof of the store. Heat rejection is usually done with air-cooled condensers with simultaneously working axial blowers mounted outside. Multiple compressors are mounted on a skid, or rack, and are piped with common suction and discharge refrigeration lines. Using multiple compressors in parallel provides a means of capacity control, since the compressors can be selected and cycled as needed to meet the refrigeration load.

Due to the system's highly integrated nature, a fault in a single unit or item of machinery can't have detrimental effects on the entire store, only decrease of system cool capacity. Failure of compressor or axial condenser blower leads to partial system failure (degradation of output cooling capacity) as well as to complete failures of the system. We treat refrigeration system as multi-state system (MSS), where components and systems have an arbitrary finite number of states. According to the generic MSS model (Lisnianski and Levitin 2003), the system can have different states corresponding to the system's performance rates. The performance rate of the system at any instant is a discrete-state continuous-time stochastic process .

In this paper, a generalized approach (Lisnianski, 2004), (Lisnianski, 2007) was extended and applied for decision making for multi-state supermarket refrigeration system structure. The approach is based on the combined Universal Generating Functions UGF) and stochastic processes method for computation of availability, output performance and performance deficiency for multi-state system.

2. The Method Description

2.1. Performance Stochastic Process for Multi-state Element

In general case any element j in MSS can have k_j different states corresponding to different performance, represented by the set $\mathbf{g}_j = \left\{ g_{j1}, \ldots, g_{jk_j} \right\}$, where g_{ji} is the performance rate of element j in the state i, $i \in \left\{ 1, 2, \ldots, k_j \right\}$.

At first stage in according to the suggested method a model of stochastic process should be built for each multi-state element in MSS. Based on this model state probabilities

$$p_{ji}(t) = \Pr\{G_j(t) = g_{ji}\},\ i \in \{1, \ldots, k_j\},$$

for every MSS's element $j \in \left\{ 1, \ldots, n \right\}$ can be obtained. These probabilities define output stochastic process $G_j(t)$ for each element j in the MSS.

At the next stage the output performance distribution for the entire MSS at each time instant t should be defined based on previously determined states probabilities for all elements and system structure function. At this stage UGF technique provides simple procedure that is based only on algebraic operation.

Without loss of generality here we consider a multi-state element with minor failures and repairs.

2.2. Markov Model for Multi-state Element

If all times to failures and repair times are exponentially distributed the performance stochastic process will have Markov property and can be represented by Markov model. Here for the simplicity we omit index *j* and assume that element has *k* different states as presented in the Fig. 1. For Markov process each transition from the state *s* to any state *m*, (*s*, *m*=1,…,*k*) has its own associated transition intensity that will be designated as a_{sm}. In our case any transition is caused by element's failure or repair. If *m*<*s*, then $a_{sm} = \lambda_{sm}$, where λ_{sm} is a failure rate for the failures that cause element transition from state *s* to state *m*. If *m*>*s*, then $a_{sm} = \mu_{sm}$, where μ_{sm} is a corresponding repair rate. With each state *s* the corresponding performance g_s is associated.

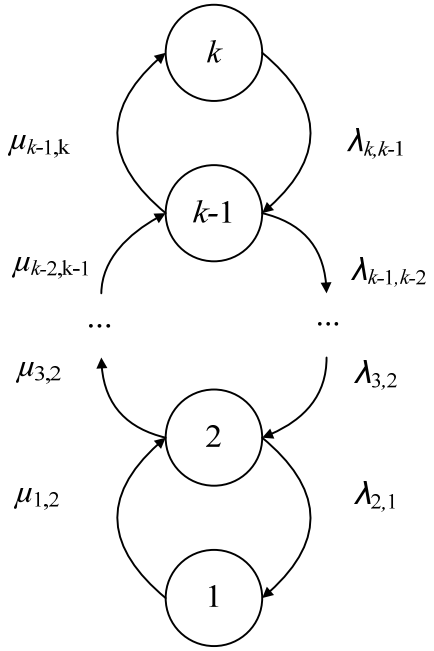


Figure 1. State-space diagram for Markov model of repairable Multi-state element

Let $p_s(t), s = 1,\ldots,k$ be the state probabilities of element's performance process $G(t)$ at time *t*: $p_s(t) = \Pr\{G(t) = g_s\},\ s = 1,\ldots,k;\ t \geq 0.$

The following system of differential equations for finding the state probabilities $p_s(t), s = 1,\ldots,k$ for the homogeneous Markov process can be written

$$\frac{dp_s(t)}{dt} = \left[\sum_{\substack{i=1 \\ i\neq s}}^{k} p_i(t)a_{is}\right] - p_s(t)\sum_{\substack{i=1 \\ i\neq s}}^{k} a_{si}\ . \qquad (1)$$

In our case all transitions are caused by element's failures and repairs. So, corresponding transition intensities a_{is} are expressed by the element's failure and repair rates. Therefore, the corresponding system of differential equations may be written

$$\begin{aligned}
&\frac{dp_1(t)}{dt} = -\mu_{12}p_1(t) + \lambda_{21}p_2(t) \\
&\frac{dp_2(t)}{dt} = \mu_{12}p_1(t) - (\lambda_{21} + \mu_{23})p_2(t) + \lambda_{32}p_3(t) \\
&\ldots \\
&\frac{dp_k(t)}{dt} = \mu_{k-1,k}p_{k-1}(t) - \lambda_{k,k-1}p_k(t)
\end{aligned} \quad . \qquad (2)$$

We assume that initial state will be the state *k* with best performance. Therefore, by solving the system (2) of differential equations under initial conditions $p_k(0)=1,\ p_{k-1}(0)=\ldots=p_2(0)=p_1(0)=0$, the states probabilities $p_s(t), s=1,\ldots,k$ can be obtained.

2.3. UGF for Multi-state System Reliability Evaluation

The generic MSS model consists of the performance stochastic processes $G_j(t)\in \mathbf{g}_j,\ j=1,\ldots,n$ for each system element *j*, and the system structure function that produces the stochastic process corresponding to the output performance of the entire MSS: $G(t)=f(G_1(t),\ldots,G_n(t))$. At the previous stage all stochastic processes $G_i(t)$, $j=1,2,\ldots,n$ were completely defined by output performance distribution at any instant *t* for each system element.

In a traditional binary-state reliability interpretation (Modarres et al 1999) a reliability block diagram shows the interdependencies among all elements. The purpose is to show, by concise visual shorthand, the various block combinations (paths) that result in system success. Each block of the reliability block diagram represents one element of function contained in the system. All blocks are configured in series, parallel, standby, or combinations thereof as appropriate. The blocks in the diagram follow a logical order which relates the sequence of events during the prescribed operation of the system. The reliability model consists of a reliability block diagram and an associated mathematical or simulation model.

In a multi-state interpretation each block of the reliability block diagram represents one multi-state element of the system. A logical order of the blocks in the diagram is defined by the system structure function $f(G_1(t),\ldots,G_n(t))$ as well as each block's *j* behavior is defined by the corresponding performance stochastic process $G_j(t)$.

At this stage based on previously determined output stochastic processes $G_j(t)$ for all elements $j=1,2,\ldots, n$, and on the given system structure function $f(G_1(t),\ldots,G_n(t))$, an output performance stochastic process $G(t)$ for the entire MSS should be defined $G(t)=f(G_1(t),\ldots,G_n(t))$. It may be done by using UGF method.

At first, individual universal generating function (UGF) for each element should be written. For each element *j* it will be UGF $u_j(z,t)$ associated with corresponding stochastic processes $G_j(t)$. Then by using composition operators over UGF of individual elements and their combinations in the entire MSS structure, one can obtain the resulting UGF $U(z,t)$ associated with output performance stochastic process $G(t)$ of the entire MSS by using simple algebraic operations. This UGF $U(z,t)$ defines the output performance distribution for the

entire MSS at each time instant *t*. MSS reliability measures can be easily derived from this output performance distribution.

The following steps should be executed:

1. Having performances g_{ji} and corresponding probabilities $p_{ji}(t)$ for each element *j*, $j=1,...n,\ i=1,...k_j$, one can define UGF $u_j(z,t)$ associated with output performance stochastic process for this element in the following form:

$$u_j(z,t) = p_{j1}(t)z^{g_{j1}} + p_{j2}(t)z^{g_{j2}} + \ldots + p_{jk_j}(t)z^{g_{jk_j}} \quad (3)$$

2. The composition operators Ω_{fser} (for elements connected in series), Ω_{fpar} (for elements connected in parallel) and $\Omega_{fbridge}$ (for elements connected in bridge structure) should be applied over the UGF of individual elements and their combinations. These operators one can find in (Lisnianski and Levitin, 2003), (Levitin, 2005), where corresponding recursive procedures for their computation were introduced for different types of systems. Based on these procedures the resulting UGF for the entire MSS can be obtained:

$$U(z,t) = \sum_{i=1}^{K} p_i(t)z^{g_i} \quad (4)$$

where *K* is the number of entire system states and g_i is the entire system performance in the corresponding state *i*, *i*=1, …,*K*.

3. Applying the operators $\delta_A, \delta_E, \delta_D$ over the resulting UGF of the entire MSS one can obtain the following MSS reliability indices:

– MSS availability *A*(*t*, *w*) at instant *t*>0 for arbitrary constant demand *w*

$$A(t,w) = \delta_A(U(z,t),w) = \delta_A(\sum_{i=1}^{K} p_i(t)z^{g_i}, w) = \sum_{i=1}^{K} p_i(t)1(g_i - w \geq 0). \quad (5)$$

– MSS expected output performance at instant t>0

$$E(t) = \delta_E(U(z,t)) = \delta_E(\sum_{i=1}^{K} p_i(t)z^{g_i}) = \sum_{i=1}^{K} p_i(t)g_i. \quad (6)$$

– MSS expected performance deficiency at t>0 for arbitrary constant demand w

$$D(t,w) = \delta_D(U(z),w) = \delta_D(\sum_{i=1}^{K} p_i(t)z^{g_i}, w) = \sum_{i=1}^{K} p_i(t)\cdot\max(w-g_i,0). \quad (7)$$

3. Numerical Example

Consider the refrigeration system used in one of the Israel supermarkets (Frenkel et al. 2010). The system consists of 2 elements: 4 compressors, situated in the machine room and 2 main axial condenser blowers. Structure scheme of the system is presented in Fig.2. It is possible to add one additional blower. In this case the structure scheme of the system is presented in Fig.4.

3.1. System with 2 Condenser Blowers

Series-parallel refrigerating multi-state system with two blowers is presented in Figure 2. State-space diagram of the elements of this system is presented in Figure 3.

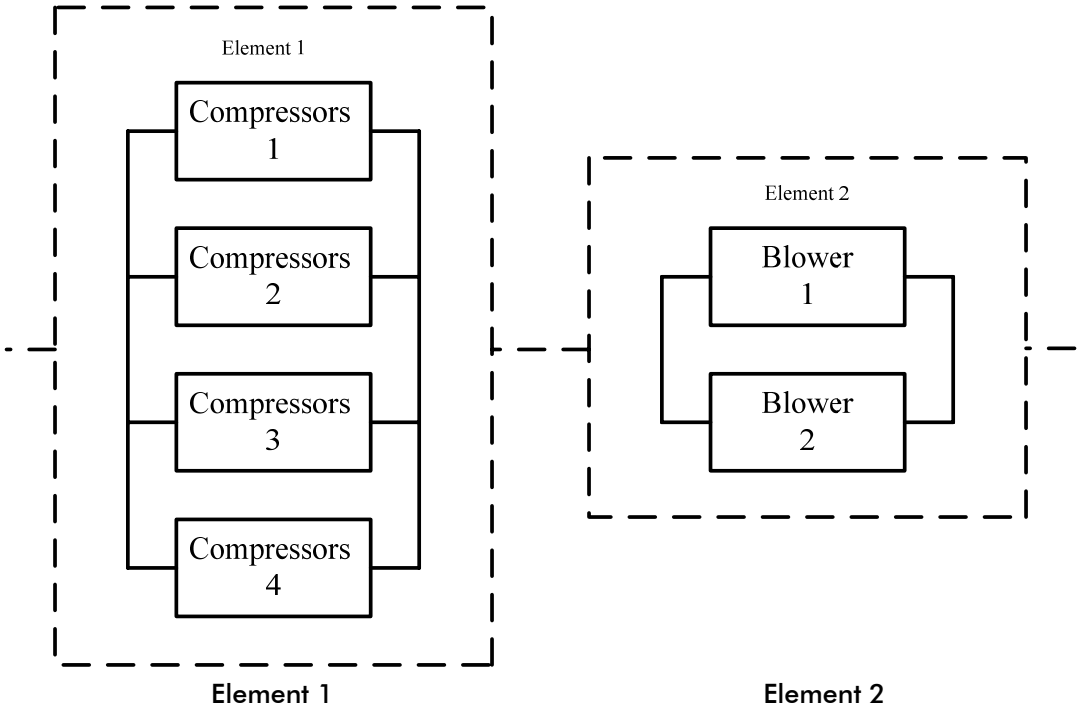


Figure 2. Series-parallel refrigerating multi-state system with two blowers

The performance of the elements is measured by their produce cold capacity (BTU per year). Times to failures and times to repairs are distributed exponentially for all elements. Elements are repairable. It is possible only minimal repair. Both elements are multi-state elements with minor failures and minor repairs. The first element can be in one of five states: a state of total failure corresponding to a capacity of 0, states of partial failures corresponding to capacities of $2.6 \cdot 10^9$, $5.2 \cdot 10^9$, $7.9 \cdot 10^9$ BTU per year and a fully operational state with a capacity of $10.5 \cdot 10^9$ BTU per year. For simplification we will present system capacity in 10^9BTU per year units. Therefore,

$$G_1(t) \in \{g_{11}, g_{12}, g_{13}, g_{14}, g_{15}\} = \{0, 2.6, 5.2, 7.9, 10.5\}. \quad (8)$$

The failure rates and repair rates corresponding to the first element are $\lambda^C = 1 year^{-1}$, $\mu^C = 12 year^{-1}$.

The second element can be in one of three states: a state of total failure corresponding to a capacity of 0, state of partial failure corresponding to capacity of $5.2 \cdot 10^9$ BTU per year and a fully operational state with a capacity of $10.5 \cdot 10^9$ BTU per year. Therefore,

$$G_2(t) \in \{g_{21}, g_{22}, g_{23}\} = \{0, 5.2, 10.5\}. \quad (9)$$

The failure rate and repair rate corresponding to the second element are $\lambda^B = 10 year^{-1}$, $\mu^B = 365 year^{-1}$.

The MSS structure function is:

$$G_s(t) = f\left(G_1(t), G_2(t)\right) = \min\left\{G_1(t), G_2(t)\right\}. \quad (10)$$

The demand is constant: $w=5.0\cdot10^9$ BTU per year.

Using combined UGF and stochastic process method we will find MSS availability $A(t,w)$, expected output performance $E(t)$ and expected performance deficiency $D(t,w)$.

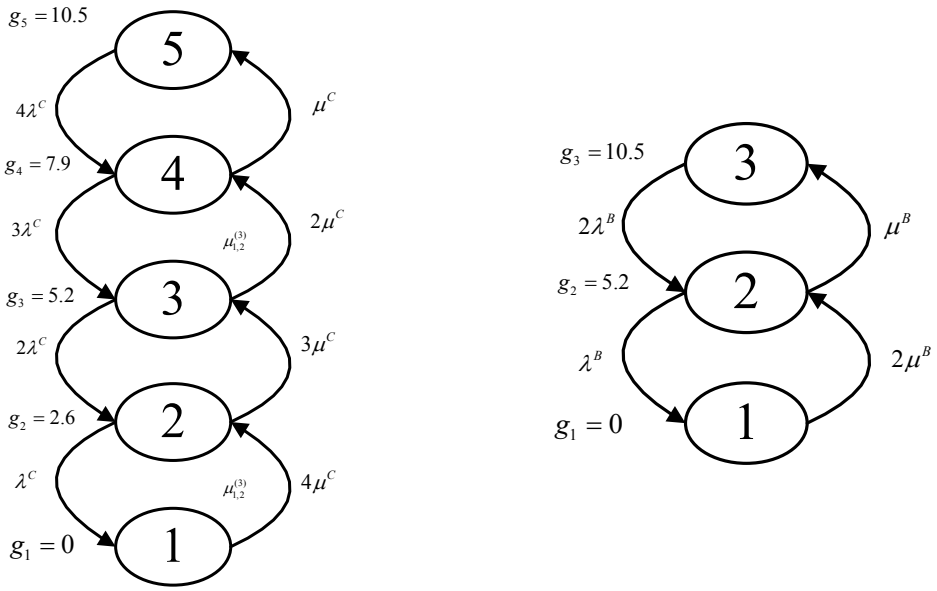


Figure 3. State-space diagram of the multi-state system with two blowers

Applying the described above two-stage procedure, we proceed as follows.

According to the Markov method we build the following systems of differential equations for each element separately (using the state-space diagrams presented in Figure 3).

- For element 1:

$$\begin{cases} \dfrac{dp_{11}(t)}{dt} = -4\mu^C p_{11}(t) + \lambda^C p_{12}(t) \\ \dfrac{dp_{12}(t)}{dt} = 4\mu^C p_{11}(t) - (\lambda^C + 3\mu^C) p_{12}(t) + 2\lambda^C p_{13}(t) \\ \dfrac{dp_{13}(t)}{dt} = 3\mu^C p_{12}(t) - (2\lambda^C + 2\mu^C) p_{13}(t) + 3\lambda^C p_{14}(t) \\ \dfrac{dp_{14}(t)}{dt} = 2\mu^C p_{13}(t) - (3\lambda^C + \mu^C) p_{14}(t) + 4\lambda^C p_{15}(t) \\ \dfrac{dp_{15}(t)}{dt} = \mu^C p_{14}(t) - 4\lambda^C p_{15}(t). \end{cases} \quad (11)$$

Initial conditions are: $p_{11}(0) = p_{12}(0) = p_{13}(0) = p_{14}(0) = 0;\ p_{15}(0) = 1..$

- For element 2:

$$\begin{cases} \dfrac{dp_{21}(t)}{dt} = -2\mu^B p_{21}(t) + \lambda^B p_{22}(t) \\ \dfrac{dp_{22}(t)}{dt} = 2\mu^B p_{21}(t) - (\lambda^B + \mu^B) p_{22}(t) + 2\lambda^B p_{23}(t) \\ \dfrac{dp_{23}(t)}{dt} = \mu^B p_{22}(t) - 2\lambda^B p_{23}(t). \end{cases} \quad (12)$$

Initial conditions are: $p_{21}(0) = p_{22}(0) = 0;\ p_{23}(0) = 1.$

A closed form solution can be obtained for each of these 2 systems of differential equations. All calculations were made using MATLAB®. Corresponding expressions for states probabilities are the following.

- For element 1:

$$\begin{aligned} p_{11}(t) &= 1/28561 - 4/28561 e^{-13t} + 6/28561 e^{-26t} - 4/28561 e^{-39t} + 1/28561 e^{-52t}, \\ p_{12}(t) &= 48/28561 - 140/28561 e^{-13t} + 132/28561 e^{-26t} - 36/28561 e^{-39t} + 4/28561 e^{-52t}, \\ p_{13}(t) &= 864/28561 - 1584/28561 e^{-13t} + 582/28561 e^{-26t} - 132/28561 e^{-39t} + 6/28561 e^{-52t}, \\ p_{14}(t) &= 6912/28561 - 5184/28561 e^{-13t} + 1584/28561 e^{-26t} - 140/28561 e^{-39t} + 4/28561 e^{-52t}, \\ p_{15}(t) &= 20736/28561 - 6912/28561 e^{-13t} + 864/28561 e^{-26t} - 48/28561 e^{-39t} + 1/28561 e^{-52t}. \end{aligned} \quad (13)$$

- For element 2:

$$\begin{aligned} p_{21}(t) &= 4/5625 + 4/5625 e^{-750t} - 8/5625 e^{-375t}, \\ p_{22}(t) &= 292/5625 - 8/5625 e^{-750t} - 284/5625 e^{-375t}, \\ p_{23}(t) &= 5329/5625 + 4/5625 e^{-750t} + 292/5625 e^{-375t}. \end{aligned} \quad (14)$$

Therefore, one obtains the following output performance stochastic processes:

– element 1: $\begin{cases} \mathbf{g}_1 = \{g_{11}, g_{12}, g_{13}, g_{14}, g_{15}\} = \{0, 2.6, 5.2, 7.9, 10.5\}, \\ \mathbf{p}_1(t) = \{p_{11}(t), p_{12}(t), p_{13}(t), p_{14}(t), p_{15}(t)\}; \end{cases}$

– element 2: $\begin{cases} \mathbf{g}_2 = \{g_{21}, g_{22}, g_{23}\} = \{0, 5.2, 10.5\}, \\ \mathbf{p}_2(t) = \{p_{21}(t), p_{22}(t), p_{23}(t)\}. \end{cases}$

Having the sets $\mathbf{g}_i$, $\mathbf{p}_i(t)$ for i=1,2 one can define for each individual element i the u-function associated with the element's output performance stochastic process:

$$\begin{aligned} u_1(z,t) &= p_{11}(t) z^{g_{11}} + p_{12}(t) z^{g_{12}} + p_{13}(t) z^{g_{13}} + p_{14}(t) z^{g_{14}} + p_{15}(t) z^{g_{15}} = \\ &\quad p_{11}(t) z^0 + p_{12}(t) z^{2.6} + p_{13}(t) z^{5.2} + p_{14}(t) z^{7.9} + p_{15}(t) z^{10.5}, \\ u_2(z,t) &= p_{21}(t) z^{g_{21}} + p_{22}(t) z^{g_{22}} + p_{23}(t) z^{g_{23}} = \\ &\quad p_{21}(t) z^0 + p_{22}(t) z^{5.2} + p_{23}(t) z^{10.5}. \end{aligned} \quad (15)$$

Using the composition operator $\Omega_{f_{ser}}$ for refrigerating MSS one obtains the resulting UGF for the entire series MSS

$$U(z,t)=\Omega_{f_{ser}}(u_1(z,t),u_2(z,t)). \tag{16}$$

In order to find the resulting UGF $U(z,t)$ for elements 1 and 2 connected in series the operator $\Omega_{f_{ser}}$ applied to individual UGF $u_1(z,t)$ and $u_2(z,t)$.

$$\begin{aligned}
U(z,t) &= \Omega_{f_{ser}}\left(u_1(z,t),u_2(z,t)\right)= \\
&= \Omega_{f_{ser}}\left(p_{11}(t)z^0+p_{12}(t)z^{2.6}+p_{13}(t)z^{5.2}+p_{14}(t)z^{7.9}+p_{15}(t)z^{10.5},\right. \\
&\quad \left. p_{21}(t)z^0+p_{22}(t)z^{5.2}+p_{23}(t)z^{10.5}\right)= \\
&= p_{11}(t)p_{21}(t)z^0+p_{11}(t)p_{22}(t)z^0+p_{11}(t)p_{23}(t)z^0+ \\
&+ p_{12}(t)p_{21}(t)z^0+p_{12}(t)p_{22}(t)z^{2.6}+p_{12}(t)p_{23}(t)z^{2.6}+ \\
&+ p_{13}(t)p_{21}(t)z^0+p_{13}(t)p_{22}(t)z^{5.2}+p_{13}(t)p_{23}(t)z^{5.2}+ \\
&+ p_{14}(t)p_{21}(t)z^0+p_{14}(t)p_{22}(t)z^{5.2}+p_{14}(t)p_{23}(t)z^{7.9}. \\
&+ p_{15}(t)p_{21}(t)z^0+p_{15}(t)p_{22}(t)z^{5.2}+p_{15}(t)p_{23}(t)z^{10.5}.
\end{aligned} \tag{17}$$

In the resulting UGF $U(z,t)$ the powers of z are found as minimum of powers of corresponding terms.

Taking into account that $p_{11}(t)+p_{12}(t)+p_{13}(t)+p_{14}(t)+p_{15}(t)=1$ and $p_{21}(t)+p_{22}(t)+p_{23}(t)=1$, one can simplify the last expression for $U(z,t)$ and obtain the resulting UGF associated with the output performance stochastic process $\mathbf{g}$, $\mathbf{p}(t)$ of the entire MSS in the following form

$$U(z,t)=\sum_{i=1}^{5}p_i(t)z^{g_i} \tag{18}$$

where

$$\begin{aligned}
&g_1=0, && p_1(t)=p_{11}(t)+\left(1-p_{11}(t)\right)p_{21}(t),\\
&g_2=2.6\cdot10^9 \text{ BTU/year}, && p_2(t)=p_{12}(t)\left[p_{22}(t)+p_{23}(t)\right],\\
&g_3=5.2\cdot10^9 \text{ BTU/year}, && p_3(t)=\left[p_{13}(t)+p_{14}(t)+p_{15}(t)\right]p_{22}(t)+p_{13}(t)p_{23}(t),\\
&g_4=7.9\cdot10^9 \text{ BTU/year}, && p_4(t)=p_{14}(t)p_{23}(t),\\
&g_5=10.5\cdot10^9 \text{ BTU/year}, && p_5(t)=p_{15}(t)p_{23}(t).
\end{aligned}$$

These two sets

$$\mathbf{g}=\{g_1,g_2,g_3,g_4,g_5\} \text{ and } \mathbf{p}(t)=\{p_1(t),p_2(t),p_3(t),p_4(t),p_5(t)\}$$

completely define output performance stochastic process for the entire MSS.

Based on resulting UGF $U(z,t)$ of the entire MSS, one can obtain the MSS reliability indices. The instantaneous MSS availability for the constant demand level $w=5.0\cdot10^9$ BTU per year

$$A(t) = \delta_A(U(z,t),w) = \delta_A(\sum_{i=1}^{5} p_i(t)z^{g_i},5) = \sum_{i=1}^{5} p_i(t)1(F(g_i,5)\geq 0) = p_3(t)+p_4(t)+p_5(t). \quad (19)$$

The instantaneous mean output performance at any instant $t>0$

$$E(t) = \delta_E(U(z,t)) = \sum_{i=1}^{5} p_i(t)g_i = 2.6p_2(t)+5.2p_3(t)+7.9p_4(t)+10.5p_5(t). \quad (20)$$

The instantaneous performance deficiency $D(t)$ at any time t for the constant demand $w=5.0\cdot10^9$ BTU per year:

$$D(t) = \delta_D(U(z),w) = \sum_{i=1}^{5} p_i(t)\cdot\max(5-g_i,0) = = p_1(t)(5-0)+p_2(t)(5-2.6) = 5p_1(t)+2.4p_2(t). \quad (21)$$

Calculated reliability indices $A(t)$, $E(t)$ and $D(t)$ are presented on the Figures 6-8.

Note that instead of solving the system of $K=5*3=15$ differential equations (as it should be done in the straightforward Markov method) here we solve just two systems. The further derivation of the entire system states probabilities and reliability indices is based on using simple algebraic equations.

3.2. System with 3 Condenser Blowers

To increase reliability level of the system Supermarket decided to add additional blower and our goal is to compare reliability indices in new structure. The new refrigerating system structure is presented in Figure 4. State-space diagram of the elements of this system is presented in Figure 5.

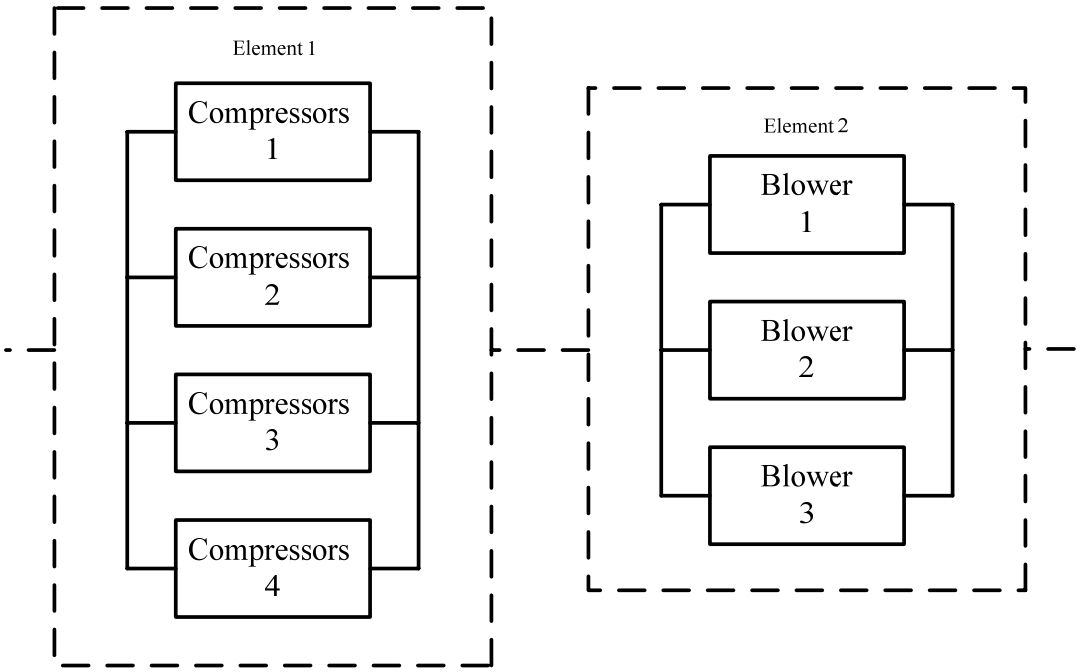


Figure 4. Series-parallel refrigerating multi-state system with 3 blowers

Like in previous case the system consists of two elements: block of 4 compressors and block of 3 blower. The performance of the elements is measured by their produce cold capacity (BTU per year). Times to failures and times to repairs are distributed exponentially for all elements. Elements are repairable. Both elements are multi-state elements with minor failures and minor repairs. The first element can be in one of five states: a state of total failure corresponding to a capacity of 0, states of partial failures corresponding to capacities of $2.6 \cdot 10^9$, $5.2 \cdot 10^9$, $7.9 \cdot 10^9$ BTU per year and a fully operational state with a capacity of $10.5 \cdot 10^9$ BTU per year. For simplification we will present system capacity in 10^9BTU per year units. Therefore,

$$G_1(t) \in \{g_{11}, g_{12}, g_{13}, g_{14}, g_{15}\} = \{0, 2.6, 5.2, 7.9, 10.5\}. \quad (22)$$

The failure rates and repair rates corresponding to the first element are $\lambda^C = 1 year^{-1}$, $\mu^C = 12 year^{-1}$.

The second element can be in one of 4 states: a state of total failure corresponding to a capacity of 0, state of partial failure corresponding to capacity of $5.2 \cdot 10^9$ BTU per year and two fully operational states with a capacity of $10.5 \cdot 10^9$ BTU per year. Therefore,

$$G_2^*(t) \in \{g_{21}^*, g_{22}^*, g_{23}^*, g_{24}^*\} = \{0, 5.2, 10.5, 10.5\}. \quad (23)$$

The failure rate and repair rate corresponding to the second element are $\lambda^B = 10 year^{-1}$, $\mu^B = 365 year^{-1}$.

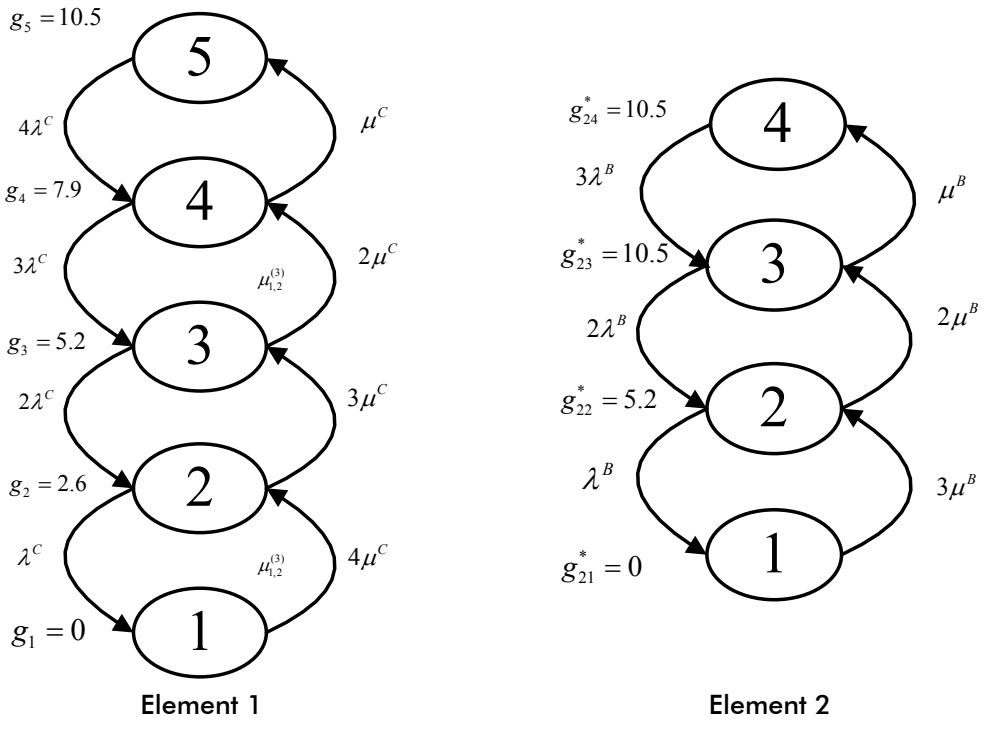


Figure 5. State-space diagram of the multi-state system with 3 blowers

The MSS structure function is:

$$G_s(t) = f\left(G_1(t), G_2^*(t)\right) = \min\left\{G_1(t), G_2^*(t)\right\}. \quad (24)$$

The demand is constant: $w = 5.0 \cdot 10^9$ BTU per year.

Using combined UGF and stochastic process method we will find MSS availability $A(t,w)$, expected output performance $E(t)$ and expected performance deficiency $D(t,w)$ for the system with additional blower.

Applying the described above two-stage procedure, we proceed as follows.

1. According to the Markov method we build the following systems of differential equations for each element separately (using the state-space diagrams presented in Figure 5).

   For element 1 all calculations were proceeded earlier (13).

   For element 2:

$$\begin{cases} \dfrac{dp_{21}^*(t)}{dt} = -3\mu^B p_{21}^*(t) + \lambda^B p_{22}^*(t) \\ \dfrac{dp_{22}^*(t)}{dt} = 3\mu^B p_{21}^*(t) - (\lambda^B + 2\mu^B) p_{22}^*(t) + 2\lambda^B p_{23}^*(t) \\ \dfrac{dp_{23}^*(t)}{dt} = 2\mu^B p_{22}^*(t) - (2\lambda^B + \mu^B) p_{23}^*(t) + 3\lambda^B p_{24}^*(t) \\ \dfrac{dp_{24}^*(t)}{dt} = \mu^B p_{23}^*(t) - 3\lambda^B p_{24}^*(t). \end{cases} \quad (25)$$

   Initial conditions are: $p_{21}^*(0) = p_{22}^*(0) = p_{23}^*(0) = 0;\ p_{24}^*(0) = 1.$

A closed form solution can be obtained for the system of differential equations. Corresponding expressions for states probabilities are the following.

For element 2:

$$\begin{aligned} p_{21}^*(t) &= 8/421875 + 8/140625 e^{-750t} - 8/140625 e^{-375t} - 8/421875 e^{-1125t}, \\ p_{22}^*(t) &= 292/140625 + 92/46875 e^{-750t} - 64/15625 e^{-375t} - 8/140625 e^{-1125t}, \\ p_{23}^*(t) &= 10658/140625 - 64/15625 e^{-750t} - 3358/46875 e^{-375t} - 8/140625 e^{-1125t}, \\ p_{24}^*(t) &= 389017/421875 + 292/140625 e^{-750t} + 10658/140625 e^{-375t} + 8/421875 e^{-1125t}. \end{aligned} \quad (26)$$

Therefore, one obtains the following output performance stochastic processes:

- element 1: $\begin{cases} \mathbf{g}_1 = \{g_{11}, g_{12}, g_{13}, g_{14}, g_{15}\} = \{0, 2.6, 5.2, 7.9, 10.5\}, \\ \mathbf{p}_1(t) = \{p_{11}(t), p_{12}(t), p_{13}(t), p_{14}(t), p_{15}(t)\}; \end{cases}$
- element 2: $\begin{cases} \mathbf{g}_2^* = \{g_{21}^*, g_{22}^*, g_{23}^*, g_{24}^*\} = \{0, 5.2, 10.5, 10.5\}, \\ \mathbf{p}_2^*(t) = \{p_{21}^*(t), p_{22}^*(t), p_{23}^*(t), p_{24}^*(t)\}. \end{cases}$

1. Having the sets $\mathbf{g}_j$, $\mathbf{p}_j(t)$ for j=1,2 one can define for each individual element j the u-function associated with the element's output performance stochastic process:

$$u_1(z,t) = p_{11}(t)z^{g_{11}} + p_{12}(t)z^{g_{12}} + p_{13}(t)z^{g_{13}} + p_{14}(t)z^{g_{14}} + p_{15}(t)z^{g_{15}} = p_{11}(t)z^{0} + p_{12}(t)z^{2.6} + p_{13}(t)z^{5.2} + p_{14}(t)z^{7.9} + p_{15}(t)z^{10.5},$$
$$u_2^*(z,t) = p_{21}^*(t)z^{g_{21}^*} + p_{22}^*(t)z^{g_{22}^*} + p_{23}^*(t)z^{g_{23}^*} + p_{24}^*(t)z^{g_{24}^*} = p_{21}^*(t)z^{0} + p_{22}^*(t)z^{5.2} + p_{23}^*(t)z^{10.5} + p_{24}^*(t)z^{10.5}. \quad (27)$$

2. Using the composition operator $\Omega_{f_{ser}}$ for refrigerating MSS one obtains the resulting UGF for the entire series MSS

$$U(z,t) = \Omega_{f_{ser}}(u_1(z,t), u_2^*(z,t)). \quad (28)$$

In order to find the resulting UGF $U(z,t)$ for elements 1 and 2 connected in series the operator $\Omega_{f_{ser}}$ applied to individual UGF $u_1(z,t)$ and $u_2(z,t)$.

$$\begin{aligned} U(z,t) &= \Omega_{f_{ser}}\left(u_1(z,t), u_2(z,t)\right) = \\ &= \Omega_{f_{ser}}\left(p_{11}(t)z^{0} + p_{12}(t)z^{2.6} + p_{13}(t)z^{5.2} + p_{14}(t)z^{7.9} + p_{15}(t)z^{10.5},\right. \\ &\left. p_{21}^*(t)z^{0} + p_{22}^*(t)z^{5.2} + \left(p_{23}^*(t) + p_{24}^*(t)\right)z^{10.5}\right) = \\ &= p_{11}(t)p_{21}^*(t)z^{0} + p_{11}(t)p_{22}^*(t)z^{0} + p_{11}(t)\left(p_{23}^*(t) + p_{24}^*(t)\right)z^{0} + \\ &+ p_{12}(t)p_{21}^*(\mathrm{t})z^{0} + p_{12}(t)p_{22}^*(t)z^{2.6} + p_{12}(t)\left(p_{23}^*(t) + p_{24}^*(t)\right)z^{2.6} + \\ &+ p_{13}(t)p_{21}^*(t)z^{0} + p_{13}(t)p_{22}^*(t)z^{5.2} + p_{13}(t)\left(p_{23}^*(t) + p_{24}^*(t)\right)z^{5.2} + \\ &+ p_{14}(t)p_{21}^*(t)z^{0} + p_{14}(t)p_{22}^*(t)z^{5.2} + p_{14}(t)\left(p_{23}^*(t) + p_{24}^*(t)\right)z^{7.9}. \\ &+ p_{15}(t)p_{21}^*(t)z^{0} + p_{15}(t)p_{22}^*(t)z^{5.2} + p_{15}(t)\left(p_{23}^*(t) + p_{24}^*(t)\right)z^{10.5}. \end{aligned} \quad (29)$$

In the resulting UGF $U(z,t)$ the powers of z are found as minimum of powers of corresponding terms.

Taking into account that $p_{11}(t) + p_{12}(t) + p_{13}(t) + p_{14}(t) + p_{15}(t) = 1$ and $p_{21}^*(t) + p_{22}^*(t) + p_{23}^*(t) + p_{24}^*(t) = 1$, one can simplify the last expression for $U(z,t)$ and obtain the resulting UGF associated with the output performance stochastic process $\mathbf{g}$, $\mathbf{p}(t)$ of the entire MSS in the following form

$$U(z,t) = \sum_{i=1}^{5} p_i(t)z^{g_i} \quad (30)$$

where

$g_1 = 0$, $\quad p_1(t) = p_{11}(t) + \left[1 - p_{11}(t)\right]p_{21}^*(t)$,

$g_2 = 2.6 \cdot 10^9$ BTU/year, $\quad p_2(t) = p_{12}(t)\left[1 - p_{21}^*(t)\right]$,

$g_3 = 5.2 \cdot 10^9$ BTU/year, $\quad p_3(t) = p_{13}(t)\left[1 - p_{21}^*(t)\right] + \left[p_{14}(t) + p_{15}(t)\right]p_{22}^*(t)$,

$g_4 = 7.9 \cdot 10^9$ BTU/year, $\quad p_4(t) = p_{14}(t)\left[p_{23}^*(t) + p_{24}^*(t)\right]$,

$g_5 = 10.5 \cdot 10^9$ BTU/year, $\quad p_5(t) = p_{15}(t)\left[p_{23}^*(t) + p_{24}^*(t)\right]$.

These two sets

$\mathbf{g} = \{g_1, g_2, g_3, g_4, g_5\}$ and $\mathbf{p}(t) = \{p_1(t), p_2(t), p_3(t), p_4(t), p_5(t)\}$

completely define output performance stochastic process for the entire MSS.

Based on resulting UGF $U(z,t)$ of the entire MSS, one can obtain the MSS reliability indices. The instantaneous MSS availability for the constant demand level $w=5.0\cdot10^9$ BTU per year

$$A(t) = \delta_A(U(z,t), w) = \delta_A(\sum_{i=1}^{5} p_i(t) z^{g_i}, 5) = \sum_{i=1}^{5} p_i(t) 1(F(g_i, 5) \geq 0) = p_3(t) + p_4(t) + p_5(t). \quad (31)$$

The instantaneous mean output performance at any instant $t>0$

$$E(t) = \delta_E(U(z,t)) = \sum_{i=1}^{5} p_i(t) g_i = 2.6 p_2(t) + 5.2 p_3(t) + 7.9 p_4(t) + 10.5 p_5(t). \quad (32)$$

The instantaneous performance deficiency $D(t)$ at any time t for the constant demand $w=5.0\cdot10^9$ BTU per year:

$$D(t) = \delta_D(U(z), w) = \sum_{i=1}^{5} p_i(t) \cdot \max(5 - g_i, 0) = = p_1(t)(5-0) + p_2(t)(5-2.6) = 5p_1(t) + 2.4 p_2(t). \quad (33)$$

Calculated reliability indices $A(t)$, $E(t)$ and $D(t)$ are presented on the Figures 6-8.

Note that instead of solving the system of $K=5*4=20$ differential equations (as it should be done in the straightforward Markov method) here we solve just two systems. The further derivation of the entire system states probabilities and reliability indices is based on using simple algebraic equations.

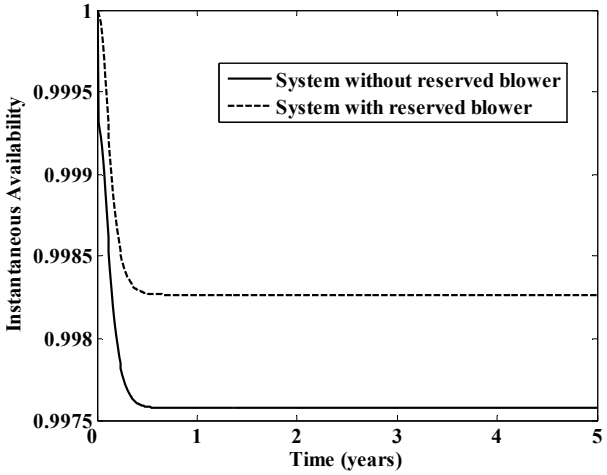


Figure 6. MSS instantaneous availability for different types of systems

Curves in Figures 6-8 support the engineering decision-making and determine the areas where required performance deficiency level of the refrigeration system can be provided by configuration "with additional blower" or by configuration "without additional blower". For example, from the Figure 6 one can conclude that the configuration "without

additional blower" cannot provide the required average availability, if it is greater than 0.998.

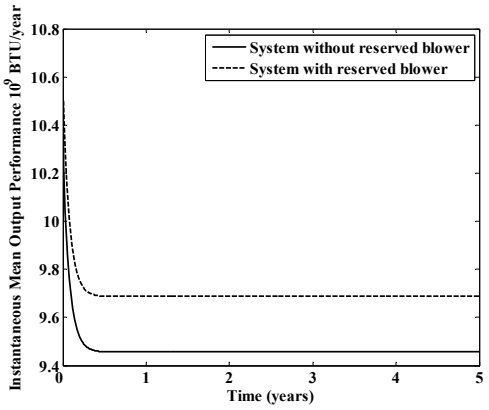


Figure 7. MSS instantaneous mean output performance for different types of systems

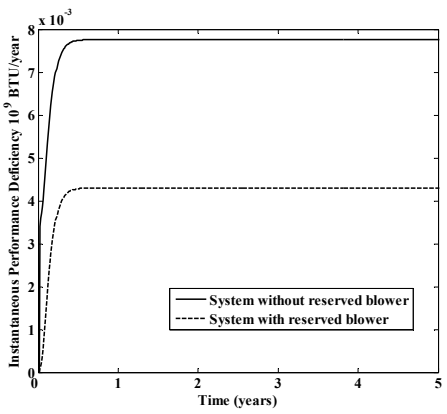


Figure 8. MSS instantaneous mean performance deficiency for different types of systems

4. Conclusions

The universal method was applied to compute MSS reliability measures: system availability, output performance and performance deficiency. The method is based on the combined Universal Generating Functions and stochastic processes method.

The case-study demonstrates that the approach is well formalized and suitable for practical application in reliability engineering. It supports the engineering decision-making and determines different system structures providing a required reliability/availability level of MSS.

References

1. Baxter, V. **Advances in supermarket refrigeration systems**, Proceedings of the 7th International Energy Agency Heat Pump Conference, Beijing, China, May 19-22, 2002
2. Frenkel, I., Khvatskin, L. and Lisnianski, A. **Markov reward model for performance deficiency calculation of refrigeration system**, in Reliability, Risk and Safety: Theory and Applications, ed. Bris, R., Soares, C.G. and Martorell, S., CRC Press, Taylor & Francis Group: London, 2009, pp. 1591-1596
3. IEA Annex 26 **Advanced Supermarket Refrigeration/Heat Recovery Systems**, Final Report Volume 1, Oak Ridge National Laboratory, Oak Ridge, TN, USA, 2003
4. Kuo, W. and Zuo, M. **Optimal Reliability Modeling Principles and Applications**, New Jersey: John Wiley & Sons, 2003
5. Lisnianski, A. **Universal generating function technique and random process methods for multi-state system reliability analysis**, in Proceedings of the 2nd International Workshop in Applied Probability (IWAP2004), Piraeus: Greece, 2004, 237-242
6. Lisnianski, A. **Extended block diagram method for a multi-state system reliability assessment**, Reliability Engineering and Systems Safety, 92(12), 2007, pp. 1601-1607
7. Lisnianski, A. and Levitin, G. **Multi-state System Reliability: Assessment, Optimization, Applications**, NY, London, Singapore: World Scientific, 2003

8. Lisnianski, A., Frenkel, I. and Ding, Y. **Multi-state System Reliability Analysis and Optimization for Engineers and Industrial Managers**, London, Springer, 2010
9. Lisnianski, A. *et al*. **Markov reward model for multi-state system reliability assessment**, in Statistical Models and Methods for Biomedical and Technical Systems, eds. Vonta, F., Nikulin, M., Limnios, N. and Huber-Carol, C., Birkhaüser: Boston, 2007, pp. 153-168
10. Modarres, M., Kaminskiy, M. and Krivtsov, V. **Reliability Engineering and Risk Analysis: A Practical Guide**, NY, Basel: Marcel Dekker, Inc., 1999

JOURNAL OF APPLIED QUANTITATIVE METHODS

International Symposium on Stochastic Models in Reliability Engineering, Life Sciences and Operations Management (SMRLO'10)

PRODUCTION PLANNING UNDER UNCERTAIN DEMANDS AND YIELDS

Zohar LASLO
Prof., Dean, Industrial Engineering and Management Department,
SCE - Shamoon College of Engineering,
Beer-Sheva, Israel

E-mail: zohar@sce.ac.il

Gregory GUREVICH
PhD, Industrial Engineering and Management Department,
SCE - Shamoon College of Engineering,
Beer-Sheva, Israel

E-mail: gregoryg@sce.ac.il

Baruch KEREN
PhD, Industrial Engineering and Management Department,
SCE - Shamoon College of Engineering,
Beer-Sheva, Israel

E-mail: baruchke@sce.ac.il

Abstract: *The periodic demands of a single product are forecasted and given by a distribution function for each period. The product can be manufactured in n plants with heterogeneous characters. Each plant has its specific stochastic production capability. The expected capability and the standard deviation of each plant can be increased by allocation of additional budgets. The problem is to determine the total budget needed and its distribution among the n plants in order to ensure a complete fulfillment of the demands according to the due dates and the pre-given confidence levels.*

Key words: *production planning; chance constrained; capability-cost trade-offs; random yield*

1. Introduction

The planning process of global production for a new product with numerous quantities addressed to anonymous customers (e.g., semiconductors, pharmaceutics, etc.) forces the corporation management to take the following principal decisions: 1) how much to produce, 2) where to produce, and 3) how to divide the production among a number of optional producers.

Mostly, actual demand fluctuates around the mean of demand distribution. Assuming that the mean of the underlying demand pattern is known, this fluctuation constitutes demand uncertainty. However, the expected demand can also vary through time, such as when seasonality is present. In such cases the true mean of the demand distribution

is not stationary through time. Demand variability over time includes both demand uncertainty and variation due to the shifting mean of the demand distribution (Enns, 2002).

The presence of random yields can considerably complicate production planning and control. When the manufacturers control their inputs but the outputs exhibit random yields, coordination in such systems becomes quite complex. Two variants of demand have been addressed in the literature: 1) rigid demand - where an order must be satisfied in its entirety (possibly necessitating multiple manufacturing runs), and 2) non-rigid demand - where there is a penalty for a shortage (only one manufacturing run). The determination of monthly productions is particularly challenging when yields are random and demand needs to be satisfied in its entirety (i.e., rigid demand). The efficient planning of monthly productions often becomes a crucial economic factor. As a result the modeling of production with random yields has attracted the attention of many researchers (for a literature review see Yano and Lee 1995).

Random yield disables satisfaction of demand in its entirety, but determining strict chance constraint enables us to attain close claim to rigid demand. Laslo (2003) clarified that when additional budget is invested in order to obtain a rigid performance, we should refer to the impact of this act on the performance fractile and not on its impact on the expected performance. Such an approach puts the delivery objectives before the objective of reducing superfluous production.

Laslo and Gurevich (2007) have developed an iterative procedure for the minimization of budget that is required for executing the activities chain with chance constrained lead-time. The procedure assumes fixed coefficient variance while budget is added in order to increase the execution speed. This iterative procedure is applicable as well for the minimization of the total budget that should be allocated among heterogeneous plants (differing by initial investment, productivity and yield variance) which are supposed to supply together a rigid known demand under strict chance constraint. Laslo et al. (2009) have introduced another procedure that resolves problems where the optimization is carried out for several known rigid demands with a common due-date but under different chance constraints. They assumed for each producer a standard deviation of the yield that increases proportionally with the production and linearly with allocated budget.

This paper introduces a solution for a comprehensive problem of operating manufacturing with heterogeneous plants that differ by their investment-capacity tradeoff curves and their yield distributions. We consider: 1) monthly rigid demands (i.e., several orders with different due days given as a time series), 2) uncertain nonnegative demands with different expected amount and different variance of demand, and 3) random yield. The objective is to establish a global production plan that minimizes the total investment in the production plants, subject to monthly rigid deliveries and under pre-given chance constraints.

2. Notation

Let us introduce the following terms:

- $\{j\}$ an index for the months, $j = 1,2,...,k$;
- O^j the j's monthly demand, $j = 1,2,...,k$ (a random variable with known distribution);

$(1-\alpha_j)$ - the lower bound probability (confidence level) for complete fulfillment of the monthly demand O_j, $0<\alpha_j<0.5$, $j=1,2,...,k$;

t_j - the time for supplying of the monthly demand O_j, $t_1 \le t_2 \le ... \le t_k$;

$\{i\}$ - an index for the plants, $i=1,2,...,n$;

p_i - the normal production quantity of plant i up to the lead time t_k (a random variable with known expectation $E(p_i)$), given that the normal budget $c_{E(p_i)}$ was allocated for plant i;

$c_{E(p_i)}$ - the known deterministic budget that enables a normal production quantity p_i, at plant i for the planning horizon $[t_1,t_k]$;

$\sigma(p_i)$ - the known standard deviation of p_i;

P_i - the crash production quantity of plant i up to the lead time t_k (a random variable with known expectation $E(P_i)$), given that the crash budget $c_{E(P_i)}$ was allocated for plant i;

$c_{E(P_i)}$ - the known deterministic budget that enables the crash production quantity P_i (capital P), at plant i for the planning horizon $[t_1,t_k]$;

q_i^k - the production quantity of plant i for the planning horizon $[t_1,t_k]$ (a random variable with expected value $E(q_i^k)$, $E(p_i) \le E(q_i^k) \le E(P_i)$, that is dependent on the deterministic budget c_i allocated to the plant i);

c_i - the budget (a decision variable) that enables q_i^k production quantity of plant i for the planning horizon $[t_1,t_k]$, $c_{E(p_i)} \le c_i \le c_{E(P_i)}$;

$\underline{c}=(c_1,...,$ - a vector of the distributed budget among all plants.

C - the total budget allocated to all plant: $C=\sum_{i=1}^{n} c_i$;

Q^k - the total production for the planning horizon $[t_1,t_k]$ (a random variable), $Q^k=\sum_{i=1}^{n} q_i^k$;

Q_α^k - the α quintile of Q^k's distribution, $0<\alpha<0.5$;

Q^j - the total production at the horizon $[t_1,t_k]$; $Q^j=\sum_{i=1}^{n} q_i^j$, $j=1,2,...,k$.

3. Problem Definition

We consider n plants (production units) that can produce the same product. Each plant i, $i=1,2,...,n$ has a stochastic production capability q_i^k and needs a deterministic budget c_i, $c_{E(p_i)} \le c_i \le c_{E(P_i)}$, in order to activate the production capability q_i^k for the planning horizon $[t_1,t_k]$.

We assume that the relation for the expected production capability, given that c_i budget was allocated for plant i, $E(q_i^k|c_i)$, is given by a continuous linear increasing curve:

$$E(q_i^k|c_i) = \varphi_i c_i + \gamma_i, \quad (1)$$

where $\varphi_i = \frac{E(P_i) - E(p_i)}{c_{E(P_i)} - c_{E(p_i)}}$, $\gamma_i = \frac{c_{P_i} E(p_i) - c_{p_i} E(P_i)}{c_{E(P_i)} - c_{E(p_i)}}$.

We also assume that the randomness in the production capability of plant i is realized only once, immediately after the investment of c_i, $i = 1,2,...,n$. Therefore the production quantity of the plant i from the beginning of the production and until the time t_j, $j = 1,2,...,k$, $i = 1,2,...,n$ is defined according to the following equation (2).

$$q_i^j = \frac{t_j}{t_k} q_i^k. \quad (2)$$

Therefore, for all $j = 1,2,...,k$, $i = 1,2,...,n$ we have

$$E(q_i^j|c_i) = \frac{t_j}{t_k}(\varphi_i c_i + \gamma_i). \quad (3)$$

In addition, we assume a normal distribution of the total output Q^j, $j = 1,2,...,k$ of all n plants, statistical independence among the plants and nonnegativity, i.e. $E(p_i) - 4\sigma(p_i) > 0$, $E(q_i^j) - 4\sigma(q_i^j) > 0$ for $i = 1,2,...,n$ and $j = 1,2,...,k$. We emphasize that despite this assumption, it is not necessary to assume any specific distribution for the random variables P_i, p_i, q_i^j, $i = 1,2,...,n$, $j = 1,2,...,k$ since normality of the random variables Q^j can be justified by its definition together with the Central Limit Theorem.

Following Laslo (2003) we presume a fixed coefficient variance (FCV) model. This model assumes that the expected production quantity and the production quantity's standard deviation are both affected by additional budget, but the production coefficient variance is constant for any budget c_i and in any time:

$$K_i = \frac{E(q_i^k|c_i)}{\sigma(q_i^k|c_i)} = \frac{E(p_i)}{\sigma(p_i)}, \; i = 1,...,n, \quad (4)$$

in other words, wherever the average performance is increased, the standard deviation is also increased and at the same rate.

By assumptions (2) and (4) for all $j = 1,2,...,k$, $i = 1,2,...,n$ we have

$$K_i = \frac{E(q_i^j|c_i)}{\sigma(q_i^j|c_i)} = \frac{E(p_i)}{\sigma(p_i)}. \quad (5)$$

Finally we assume that for each point of time t_j, $j = 1,2,...,k$, a new delivery order O^j with random demand is set for the product. Hence there are $k \geq 1$ stochastic delivery orders for the product of the plants. An order O^j is a random variable with known

distribution and must be supplied with probability of at least $(1-\alpha_j)$, $j=1,2,...,k$. The delivery order O^j does not depend on the production quantities of plants. We study here in details the case where for all $j=1,2,...,k$, the distribution of the O^j are normal distributed with known expected value and variance. For other situations (non-normal distribution for the delivery orders O^j) the analysis can be more complex but is based on similar considerations.

The main objective of the problem is to find the minimal budget and its distribution among all plants in order to ensure the fulfillment of all orders subject to the required probabilities.

4. The Solution

For all $j=1,2,...,k$ we need to fulfil the following inequalities

$$P\left(Q^j > O^1+...+O^j \middle| \underline{c}\right) \geq 1-\alpha_j \ , \ j=1,2,...,k \ ,$$

or equivalently

$$P\left(Q^j - \sum_{m=1}^{j} O^m > 0 \middle| \underline{c}\right) \geq 1-\alpha_j \ , \ j=1,2,...,k \ . \qquad (6)$$

Since Q^j and $\sum_{m=1}^{j} O^m$ are independent normally distributed random variables, the random variable $Q^j - \sum_{m=1}^{j} O^m$ also has a normal distribution with the following expectation and variance:

$$E\left(Q^j - \sum_{m=1}^{j} O^m \middle| \underline{c}\right) = \sum_{i=1}^{n} E\left(q_i^j \middle| \underline{c}\right) - \sum_{m=1}^{j} E\left(O^m\right) = \frac{t_j}{t_k}\sum_{i=1}^{n}\left(\varphi_i c_i + \gamma_i\right) - \sum_{m=1}^{j} E\left(O^m\right),$$

$$V\left(Q_j - \sum_{m=1}^{j} O^m \middle| \underline{c}\right) = \sum_{i=1}^{n} V\left(q_i^j \middle| \underline{c}\right) + \sum_{m=1}^{j} V\left(O^m\right) = \left(\frac{t_j}{t_k}\right)^2 \sum_{i=1}^{n} \frac{\left(\varphi_i c_i + \gamma_i\right)^2}{K_i^{\,2}} + \sum_{m=1}^{j} V\left(O^m\right).$$

By (6) we get $\left(Q^j - \sum_{m=1}^{j} O^m\right)_{\alpha_j} \geq 0$, where $\left(Q^j - \sum_{m=1}^{j} O^m\right)_{\alpha_j}$ is the α_j quintile of the $Q^j - \sum_{m=1}^{j} O^m$ distribution. Straightforwardly we have:

$$\left(Q^j - \sum_{m=1}^{j} O^m\right)_{\alpha_j} = \frac{t_j}{t_k}\sum_{i=1}^{n}\left(\varphi_i c_i + \gamma_i\right) - \sum_{m=1}^{j} E\left(O^m\right) + Z_{\alpha_j}\sqrt{\left(\frac{t_j}{t_k}\right)^2 \sum_{i=1}^{n} \frac{\left(\varphi_i c_i + \gamma_i\right)^2}{K_i^{\,2}} + \sum_{m=1}^{j} V\left(O^m\right)}, \qquad (7)$$

where Z_{α_j} is the α_j quintile of the normal standard distribution.

Therefore, by (6), (7), the solution for the problem is equivalent finding the vector of optimal budgets $\underline{c} = (c_1,...,c_n)$ that minimizes the total budget $C = \sum_{i=1}^{n} c_i$,

subject to:

$$\frac{t_j}{t_k}\sum_{i=1}^{n}(\varphi_i c_i+\gamma_i)+Z_{\alpha_j}\sqrt{\left(\frac{t_j}{t_k}\right)^2\sum_{i=1}^{n}\frac{(\varphi_i c_i+\gamma_i)^2}{K_i^{\,2}}+\sum_{m=1}^{j}V(O^m)}\geq\sum_{m=1}^{j}E(O^m),\quad j=1,2,...,k,\qquad(8)$$

and subject to the budget constraints:

$$c_{E(p_i)}\leq c_i\leq c_{E(P_i)}\quad,\quad i=1,...,n$$

The following inequality (9) guarantees that any additional budget in each plant i, $i=1,2,...,n$ increases the probability that the total production will fulfil the cumulative delivery order constraints until time t_j for all $j=1,2,...,k$.

Statement 1. If

$$\sqrt{\sum_{i=1}^{n}\left(\frac{\varphi_i c_i+\gamma_i}{K_i}\right)^2+\left(\frac{t_k}{t_j}\right)^2\sum_{m=1}^{j}V(O^m)}\geq -Z_{\alpha_j}\left(\frac{\varphi_i c_i+\gamma_i}{K_i^{\,2}}\right),\qquad(9)$$

for all $i=1,2,...,n$, $j=1,2,...,k$, then the quintile $\left(Q^j-\sum_{m=1}^{j}O^m\right)_{\alpha_j}$ is an increasing function of c_i for all $i=1,2,...,n$, $j=1,2,...,k$.

The following *Proposition 1* provides the necessary and sufficient conditions for the existence of a unique solution for the considered problem.

Proposition 1.

If for all $i=1,2,...,n$, $j=1,2,...,k$, equation (9) holds then the considered problem has a unique solution if and only if for all $j=1,2,...,k$

$$\frac{t_j}{t_k}\sum_{i=1}^{n}E(P_i)+Z_{\alpha_j}\sqrt{\left(\frac{t_j}{t_k}\right)^2\sum_{i=1}^{n}\left(\frac{E(P_i)}{K_i}\right)^2+\sum_{m=1}^{j}V(O^m)}\geq\sum_{m=1}^{j}E(O^m).\qquad(10)$$

Proposition 1, i.e., inequalities (9) and (10) guarantee the existence of a unique optimal solution for the considered problem.

In order to solve the considered chance-constrained programming problem, we can solve its certainty equivalent as a mathematical programming problem as defined by (8). But since the total budget is bounded: $\sum_{i=1}^{n}c_{E(p_i)}\leq C\leq\sum_{i=1}^{n}c_{E(P_i)}$ and since in any real life problem a budget is not a continuous entity, the budget can be considered as if it has a finite number of alternative values. Hence, after verifying by (9), (10) the existence and uniqueness of the optimal solution, one can attain it by examination of all the finite integer possibilities for budget allocation (dollars or cents), satisfying the constraints. Alternatively, the optimal solution can be obtained by optimization software package.

Remark 1.

A deterministic delivery order O^j can be considered as a "normal" random variable with expectation $E(O^j)=O^j$ and variance $V(O^j)=0$. Therefore the case where

all the delivery orders are deterministic is only a special case of the considered problem and our analysis is also valid for this case.

5. Numerical Examples

We consider a situation with 3 plants and 2 monthly demands: $n=3$, $k=2$.

The upper bound probabilities α_j for full supplying of demands O^j, $j=1,2$ are:

$\alpha_1=0.001$, $\alpha_2=0.025$. That is, $Z_{0.001}=-3.09023$, $Z_{0.05}=-1.95996$.

Also given:

$E(p_1)=25.00$, $E(P_1)=220.00$, $c_{E(p_1)}=75.00$, $c_{E(P_1)}=250.00$, $\sigma(p_1)=8.00$,

$E(p_2)=50.00$, $E(P_2)=250.00$, $c_{E(p_2)}=100.00$, $c_{E(P_2)}=350.00$, $\sigma(p_2)=2.00$,

$E(p_3)=50.00$, $E(P_3)=200.00$, $c_{E(p_3)}=25.00$, $c_{E(P_3)}=450.00$, $\sigma(p_3)=5.00$,

$t_1=50$, $t_2=100$.

Then we have:

$\varphi_1=1.11429$, $\gamma_1=-58.57140$, $K_1=\dfrac{25}{8}=3.12500$,

$\varphi_2=0.80000$, $\gamma_2=-30.00000$, $K_2=25.00000$,

$\varphi_3=0.35294$, $\gamma_3=41.17650$, $K_3=10.00000$.

First we consider a situation with deterministic monthly demands:

$O^1=200.00$, $O^2=150.00$. That is, by Remark 1, $E(O^j)=O^j$, $V(O^j)=0$, $j=1,2$.

By a straightforward calculation we find that the equations (9), (10) are valid for this example. Therefore by Proposition 1 there is a unique optimal solution of the considered problem. By examination of all the finite integer possibilities for budget allocation, satisfying the constraints (8), we get this optimal vector of the budget allocation among all plants:

$$(c_1,c_2,c_3)=(114.70,350.00,373.37),$$

and the total optimal (minimal) budget allocated is $C=\sum_{i=1}^{3}c_i=838.07$.

Secondly we consider the same situation as in the previous case, but with stochastic normal distributed demands such that: $O^1\sim N(200,20^2)$, $O^2\sim N(150,15^2)$.

By a straightforward calculation we find that the equations (9), (10) are valid for this example too. Therefore by Proposition 1 there is a unique optimal solution for the considered problem. By examination of all the finite integer possibilities for budget allocation, satisfying the constraints (8), we get this optimal vector of budget allocation among all factories:

$$(c_1,c_2,c_3)=(204.42,350.00,450.00),$$

the total optimal (minimal) budget allocated to all factories is $C=\sum_{i=1}^{3}c_i=1,004.42$.

Finally, we consider the same situation as in the previous case, but with stochastic uniform distributed delivery orders such that: $O^1 \sim Uni(170,230)$, $O^2 \sim Uni(125,175)$.

Based on analysis which is similar to that presented for two previous examples, and by examination of all the finite integer possibilities for budget allocation we find that the optimal vector of the budget allocation among all plants:

$$(c_1, c_2, c_3) = (148.98, 350.00, 450.00),$$

the total optimal (minimal) budget allocated to all plants is $C = \sum_{i=1}^{3} c_i = 948.98$.

6. Summary and Conclusions

This paper gives a comprehensive analysis for the problem and a procedure that can help management to solve it, i.e., to determine how much budget is needed and how to distribute the budget among the plants in order to increase its capabilities and to guarantee the fulfillment of the orders under some chance constraints.

The first step is to verify that the problem has a feasible solution. This can be done by Proposition 1 that states roughly that the expected total capabilities of all plants must be sufficiently larger than the total cumulative orders at any time. We have proved that if the allocation of the crashed budgets $c_{E(P_i)}$ to each plant $i = 1,...,n$ is a feasible solution, then the problem has a unique optimal solution. The optimal solution can be obtained by a discrete search among the bounded budget or by optimization software package.

Although we assumed normal distributions for all the random variables, we demonstrated that even if the order quantities have non normal distributions, the considered problem can be solved in a similar way. Solutions for normal distributions and uniform distributions were presented through numerical examples.

References

1. Enns, S.T. **MRP performance effects due to forecast bias and demand uncertainty**, European Journal of Operational Research, 138, 2002, pp. 87-102
2. Laslo, Z. **Activity time-cost tradeoffs under time and cost chance constraints**, Computers & Industrial Engineering, 44, 2003, pp. 365-384
3. Laslo, Z. and Gurevich, G. **Minimal budget for activities chain with chance constrained lead-time**, International Journal of Production Economics, 107, 2007, pp. 164-172
4. Laslo, Z., Gurevich, G. and Keren, B. **Economic distribution of budget among producers for fulfilling orders under delivery chance constraints**, International Journal of Production Economics, 122, 2009, pp. 656-662
5. Yano, C. and Lee, H.L. **Lot sizing with random yields: a review**, Operations Research, 43, 1995, pp. 311-334

JOURNAL OF APPLIED QUANTITATIVE METHODS

International Symposium on Stochastic Models in Reliability Engineering, Life Sciences and Operations Management (SMRLO'10)

PRODUCTIVITY ASSESSMENT AND IMPROVEMENT MEASUREMENT OF DECISION MAKING UNITS - AN APPLICATION FOR RANKING CITIES IN ISRAEL

Yossi HADAD
Prof., Industrial Engineering and Management Department,
SCE - Shamoon College of Engineering,
Beer-Sheva, Israel

E-mail: yossi@sce.ac.il

Baruch KEREN
PhD, Industrial Engineering and Management Department,
SCE - Shamoon College of Engineering,
Beer-Sheva, Israel

E-mail: baruchke@sce.ac.il

Avner BEN-YAIR
PhD, Center for Reliability and Risk Management,
SCE - Shamoon College of Engineering,
Beer-Sheva, Israel

E-mail: avner2740@gmail.com

Abstract: *In this paper we will demonstrate how productivity and improvement rate of urban organizational units (called also Decision Making Units - DMU's) may be assessed when measured along several time periods. The assessment and subsequent ranking of cities is achieved by means of the Data Envelopment Analysis (DEA) methodology to determine DMU's efficiency for each period, the Cross Efficiency ranking method to rank DMU's and the Malmquist Index approach which measures changes in productivity relative to a base period. The above combined methodology will be applied to a case study of 70 Israeli cities in years 2006, 2007 and 2008.*

Key words: *Data Envelopment Analysis (DEA); Malmquist Index; ranking methods*

1. Introduction

Banks, insurance companies, widespread food chains, police stations, etc., are organizations that have several branches and subunits (Decision Making Units - DMUs). Such organizations are often interested in assessing the productivity of their DMU's in two main aspects: 1) Relative efficiency of each DMU per every time period; 2) General improvement trends among DMU's. The importance of this assessment boils down to first know better relative productivity of each DMU as compared to other structural units, and in addition to be aware of improvement trends characteristic of every DMU. Profound knowledge of the above parameters enables decision makers in the regarded organizations to assess better specific performance of each division unit as well as their timely changes, thus contributing to a broader managerial view on every DMU within the organization. In our application we assume that cities are subunits of the organization, and therefore their relative productivities and their improvement trends should be estimated.

Relative efficiency of each DMU (productivity assessment) for every time period may be investigated by means of the DEA methodology (the so-called CCR or BCC models) which have been primarily suggested by Charnes, Cooper and Rhodes (CCR) in 1978 and subsequently developed and expanded by Banker, Charnes and Cooper (BCC) in 1984. The CCR model calculates the Technical and Scale Efficiency (TSE) while BCC determines Technical Efficiency (TE). In addition to that, productivity assessment may be facilitated by means of conventional ranking methods, like the Super Efficiency method (SE) developed by Anderson and Peterson (1993), the Cross Efficiency method (CE) introduced by Sexton et al. (1994), the bi-criteria method for efficient DMU ranking suggested by Hadad and Friedman (2004), as well as a combination of the AHP method (Analytical Hierarchic Process) and the DEA methodology described by Sinuany-Stern et al. (2000). A comprehensive review of contemporary ranking methods can be found in a study by Adler et al. (2002).

In recent years, a variety of scientific papers tackling the problem of productivity assessment by means of DEA and ranking methods, have been published and are available to the broad scientific community. Among others, one should mention Sueyoshi (1992) who measured the industrial performance of 35 Chinese cities by means of Data Envelopment Analysis, Doyle and Green (1994) who ranked 20 universities in the UK; Hadad, et al. (2009), to carry out comparative efficiency assessment and ranking of public defence authority in Israel; Malul, et al. (2009), measuring and ranking of economic, environmental and social efficiency of countries; Hadad, et al. (2007), measuring efficiency of restaurants; Hadad, et al. (2004), evaluating hotel advertisements efficiency using DEA; Hadad, et al. (2004), ranking fish farms.

Relative improvement trend for each DMU may be assessed when comparing productivities determined for every pair of consecutive time periods, or by means of Malmquist Index approach primarily suggested by Caves, et al. (1982) and subsequently developed by Fare, et al. (1985) and Fare, et al. (1994). This method investigates the improvement measure of each DMU during every pair of consecutive time periods. Should the amount of time periods exceed 2, the procedure boils down to determining improvement levels for consecutive time periods with subsequently calculating the mean geometrical product for all values obtained during the regarded complex time period (Coelli, (1996)). Practical implementation of the Malmquist Index approach have been demonstrated by Barros (2006) who investigated relative efficiency of 33 police stations in Lisbon during

2001-2002. In his research, Barros also estimates the total productivity change of the Lisbon Police Force.

In the present paper, we will demonstrate joint implementation of the DEA methodology, the Cross Efficiency ranking method and the Malmquist Index approach for productivity assessment of 70 Israeli cities between 2006-2008. The cities will be estimated every year separately upon a variety of parameters which we will regard as inputs or outputs. In addition, we will verify existence of positive correlation between the productivity ratios calculated by means of the Cross Efficiency method versus the Malmquist Index approach.

Our paper is organized in the following way. The next section introduces the presentation and formulation of DEA procedures, which will be employed in the analysis and the Super Efficiency. The third section presents the Malmquist Index approach. Part four illustrates how to evaluate city advertisements' efficiency using the models that have been described in sections two and three. Finally, the findings are presented along with conclusions and recommendations for future research.

2. Data Envelopment Analysis and Cross Efficiency

2.1. Data Envelopment Analysis

DEA is a procedure designed to measure the relative efficiency in situations when there are multiple inputs and multiple outputs and no obvious objective how to aggregate both inputs and outputs into a meaningful index of productive efficiency DEA was developed by Charnes Cooper and Rhodes (CCR) (1978). The method provides a mechanism for measuring the efficiency of each Decision-Making Unit (DMU).

The efficiency in CCR model is termed Technical and Scale Efficiency (TSE) and the relative efficiency of a DMU is defined as the ratio of its total weighted output to its total weighted input. The BCC model, named after Banker, Charnes and Cooper (1984) allow the production function to exhibit non-constant return to scale (Banker and Chang 1995)) whiles the CCR model imposes the additional assumption of constant returns to scale on the production function.

The formulation of CCR model for unit k is:

$$\text{Maximize } h_k = \sum_{r=1}^{s} U_r^k Y_{rk}$$

subject to

$$\sum_{r=1}^{s} U_r^k Y_{rj} - \sum_{i=1}^{m} V_i^k X_{ij} \le 0 \quad \text{for } j=1,2,..n, \qquad (1)$$

$$\sum_{i=1}^{m} V_i^k X_{ik} = 1,$$

$$U_r^k \ge \varepsilon > 0 \qquad r=1,2,..s,$$

$$V_i^k \ge \varepsilon > 0 \qquad i=1,2,...m.$$

where ε is defined as an infinitesimal constant (a non-Archimedean quantity).

2.2. The Cross Efficiency

The Cross Evaluation matrix was first developed by Sexton et al. (1994). This method calculated the efficiency score of each unit n times using the optimal weights evaluated by each run. The results of all the DEA cross efficiency are summarized in a matrix as given in (2)

$$h_{kj} = \frac{\sum_{r=1}^{s} U_r^k Y_{rj}}{\sum_{i=1}^{m} V_i^k X_{ij}}, \quad j = 1,2,...,n, \quad k = 1,2,...,n. \tag{2}$$

Thus h_{kj} represents the score given to unit j by the optimal weights of unit k. The elements in the diagonal h_{kk} represent the standard DEA scores h_k. The Cross Efficiency ranking method utilized the matrix h_{kj} for ranking the units one scale.

Ranking of DMUs is thus based on the average cross efficiency score being calculated as

$$h_k = \frac{\sum_{j=1}^{n} h_{kj}}{n}.$$

3. The Malmquist Index Approach

To investigate improvement of productivity, Fare et al. (1994) have demonstrated that DEA methodology may be applied to assess Malmquist Total Factor Productivity (TFP) index numbers. As a matter of fact, Malmquist index is an approach enabling relative measurement of productivity changes between consecutive periods of time (e.g., a year). Those productivity changes may be broken down to structural elements depending on technical efficiency enhancement as well as technology changes and progress. The Malmquist DEA approach determines efficiency level in a certain year relatively to the previous one, thus enabling evaluation of productivity improvement between the two consecutive periods.

The Malmquist TFP index measures efficiency in each period t related to the base period s in terms of productivity improvement. Fare et al. (1994) specifies an output based Malmquist productivity change index:

$$m_0(y_s, x_s, y_t, x_t) = \left[\frac{d_0^s(y_t, x_t)}{d_0^s(y_s, x_s)} \times \frac{d_0^t(y_t, x_t)}{d_0^t(y_s, x_s)} \right]^{\frac{1}{2}} \tag{3}$$

where the notations $d_0^s(y_t, x_t)$, $d_0^t(y_t, x_t)$, $d_0^s(y_s, x_s)$, $d_0^t(y_s, x_s)$ are distance functions and x, y are the output and input vector.

An equivalent way of writing this would be

$$m_0(y_s, x_s, y_t, x_t) = \frac{d_0^t(y_t, x_t)}{d_0^s(y_s, x_s)} \left[\frac{d_0^s(y_t, x_t)}{d_0^s(y_t, x_t)} \times \frac{d_0^t(y_s, x_s)}{d_0^t(y_s, x_s)} \right]^{\frac{1}{2}}. \tag{4}$$

The above equation can be broken into two parts, namely the efficiency change component and the technical change component:

$$\text{Efficiency change} = \frac{d_0^t(y_t, x_t)}{d_0^s(y_s, x_s)} \tag{5}$$

and

$$\text{Technical change} = \left[\frac{d_0^s(y_t, x_t)}{d_0^s(y_t, x_t)} \times \frac{d_0^t(y_s, x_s)}{d_0^t(y_s, x_s)}\right]^{\frac{1}{2}}. \tag{6}$$

The Malmquist productivity change index may be determined as the result of solving 4 linear programming problems as listed below:

$$\begin{cases} d_0^s(y_t, x_t) = \min\theta \\ subject\ to \\ -y_{ot} + Y_s\lambda \geq 0 \\ \theta x_{0t} - X_s\lambda \geq 0 \\ \lambda \geq 0 \end{cases} \tag{7}$$

$$\begin{cases} d_0^s(y_s, x_s) = \min\theta \\ subject\ to \\ -y_{os} + Y_s\lambda \geq 0 \\ \theta x_{0s} - X_s\lambda \geq 0 \\ \lambda \geq 0 \end{cases} \tag{8}$$

$$\begin{cases} d_0^t(y_t, x_t) = \min\theta \\ subject\ to \\ -y_{os} + Y_s\lambda \geq 0 \\ \theta x_{0s} - X_s\lambda \geq 0 \\ \lambda \geq 0 \end{cases} \tag{9}$$

$$\begin{cases} d_0^t(y_t, x_t) = \min\theta \\ subject\ to \\ -y_{ot} + Y_t\lambda \geq 0 \\ \theta x_{0t} - X_t\lambda \geq 0 \\ \lambda \geq 0 \end{cases} \tag{10}$$

In (7-10), θ is a scalar, λ is a vector that representative the constants. The value of θ will be the efficiency score for the i-th DMU. It will satisfy θ less than or equal to 1, with a value of 1 indicating a point on the frontier and hence a technically efficient DMU. These four LPs must be solved for each DMU in the sample.

4. The Case Study on Israeli Cities

4.1. Determining DMUs

In order to proceed with the DEA procedure one has to determine first Decision Making Units (DMUs). In our research, we decided to determine DMUs as Israeli cities comprising a total of 10,000 inhabitants at least. The latter data has been adopted from the Central Bureau Statistics database published in 2006, 2007, 2008.

4.2. Selection of outputs and inputs

An important issue in employing DEA is the selection of inputs and outputs. In order to calculate the efficiency and the score of each city entering this study the following outputs and inputs have been implemented:

Inputs

X_1 - negative emigration percentage - the ratio between citizens that left the city to the total number of inhabitants (negative emigration ratio);

X_2 - percentage of unemployed citizens and obtaining minimal income insurance;

X_3 - percentage of deceased throughout the year;

X_4 - average number of schoolchildren in a classroom;

X_5 - average number of apartments per citizen;

X_6 - average spending by the local authority per citizen (in thousands NIS per citizen).

Outputs

Y_1 - positive immigration percentage - the ratio between citizens that joint the city to the total number of inhabitants (positive immigration ratio);

Y_2 - average monthly income per citizen (in thousands NIS per month);

Y_3 - percentage of successfully graduating from the city school system that complied with university entrance requirements;

Y_4 - average number of private vehicles per citizen;

Y_5 - average city income per citizen (including donations from the state).

We will demonstrate the regarded procedure for Year 2008. The data on 70 cities with 5 outputs and 6 inputs are given in Table 1:

Table 1. The numerical data for 2008

City	Year	X1	X2	X3	X4	X5	X6	Y1	Y2	Y3	Y4	Y5
Umm el-Faheim	2008	0.41%	16.21%	0.30%	30.25	0.28	4107.84	0.11%	3588.11	23.00%	0.1977	12924.81
OFAKIM	2008	3.77%	14.35%	0.55%	20.91	0.24	5863.73	0.09%	4182.77	37.60%	0.1604	17231.31
Or Yehuda	2008	3.53%	5.73%	0.49%	25.19	0.30	6250.42	0.11%	5706.27	43.60%	0.3121	20275.02
Or Akiva	2008	3.89%	8.03%	0.88%	21.25	0.31	6461.46	0.08%	4445.78	43.60%	0.2429	20919.81
Eilat	2008	10.99%	5.12%	0.32%	25.94	0.35	13736.39	0.10%	5334.78	51.80%	0.2974	39797.78
Ariel	2008	4.76%	3.51%	0.44%	27.52	0.28	6098.48	0.07%	5480.70	39.90%	0.2406	20728.72
Ashdod	2008	2.37%	8.54%	0.57%	26.61	0.29	5156.99	0.10%	5590.69	46.50%	0.1952	17595.95
Ashkelon	2008	2.84%	11.52%	0.69%	28.62	0.32	4692.72	0.10%	4989.72	49.40%	0.2343	16448.67
BakaJat	2008	0.76%	4.28%	0.34%	30.89	0.23	15958.20	0.08%	4275.78	41.00%	0.2457	55935.87
Beer-Sheva	2008	3.46%	10.95%	0.68%	26.22	0.37	5711.99	0.06%	5515.36	43.60%	0.2252	19125.60
Beit-Shean	2008	2.54%	9.35%	0.51%	20.99	0.28	9391.22	0.10%	4530.72	42.20%	0.2471	23928.66
Beit-Shemesh	2008	3.37%	4.07%	0.27%	24.07	0.22	3904.86	0.13%	5187.96	31.30%	0.1227	13059.96
Beitar-Illit	2008	2.49%	4.11%	0.10%	23.42	0.19	4100.62	0.19%	3345.32	25.30%	0.0523	10698.04
Bney-Brak	2008	3.74%	4.25%	0.46%	25.73	0.26	5249.03	0.08%	4613.07	36.00%	0.5272	17272.95
Bat-Yam	2008	4.73%	5.73%	0.99%	26.37	0.37	5314.27	0.09%	4755.76	42.50%	0.2486	17487.41
Givatayim	2008	6.40%	1.76%	0.92%	30.63	0.45	6798.70	0.11%	8775.23	66.30%	0.3988	21178.30
Dimona	2008	3.19%	15.87%	0.70%	24.51	0.33	5676.74	0.06%	5677.29	35.70%	0.1778	18799.31
Hod-Hasharon	2008	3.74%	1.42%	0.36%	28.36	0.31	6234.17	0.14%	9580.47	66.10%	0.3808	20975.07
Herzliya	2008	4.48%	1.90%	0.70%	27.34	0.39	8773.56	0.09%	8637.24	64.10%	0.4884	29958.56
Hadera	2008	3.06%	5.82%	0.75%	26.32	0.34	6128.19	0.10%	5492.21	44.60%	0.2937	19964.86
Holon	2008	3.67%	3.46%	0.72%	27.65	0.36	5580.70	0.09%	6013.82	53.10%	0.3474	18974.67
Haifa	2008	3.18%	7.38%	0.99%	25.49	0.42	8035.95	0.08%	7030.48	60.00%	0.3541	26247.90
Tiberias	2008	4.20%	12.19%	0.55%	24.13	0.35	8101.73	0.06%	4381.15	35.90%	0.2428	22693.33
Taibe	2008	0.34%	14.00%	0.38%	31.61	0.20	9559.56	0.13%	3898.18	30.90%	0.2393	26776.78
Tierra	2008	0.44%	3.60%	0.35%	28.73	0.25	3817.01	0.14%	3934.20	40.60%	0.2806	11242.56
Tirat-Carmel	2008	2.93%	10.66%	0.84%	22.98	0.33	7091.02	0.06%	4829.38	46.90%	0.2474	25790.75

City	Year	X1	X2	X3	X4	X5	X6	Y1	Y2	Y3	Y4	Y5
Tamra	2008	0.44%	21.37%	0.33%	29.18	0.23	5344.81	0.11%	3621.48	36.30%	0.2158	16442.14
Yavne	2008	3.66%	6.82%	0.42%	26.17	0.28	6130.53	0.07%	6679.61	51.70%	0.3398	22593.32
Yehud	2008	4.30%	2.25%	0.48%	28.56	0.32	10237.37	0.10%	8353.91	61.50%	0.3870	26502.27
Jerusalem	2008	2.35%	3.66%	0.44%	26.23	0.25	4900.07	0.06%	5699.61	31.60%	0.2057	15893.69
Kfar-Saba	2008	4.13%	1.95%	0.68%	28.27	0.31	6634.08	0.09%	7856.73	68.10%	0.4168	20607.02
Karmiel	2008	3.49%	6.95%	0.68%	27.16	0.34	4962.59	0.09%	5732.54	51.80%	0.2510	17681.90
Lod	2008	4.09%	9.07%	0.69%	25.79	0.30	4856.54	0.07%	5015.39	37.10%	0.4340	16640.97
Migdal-Haemeq	2008	3.37%	10.14%	0.66%	23.05	0.32	7704.21	0.06%	4805.62	37.80%	0.2241	23104.41
Modiin	2008	3.59%	1.23%	0.13%	29.86	0.28	6999.50	0.23%	9856.76	74.30%	0.2941	17710.30
Maale-Adummim	2008	3.36%	2.78%	0.23%	27.71	0.26	5552.77	0.19%	6398.11	59.40%	0.2623	18134.97
Ma'alot Tarshiha	2008	3.13%	9.39%	0.51%	25.94	0.30	5882.31	0.08%	5169.07	45.70%	0.2417	19837.36
Nahariya	2008	3.61%	8.01%	0.76%	27.31	0.36	5422.02	0.09%	6357.79	54.00%	0.2948	17830.30
Ness-Ziona	2008	2.98%	3.44%	0.56%	28.02	0.33	7315.87	0.25%	8289.72	55.80%	0.3527	24775.00
Nazareth	2008	1.96%	12.21%	0.37%	29.79	0.28	4128.86	0.04%	4355.33	42.90%	0.2844	13355.96
Nazareth-Illit	2008	3.45%	10.89%	0.98%	24.88	0.39	5265.44	0.07%	4782.66	50.00%	0.2413	18210.00
Nesher	2008	5.42%	6.20%	0.65%	27.01	0.42	7332.59	0.08%	6431.81	62.90%	0.3021	24611.84
Netivot	2008	4.30%	11.45%	0.34%	22.59	0.25	5861.29	0.09%	4054.92	33.00%	0.1681	18952.17
Netanya	2008	2.51%	7.29%	0.80%	26.80	0.35	5559.58	0.11%	5564.58	45.20%	0.2623	18538.35
Skhnen	2008	0.56%	18.27%	0.24%	31.15	0.24	5999.12	0.10%	3836.95	37.10%	0.2295	15653.16
Akko	2008	3.00%	17.49%	0.72%	25.99	0.33	6689.17	0.07%	4503.33	35.10%	0.2142	21268.37
Afula	2008	3.08%	8.84%	0.60%	24.36	0.36	7129.49	0.09%	4758.94	45.00%	0.2606	22146.03
Arad	2008	4.95%	9.87%	0.83%	23.61	0.37	5793.13	0.08%	5686.36	48.10%	0.2074	19617.81
Petah-Tikva	2008	2.98%	3.36%	0.68%	26.32	0.35	6035.23	0.15%	6473.14	52.80%	0.6057	20679.46
Zfat	2008	8.03%	13.13%	0.65%	23.04	0.34	6214.73	0.04%	4377.43	30.60%	0.1903	19933.67
Qalansuwa	2008	0.53%	12.99%	0.26%	32.89	0.18	3428.85	0.11%	3866.99	27.90%	0.1958	9909.31
Kiryat- Ono	2008	4.48%	1.78%	0.59%	29.81	0.38	6755.53	0.14%	9339.45	61.30%	0.3952	21758.02
kiryat-ata	2008	3.15%	8.42%	0.77%	26.02	0.36	6070.42	0.10%	5749.87	45.90%	0.2808	20550.04
Qiryat-Bialik	2008	5.06%	6.71%	0.98%	28.92	0.40	5310.60	0.09%	6339.89	49.10%	0.3205	18057.01
Qiryat-Gat	2008	3.27%	12.51%	0.70%	25.30	0.31	6117.45	0.07%	4283.38	46.00%	0.1921	21587.22
Qiryat-Yam	2008	4.99%	11.57%	1.09%	26.24	0.39	5526.85	0.06%	5020.33	38.00%	0.2246	18136.94
Kiriat-Motzkin	2008	4.88%	6.53%	0.83%	30.82	0.37	4963.27	0.09%	6567.83	53.80%	0.2832	17301.74
kiryat-malchy	2008	4.02%	16.59%	0.40%	23.68	0.28	9024.10	0.07%	4001.12	29.10%	0.2242	26195.74
kiryat-shmona	2008	3.79%	7.56%	0.56%	23.22	0.35	8378.68	0.07%	4843.79	43.40%	0.2830	24846.45
Rosh-Haayin	2008	3.25%	3.14%	0.35%	27.73	0.27	6932.40	0.11%	7142.81	50.70%	0.3206	22670.03
Rishon Lezion	2008	3.30%	3.95%	0.55%	28.48	0.31	4757.82	0.10%	7072.96	54.60%	0.3509	17637.87
Rahat	2008	0.49%	28.86%	0.27%	30.92	0.08	4028.05	0.17%	3584.18	24.70%	0.1333	12866.80
Rehovot	2008	3.90%	5.64%	0.63%	27.88	0.35	5655.50	0.10%	6996.33	50.20%	0.3036	19575.62
Ramla	2008	3.54%	7.09%	0.63%	25.46	0.27	5155.45	0.06%	4692.08	36.10%	0.2588	17390.87
Ramat-Gan	2008	5.47%	2.31%	0.89%	27.10	0.43	6174.25	0.10%	7621.22	65.60%	0.3820	22463.48
Ramat-Hasharon	2008	4.37%	1.20%	0.54%	27.56	0.36	9373.42	0.11%	10676.27	69.20%	0.4843	31185.27
Raanana	2008	4.59%	1.23%	0.45%	27.94	0.29	7383.13	0.08%	10102.80	70.00%	0.3767	27198.43
Sderot	2008	5.28%	10.38%	0.68%	21.52	0.32	7895.45	0.06%	4417.50	40.90%	0.2264	25164.65
Shefaram	2008	0.62%	17.28%	0.35%	31.09	0.24	4319.61	0.15%	4272.76	42.20%	0.2464	14148.12
Tel-Aviv	2008	5.45%	4.20%	0.93%	24.14	0.47	10503.32	0.10%	7780.98	58.40%	0.5925	35566.68
Umm el-Faheim	2007	0.52%	16.49%	0.27%	30.36	0.28	4232.17	0.07%	3503.46	30.10%	0.1942	9168.60
OFAKIM	2007	3.84%	14.81%	0.68%	21.18	0.25	5765.02	0.08%	4226.35	31.00%	0.1550	11569.14
Or Yehuda	2007	3.38%	6.33%	0.61%	25.14	0.31	6571.42	0.12%	5478.12	37.20%	0.3148	13865.64
Or Akiva	2007	3.81%	11.68%	0.64%	22.02	0.30	6749.95	0.08%	4194.08	41.80%	0.2368	13992.22

City	Year	X1	X2	X3	X4	X5	X6	Y1	Y2	Y3	Y4	Y5
Eilat	2007	10.49%	5.45%	0.38%	25.74	0.34	9775.43	0.11%	5150.50	49.30%	0.2864	22091.07
Ariel	2007	4.10%	3.66%	0.43%	26.77	0.27	6054.18	0.09%	5243.58	42.60%	0.2378	14106.72
Ashdod	2007	2.36%	9.72%	0.59%	26.48	0.29	4758.08	0.11%	5091.81	47.60%	0.1932	12479.80
Ashkelon	2007	2.55%	13.87%	0.74%	28.88	0.31	4781.68	0.10%	4780.43	46.80%	0.2312	11688.01
BakaJat	2007	0.55%	8.54%	0.35%	31.34	0.24	4643.00	0.13%	4365.51	41.50%	0.2409	8989.90
Beer-Sheva	2007	3.20%	12.08%	0.69%	26.41	0.37	5514.61	0.06%	5620.23	43.10%	0.2243	12955.26
Beit-Shean	2007	2.91%	10.29%	0.49%	21.24	0.28	9637.81	0.09%	4624.84	35.10%	0.2393	15766.02
Beit-Shemesh	2007	3.09%	4.96%	0.25%	24.27	0.23	4466.11	0.16%	4912.60	32.50%	0.1267	9050.65
Beitar-Illit	2007	2.50%	4.42%	0.08%	22.39	0.18	4823.16	0.20%	3197.23	32.50%	0.0533	6332.90
Bney-Brak	2007	3.74%	4.24%	0.48%	25.66	0.26	5173.50	0.08%	4514.19	32.50%	0.4731	11729.91
Bat-Yam	2007	4.59%	6.08%	1.04%	26.40	0.37	4958.59	0.08%	4586.62	43.60%	0.2466	11771.83
Givatayim	2007	6.10%	1.96%	1.05%	30.92	0.45	5998.44	0.12%	8259.25	66.60%	0.4071	13700.92
Dimona	2007	3.62%	17.95%	0.69%	24.59	0.33	5903.85	0.06%	5677.53	34.10%	0.1725	12518.99
Hod-Hasharon	2007	3.40%	1.69%	0.45%	28.69	0.31	5959.08	0.15%	8854.76	65.60%	0.3821	13155.74
Herzliya	2007	4.43%	1.99%	0.64%	27.40	0.39	7836.23	0.08%	8076.68	65.00%	0.5061	18810.79
Hadera	2007	3.08%	7.98%	0.69%	26.08	0.34	5741.28	0.09%	5301.80	45.60%	0.2895	13346.27
Holon	2007	3.67%	3.78%	0.75%	27.83	0.36	5003.69	0.09%	5678.08	48.80%	0.3517	12403.61
Haifa	2007	3.21%	7.98%	0.98%	25.75	0.42	7618.54	0.07%	6938.97	58.90%	0.3502	17267.25
Tiberias	2007	3.94%	12.62%	0.56%	24.46	0.35	6387.78	0.06%	4438.59	36.70%	0.2402	14632.41
Taibe	2007	0.40%	13.75%	0.34%	31.72	0.20	5383.81	0.12%	3938.88	32.50%	0.2290	4700.83
Tierra	2007	0.35%	4.15%	0.37%	28.66	0.25	4194.86	0.10%	4181.62	36.90%	0.2668	8477.00
Tirat-Carmel	2007	2.91%	11.68%	0.73%	22.31	0.33	7304.29	0.08%	4710.22	46.40%	0.2462	18785.79
Tamra	2007	0.50%	23.63%	0.23%	29.03	0.22	4873.81	0.12%	3621.66	42.90%	0.2131	9692.01
Yavne	2007	3.48%	7.40%	0.42%	26.78	0.27	6023.27	0.07%	6491.69	52.70%	0.3390	15322.93
Yehud	2007	4.00%	2.38%	0.49%	29.09	0.32	5833.91	0.08%	7496.08	58.10%	0.3859	13361.17
Jerusalem	2007	2.25%	4.27%	0.46%	26.18	0.26	5272.54	0.06%	5575.72	31.50%	0.2034	11294.58
Kfar-Saba	2007	3.69%	2.20%	0.59%	28.36	0.31	5645.17	0.10%	7502.91	65.40%	0.4124	13044.66
Karmiel	2007	3.82%	7.98%	0.69%	27.07	0.34	4790.15	0.08%	5664.58	49.20%	0.2502	11648.05
Lod	2007	4.08%	9.85%	0.65%	25.56	0.30	5331.43	0.06%	4720.34	38.40%	0.4202	12376.37
Migdal-Haemeq	2007	3.35%	11.08%	0.58%	22.64	0.32	5887.32	0.05%	4735.26	40.10%	0.2219	13558.24
Modiin	2007	3.35%	1.31%	0.09%	29.45	0.29	4988.70	0.28%	9280.00	76.70%	0.2987	11201.16
Maale-Adummim	2007	3.44%	2.59%	0.23%	27.49	0.26	5555.60	0.16%	6490.48	59.10%	0.2645	12036.37
Ma'alot Tarshiha	2007	2.90%	10.58%	0.55%	24.86	0.30	5747.82	0.09%	4989.13	45.80%	0.2367	12891.54
Nahariya	2007	3.86%	8.47%	0.88%	26.69	0.36	5404.66	0.10%	6190.95	53.00%	0.2937	11349.63
Ness-Ziona	2007	2.75%	4.36%	0.49%	27.55	0.32	7491.69	0.24%	7556.07	55.00%	0.3565	16905.03
Nazareth	2007	2.17%	18.49%	0.37%	29.67	0.28	3942.81	0.05%	4383.34	42.00%	0.2739	8725.22
Nazareth-Illit	2007	3.51%	13.06%	0.91%	26.05	0.39	5431.68	0.07%	4917.52	47.00%	0.2362	11834.19
Nesher	2007	4.41%	6.77%	0.68%	27.19	0.41	6286.14	0.09%	6251.48	60.70%	0.2994	14693.95
Netivot	2007	4.33%	12.97%	0.43%	22.56	0.26	5736.80	0.09%	3983.00	33.20%	0.1611	11997.82
Netanya	2007	2.61%	8.30%	0.80%	27.19	0.35	4993.63	0.10%	5241.44	43.90%	0.2584	12054.83
Skhnen	2007	0.51%	19.13%	0.28%	31.51	0.21	5662.27	0.11%	3753.30	37.10%	0.2223	10697.09
Akko	2007	2.88%	18.51%	0.78%	25.73	0.33	5346.91	0.07%	4531.23	37.10%	0.2102	12467.09
Afula	2007	3.12%	9.88%	0.65%	25.08	0.36	6248.25	0.09%	5077.78	46.70%	0.2579	12842.25
Arad	2007	5.30%	13.51%	0.75%	23.64	0.37	5901.65	0.06%	5531.71	50.20%	0.2084	12458.49
Petah-Tikva	2007	3.01%	3.25%	0.67%	26.49	0.35	5800.08	0.12%	6219.70	53.30%	0.5726	13728.57
Zfat	2007	5.51%	12.53%	0.62%	23.50	0.34	6438.91	0.07%	4427.93	32.80%	0.1839	13103.47
Qalansuwa	2007	0.48%	13.77%	0.25%	32.78	0.18	3460.11	0.15%	4093.00	27.80%	0.1903	6705.55
Kiryat- Ono	2007	4.28%	1.99%	0.67%	29.35	0.38	6819.38	0.13%	8539.01	63.20%	0.3971	14937.99

City	Year	X1	X2	X3	X4	X5	X6	Y1	Y2	Y3	Y4	Y5
kiryat-ata	2007	3.07%	9.24%	0.75%	26.65	0.36	5536.75	0.10%	5572.76	46.70%	0.2775	13507.75
Qiryat-Bialik	2007	4.86%	7.61%	0.85%	29.52	0.40	5137.21	0.09%	6257.44	46.90%	0.3180	11811.60
Qiryat-Gat	2007	3.67%	13.80%	0.66%	25.18	0.31	6192.98	0.05%	4189.56	43.20%	0.1897	14900.79
Qiryat-Yam	2007	4.83%	12.47%	1.12%	26.91	0.39	5721.28	0.07%	5010.57	35.20%	0.2218	12345.43
Kiriat-Motzkin	2007	4.63%	7.14%	0.87%	30.98	0.37	4721.07	0.11%	6451.54	49.80%	0.2815	11689.02
kiryat-malchy	2007	4.34%	17.79%	0.61%	23.76	0.29	7686.50	0.07%	3920.29	29.90%	0.2186	15440.64
kiryat-shmona	2007	3.69%	8.20%	0.58%	24.08	0.35	7867.33	0.07%	4830.48	51.20%	0.2725	15627.54
Rosh-Haayin	2007	3.32%	3.45%	0.28%	27.07	0.27	6189.53	0.10%	6907.95	53.70%	0.3202	14317.89
Rishon Lezion	2007	3.23%	4.24%	0.55%	28.20	0.31	4800.22	0.10%	6620.68	55.60%	0.3494	12667.93
Rahat	2007	0.66%	29.91%	0.30%	30.84	0.08	3721.42	0.07%	3596.92	26.40%	0.1303	8398.82
Rehovot	2007	3.77%	6.61%	0.65%	27.94	0.35	6453.53	0.10%	6751.53	54.40%	0.3044	13353.42
Ramla	2007	3.78%	8.19%	0.67%	25.72	0.27	8669.89	0.06%	4459.08	33.50%	0.2541	14344.04
Ramat-Gan	2007	5.38%	2.62%	0.92%	26.75	0.44	6128.87	0.10%	7211.83	64.90%	0.3841	15008.28
Ramat-Hasharon	2007	4.20%	1.39%	0.52%	27.82	0.36	9199.04	0.10%	10041.18	71.10%	0.4964	21141.94
Raanana	2007	4.08%	1.42%	0.42%	27.92	0.30	7856.34	0.10%	9419.10	68.20%	0.3788	17121.83
Sderot	2007	4.59%	13.26%	0.64%	21.38	0.32	9007.15	0.08%	4264.63	36.60%	0.2197	19469.61
Shefaram	2007	0.74%	18.79%	0.34%	30.80	0.24	4376.51	0.15%	4372.88	36.70%	0.2429	9638.01
Tel-Aviv	2007	4.95%	4.66%	0.99%	24.25	0.47	9952.09	0.12%	7275.70	56.10%	0.5826	24263.23
Umm el-Faheim	2006	0.58%	16.96%	0.29%	29.60	0.26	4031.74	0.07%	3355.61	23.00%	0.1832	3973.11
OFAKIM	2006	3.97%	15.43%	0.55%	21.93	0.25	6140.57	0.09%	4021.56	25.30%	0.1502	5746.55
Or Yehuda	2006	3.91%	6.78%	0.62%	24.89	0.31	5717.50	0.12%	5557.37	37.20%	0.3075	5401.72
Or Akiva	2006	3.92%	13.33%	0.58%	22.99	0.30	7337.37	0.09%	4124.33	35.60%	0.2324	6592.19
Eilat	2006	11.45%	5.42%	0.32%	25.13	0.35	8831.58	0.10%	5170.14	46.80%	0.2827	9009.97
Ariel	2006	5.31%	4.08%	0.46%	27.02	0.27	5536.68	0.07%	5306.77	42.40%	0.2352	5278.97
Ashdod	2006	2.35%	10.73%	0.59%	26.41	0.29	4720.83	0.13%	5168.95	44.90%	0.1903	4535.28
Ashkelon	2006	3.10%	15.67%	0.71%	29.05	0.32	4345.39	0.09%	4645.88	48.10%	0.2276	4373.55
BakaJat	2006	0.71%	9.53%	0.26%	31.44	0.23	4919.84	0.11%	3900.28	36.50%	0.2287	4234.01
Beer-Sheva	2006	3.55%	15.19%	0.71%	25.73	0.37	5368.67	0.07%	5467.78	41.30%	0.2214	5229.46
Beit-Shean	2006	3.70%	11.19%	0.42%	21.38	0.28	7908.85	0.04%	4329.42	35.00%	0.2359	6270.90
Beit-Shemesh	2006	3.62%	5.43%	0.27%	24.48	0.24	4493.94	0.20%	5016.11	34.90%	0.1304	4245.61
Beitar-Illit	2006	3.37%	4.68%	0.09%	22.34	0.19	4761.60	0.17%	3256.01	4.50%	0.0544	4532.36
Bney-Brak	2006	4.41%	4.50%	0.48%	25.01	0.26	5198.57	0.08%	4552.57	7.80%	0.4267	4694.89
Bat-Yam	2006	5.68%	7.05%	0.96%	26.12	0.37	4571.60	0.07%	4638.47	43.20%	0.2424	4596.63
Givatayim	2006	7.01%	2.23%	1.01%	29.67	0.46	5702.59	0.11%	8269.47	67.00%	0.4086	5315.72
Dimona	2006	3.84%	20.47%	0.77%	24.83	0.33	5711.13	0.05%	5330.98	37.70%	0.1653	5294.57
Hod-Hasharon	2006	3.81%	1.79%	0.41%	28.86	0.32	6273.68	0.16%	8823.08	62.60%	0.3808	6029.56
Herzliya	2006	4.86%	2.19%	0.65%	27.51	0.39	7273.56	0.08%	7997.90	65.00%	0.4821	7535.40
Hadera	2006	3.53%	9.50%	0.75%	26.36	0.34	5375.72	0.09%	5162.32	40.90%	0.2830	5235.46
Holon	2006	4.38%	4.39%	0.72%	27.70	0.35	4967.89	0.08%	5718.21	47.50%	0.3571	5035.63
Haifa	2006	3.48%	8.53%	0.95%	25.96	0.41	7408.64	0.08%	6643.68	57.40%	0.3455	7508.53
Tiberias	2006	4.84%	13.74%	0.58%	23.28	0.35	7001.30	0.05%	4309.33	30.30%	0.2371	6715.83
Taibe	2006	0.64%	13.49%	0.26%	31.39	0.19	2906.52	0.08%	3550.05	28.10%	0.2192	2801.85
Tierra	2006	0.56%	4.50%	0.32%	28.10	0.25	4657.80	0.15%	3845.52	37.80%	0.2471	3653.06
Tirat-Carmel	2006	3.22%	12.39%	0.78%	23.18	0.33	6998.89	0.08%	4541.03	38.50%	0.2440	6937.71
Tamra	2006	0.53%	24.50%	0.36%	28.55	0.22	4644.45	0.10%	3427.45	38.70%	0.2023	4291.62
Yavne	2006	4.47%	8.01%	0.43%	27.68	0.27	6024.00	0.07%	6442.54	50.40%	0.3338	6110.68
Yehud	2006	4.69%	2.71%	0.56%	28.72	0.32	5899.54	0.09%	7613.75	57.50%	0.3860	5259.71
Jerusalem	2006	2.56%	4.49%	0.45%	25.86	0.26	4424.60	0.06%	5669.84	31.70%	0.2023	4321.84

City	Year	X1	X2	X3	X4	X5	X6	Y1	Y2	Y3	Y4	Y5
Kfar-Saba	2006	3.94%	2.41%	0.60%	28.72	0.31	5668.50	0.10%	7544.79	65.30%	0.4167	6345.94
Karmiel	2006	3.54%	8.72%	0.66%	27.13	0.33	5069.12	0.11%	5434.60	48.30%	0.2459	5260.46
Lod	2006	5.09%	10.83%	0.61%	25.16	0.30	5035.33	0.06%	4800.29	34.70%	0.4074	4427.78
Migdal-Haemeq	2006	3.40%	11.75%	0.64%	22.70	0.32	5876.33	0.07%	4505.44	36.50%	0.2216	5287.48
Modiin	2006	4.03%	1.33%	0.14%	30.21	0.30	5082.54	0.24%	9652.30	72.90%	0.3005	5016.07
Maale-Adummim	2006	3.53%	2.70%	0.31%	27.69	0.26	6173.08	0.22%	6565.38	55.70%	0.2652	6310.10
Ma'alot Tarshiha	2006	3.77%	11.21%	0.52%	25.30	0.30	6158.73	0.08%	4791.92	47.70%	0.2351	6012.77
Nahariya	2006	3.89%	9.04%	0.80%	27.70	0.36	5400.13	0.12%	6111.42	50.10%	0.2873	5429.66
Ness-Ziona	2006	3.35%	4.78%	0.50%	26.91	0.34	7458.10	0.21%	7484.89	55.20%	0.3497	7623.72
Nazareth	2006	1.78%	20.65%	0.34%	29.27	0.28	4140.83	0.07%	4198.90	40.00%	0.2626	3436.44
Nazareth-Illit	2006	3.73%	14.34%	0.93%	27.36	0.38	7506.09	0.08%	4621.52	51.00%	0.2303	7612.51
Nesher	2006	5.53%	7.30%	0.61%	26.68	0.41	6298.01	0.09%	6044.25	51.80%	0.2998	6909.04
Netivot	2006	4.21%	13.40%	0.40%	22.28	0.26	5400.69	0.11%	3759.65	26.30%	0.1582	5582.78
Netanya	2006	3.00%	9.15%	0.76%	27.18	0.35	4885.10	0.11%	5210.34	45.90%	0.2517	4812.54
Skhnen	2006	0.68%	19.65%	0.25%	31.02	0.21	4255.24	0.10%	3434.83	24.50%	0.2104	4294.47
Akko	2006	3.60%	19.09%	0.82%	25.02	0.32	5880.38	0.07%	4435.51	32.90%	0.2079	5215.99
Afula	2006	3.56%	11.49%	0.66%	24.96	0.36	5958.57	0.08%	4820.73	42.20%	0.2518	5931.07
Arad	2006	6.76%	22.37%	0.66%	23.24	0.36	5529.28	0.04%	5358.44	47.70%	0.2042	4854.75
Petah-Tikva	2006	3.41%	3.36%	0.71%	26.37	0.35	6125.10	0.11%	6206.13	53.00%	0.5408	5911.18
Zfat	2006	7.29%	13.38%	0.64%	23.32	0.35	6705.75	0.06%	4238.11	30.70%	0.1866	6778.28
Qalansuwa	2006	0.68%	13.76%	0.24%	31.90	0.19	3627.31	0.12%	3832.13	30.70%	0.1783	3448.67
Kiryat- Ono	2006	5.21%	2.21%	0.57%	29.28	0.38	6409.95	0.12%	8345.80	60.90%	0.3969	5982.21
kiryat-ata	2006	3.76%	9.79%	0.76%	26.21	0.36	5897.98	0.09%	5307.11	42.50%	0.2722	5691.01
Qiryat-Bialik	2006	5.60%	8.12%	0.87%	29.56	0.39	4860.59	0.08%	5926.38	50.10%	0.3141	4743.52
Qiryat-Gat	2006	4.04%	14.72%	0.66%	25.29	0.30	6221.10	0.05%	4254.37	43.30%	0.1839	5850.21
Qiryat-Yam	2006	5.40%	12.99%	1.04%	26.55	0.38	5169.58	0.07%	4782.72	40.60%	0.2194	4797.98
Kiriat-Motzkin	2006	5.47%	7.73%	0.87%	29.45	0.37	4507.54	0.08%	6109.22	50.50%	0.2846	4224.28
kiryat-malchy	2006	4.71%	18.84%	0.52%	23.61	0.29	7513.38	0.08%	3937.17	28.50%	0.2119	6970.30
kiryat-shmona	2006	4.25%	8.91%	0.52%	23.37	0.34	7824.78	0.08%	4613.46	42.30%	0.2677	7282.38
Rosh-Haayin	2006	3.95%	3.57%	0.34%	27.45	0.28	6158.39	0.09%	7009.25	48.10%	0.3143	6258.82
Rishon Lezion	2006	3.81%	4.70%	0.52%	28.34	0.31	4941.75	0.11%	6790.28	54.80%	0.3456	5211.72
Rahat	2006	0.78%	32.02%	0.35%	30.36	0.08	3369.27	0.09%	3621.49	16.30%	0.1230	3593.39
Rehovot	2006	3.96%	7.57%	0.73%	28.40	0.35	5818.41	0.12%	6736.51	52.30%	0.2984	5713.27
Ramla	2006	3.96%	9.31%	0.60%	26.60	0.27	5506.30	0.08%	4533.86	31.00%	0.2481	5131.61
Ramat-Gan	2006	6.22%	3.01%	0.87%	26.89	0.44	5935.48	0.10%	7191.15	63.20%	0.3875	5987.96
Ramat-Hasharon	2006	4.86%	1.42%	0.55%	27.69	0.36	7390.71	0.09%	9664.47	66.10%	0.5069	7373.36
Raanana	2006	4.72%	1.56%	0.47%	27.82	0.30	6437.40	0.09%	9269.74	68.40%	0.3848	6786.28
Sderot	2006	4.13%	15.51%	0.64%	21.61	0.32	7718.35	0.09%	4135.18	30.30%	0.2074	7083.78
Shefaram	2006	0.97%	19.29%	0.33%	30.99	0.24	5364.56	0.13%	4233.62	40.60%	0.2368	5057.73
Tel-Aviv	2006	5.52%	5.24%	0.97%	24.20	0.48	9785.80	0.12%	7208.46	55.70%	0.5957	9799.20

Table 2. Score and Malmquist productivity change index

City	Cross Efficiency					Malmquist productivity		
	2008	2007	2006	Mean	Rank	2006-2007	2007-2008	Mean
Umm el-Faheim	0.7547	0.7442	0.7667	0.755	54	0.777	0.561	0.660
OFAKIM	0.7228	0.7201	0.7491	0.731	63	0.723	0.570	0.642

City	Cross Efficiency					Malmquist productivity		
	2008	2007	2006	Mean	Rank	2006-2007	2007-2008	Mean
Or Yehuda	0.8159	0.7957	0.7897	0.800	31	0.738	0.615	0.674
Or Akiva	0.7936	0.7644	0.7603	0.773	47	0.742	0.587	0.660
Eilat	0.7003	0.8116	0.793	0.768	49	0.631	0.546	0.587
Ariel	0.7957	0.8402	0.7509	0.796	35	0.724	0.528	0.618
Ashdod	0.8472	0.9064	0.7849	0.846	17	0.852	0.534	0.675
Ashkelon	0.8466	0.8264	0.7591	0.811	27	0.803	0.584	0.685
BakaJat	0.8778	0.7549	0.7564	0.796	33	0.406	0.644	0.511
Beer-Sheva	0.7977	0.8083	0.7616	0.789	39	0.752	0.548	0.642
Beit-Shean	0.6782	0.6852	0.717	0.693	68	0.696	0.504	0.592
Beit-Shemesh	0.8029	0.7389	0.7532	0.765	50	0.781	0.710	0.745
Beitar-Illit	0.6603	0.542	0.7139	0.639	69	1.018	0.669	0.826
Bney-Brak	0.8154	0.8219	0.7441	0.794	36	0.824	0.668	0.742
Bat-Yam	0.769	0.7696	0.7162	0.752	55	0.794	0.580	0.679
Givatayim	0.8018	0.811	0.7614	0.791	37	0.921	0.681	0.792
Dimona	0.781	0.7421	0.7246	0.749	56	0.701	0.572	0.633
Hod-Hasharon	0.9233	0.8963	0.8867	0.902	8	0.849	0.644	0.740
Herzliya	0.8883	0.9334	0.913	0.912	5	0.833	0.615	0.716
Hadera	0.8083	0.8322	0.7795	0.807	29	0.775	0.555	0.656
Holon	0.8495	0.8551	0.8034	0.836	20	0.833	0.611	0.713
Haifa	0.8436	0.8637	0.8707	0.859	16	0.828	0.613	0.712
Tiberias	0.6754	0.7921	0.7588	0.742	60	0.785	0.515	0.636
Taibe	0.7184	0.4331	0.7628	0.638	70	0.552	0.844	0.682
Tierra	0.8151	0.7819	0.73	0.776	45	0.944	0.608	0.758
Tirat-Carmel	0.8799	0.9465	0.8279	0.885	11	0.781	0.475	0.609
Tamra	0.782	0.7622	0.7889	0.778	44	0.869	0.573	0.705
Yavne	0.915	0.9477	0.8557	0.906	6	0.784	0.592	0.681
Yehud	0.7135	0.8669	0.792	0.791	38	0.847	0.642	0.738
Jerusalem	0.7832	0.7935	0.7899	0.789	40	0.779	0.606	0.687
Kfar-Saba	0.8421	0.893	0.9577	0.898	10	0.914	0.679	0.788
Karmiel	0.8669	0.8245	0.8146	0.835	21	0.802	0.581	0.682
Lod	0.8172	0.811	0.7231	0.784	41	0.778	0.629	0.699
Migdal-Haemeq	0.7331	0.8169	0.7389	0.763	52	0.792	0.526	0.646
Modiin	0.7887	0.9351	0.9014	0.875	13	1.063	0.556	0.769
Maale-Adummim	0.8798	0.851	0.9015	0.877	12	0.815	0.667	0.737
Ma'alot Tarshiha	0.823	0.8187	0.8017	0.814	25	0.755	0.598	0.672
Nahariya	0.829	0.7635	0.8065	0.800	32	0.815	0.653	0.730
Ness-Ziona	0.9108	0.9168	0.9268	0.918	3	0.843	0.603	0.713
Nazareth	0.7881	0.7453	0.6936	0.742	59	0.852	0.650	0.744
Nazareth-Illit	0.829	0.7452	0.8145	0.796	34	0.739	0.647	0.692
Nesher	0.8327	0.8329	0.8452	0.837	19	0.832	0.616	0.716
Netivot	0.7583	0.734	0.7855	0.759	53	0.668	0.570	0.617
Netanya	0.8271	0.833	0.7818	0.814	26	0.787	0.554	0.661
Skhnen	0.686	0.7354	0.808	0.743	58	0.794	0.573	0.675
Akko	0.7486	0.7878	0.7028	0.746	57	0.750	0.487	0.604

City	Cross Efficiency					Malmquist productivity		
	2008	2007	2006	Mean	Rank	2006-2007	2007-2008	Mean
Afula	0.7735	0.7568	0.7951	0.775	46	0.770	0.622	0.692
Arad	0.8106	0.7346	0.6671	0.737	62	0.783	0.624	0.699
Petah-Tikva	0.9157	0.9086	0.8788	0.901	9	0.846	0.653	0.743
Zfat	0.7006	0.6958	0.7388	0.712	66	0.659	0.571	0.613
Qalansuwa	0.7304	0.7151	0.7688	0.738	61	0.949	0.667	0.796
Kiryat- Ono	0.8552	0.8515	0.8136	0.840	18	0.823	0.645	0.728
kiryat-ata	0.8318	0.8541	0.7754	0.820	23	0.785	0.528	0.644
Qiryat-Bialik	0.8093	0.769	0.7334	0.771	48	0.786	0.637	0.707
Qiryat-Gat	0.834	0.828	0.7431	0.802	30	0.707	0.527	0.610
Qiryat-Yam	0.7454	0.7074	0.6776	0.710	67	0.672	0.565	0.616
Kiriat-Motzkin	0.8299	0.8101	0.7097	0.783	42	0.848	0.608	0.718
kiryat-malchy	0.6894	0.7275	0.7591	0.725	65	0.630	0.533	0.580
kiryat-Shmona	0.7394	0.7617	0.7934	0.765	51	0.809	0.630	0.714
Rosh-Haayin	0.8454	0.9006	0.8781	0.875	14	0.792	0.601	0.690
Rishon Lezion	0.9341	0.9419	0.8739	0.917	4	0.847	0.611	0.720
Rahat	0.7923	0.8102	0.8267	0.810	28	0.674	0.662	0.668
Rehovot	0.8574	0.7841	0.812	0.818	24	0.741	0.652	0.695
Ramla	0.7926	0.6538	0.7399	0.729	64	0.672	0.546	0.606
Ramat-Gan	0.9114	0.8701	0.8126	0.865	15	0.822	0.649	0.730
Ramat-Hasharon	0.9001	0.9564	0.9233	0.927	2	0.779	0.601	0.684
Raanana	0.9631	0.9093	0.9432	0.939	1	0.791	0.612	0.695
Sderot	0.7598	0.8104	0.7684	0.780	43	0.786	0.469	0.607
Shefaram	0.8488	0.8041	0.8143	0.822	22	0.793	0.639	0.712
Tel-Aviv	0.8686	0.9401	0.9029	0.904	7	0.820	0.616	0.711
Mean	0.808	0.805	0.793	0.802		0.778	0.598	0.682

- Note that all Malmquist index averages are geometric means.
- The results of Malmquist index from DEAP version 2.1

Table 3. 10 top ranked Israeli cities

City	Cross Efficiency					Malmquist productivity		
	2008	2007	2006	Mean	Rank	2006-2007	2007-2008	Mean*
Raanana	0.9631	0.9093	0.9432	0.939	1	0.791	0.612	0.695
Ramat-Hasharon	0.9001	0.9564	0.9233	0.927	2	0.779	0.601	0.684
Ness-Ziona	0.9108	0.9168	0.9268	0.918	3	0.843	0.603	0.713
Rishon Lezion	0.9341	0.9419	0.8739	0.917	4	0.847	0.611	0.720
Herzliya	0.8883	0.9334	0.913	0.912	5	0.833	0.615	0.716
Yavne	0.915	0.9477	0.8557	0.906	6	0.784	0.592	0.681
Tel-Aviv	0.8686	0.9401	0.9029	0.904	7	0.82	0.616	0.711
Hod-Hasharon	0.9233	0.8963	0.8867	0.902	8	0.849	0.644	0.740
Petah-Tikva	0.9157	0.9086	0.8788	0.901	9	0.846	0.653	0.743

City	Cross Efficiency					Malmquist productivity		
	2008	2007	2006	Mean	Rank	2006-2007	2007-2008	Mean*
Kfar-Saba	0.8421	0.893	0.9577	0.898	10	0.914	0.679	0.788

- Note that according to the Cross Efficiency method results Raanana, Ramat-Hasharon and Ness-Ziona turned out to be ranked first. The common feature of these cities is their geographical position in the centre of the country and along the sea coast line.

Table 4. 10 bottom ranked Israeli cities

City	Cross Efficiency					Malmquist productivity		
	2008	2007	2006	Mean	Rank	2006-2007	2007-2008	Mean*
Qalansuwa	0.7304	0.7151	0.7688	0.738	61	0.949	0.667	0.796
Arad	0.8106	0.7346	0.6671	0.737	62	0.783	0.624	0.699
Ofakim	0.7228	0.7201	0.7491	0.731	63	0.723	0.57	0.642
Ramla	0.7926	0.6538	0.7399	0.729	64	0.672	0.546	0.606
Kiryat-Mal'achy	0.6894	0.7275	0.7591	0.725	65	0.63	0.533	0.580
Zfat	0.7006	0.6958	0.7388	0.712	66	0.659	0.571	0.613
Qiryat-Yam	0.7454	0.7074	0.6776	0.71	67	0.672	0.565	0.616
Beit-Shean	0.6782	0.6852	0.717	0.693	68	0.696	0.504	0.592
Beitar-Illit	0.6603	0.542	0.7139	0.639	69	1.018	0.669	0.826
Taibe	0.7184	0.4331	0.7628	0.638	70	0.552	0.844	0.682

- Note that bottom ranked cities listed in Table 4 belong generally to the northern and southern outskirts of the country.

5. Conclusions and Future Research

In this paper we intended to demonstrate an important application area of the DEA methodology enabling relative effectively assessment of DMUs in conjunction with the CE method to carry out fully ranking for the same, all this compared with the Malmquist productivity index capable of evaluating relative improvement of each DMU per every pair of consecutive time periods. These methods have been applied to evaluating 70 Israeli cities within years 2006, 2007 and 2008. The results obtained which have been reported in the present study may lead to the following conclusions:

- No correlation whatsoever has been identified between the ranking position of the city and its relative improvement through years. As a matter of fact, the majority of cities investigated in our research show actually worsening rather than improvement, no matter whether being ranked at the top of the list or close to the bottom.
- Ranking positions obtained as well as improvement rates refer of cause to the set of input / output criteria chosen for the research. A different choice of input / output criteria might cause other results, accordingly.

- There is no consistency within the ranking obtained on the basis of existing ranking methods. The substance of the ranking method is important for obtaining specific ranking results whatsoever.
- No correlation whatsoever has been identified between the size of the city (number of inhabitants) and its ranking / relative improvement through years. In other words, one cannot claim those features to be size dependent.

Ranking and improvement as reflected in our research relate to a broad spectrum of instances such as: education, health care, the local authority's municipal spending, etc. We suggest further research to be undertaken to estimate ranking of cities according to each criterion separately, while calculating relative weights for every criterion chosen. Then, efficiency score and ranking position for every participating city might be re-calculated with reference to the established weights. In addition, further investigation has to be undertaken as to specific reasons for cities' productivity worsening as detected in the framework of the current research.

References

1. Adler, N., Friedman, L. and Sinuany-Stern, Z. **Review of ranking methods in the DEA context**, European Journal of Operational Research, 140, 2002, pp. 249-265
2. Anderson, P. and Peterson, N.C. **A procedure for ranking efficient units in DEA**, Management Science, 39, 10, 1993, pp. 1261-1264
3. Banker R.D., Charnes, A. and Cooper, W.W. **Some models for scale inefficiencies in DEA analysis**, Management Science, 30, 9, 1984, pp. 1078-1092
4. Banker, R.D. and Chang, H. **A simulation study of hypothesis tests for differences in efficiencies**, International Journal of Production Economics, 39, 1995, pp. 37-54
5. Barros, C.P. **Productivity growth in the Lisbon police force**, Article Public Organization Review, 6, 1, 2006, pp. 21-35
6. Caves, D.W., Christensen, L.R. and Diewert, W.E. **Multilateral comparison of output, input, and productivity using superlative index numbers**, The Economic Journal, 92, 1982, pp. 73-86
7. Coelli, T.A. **Guide to DEAP, version 2.1: A data envelopment analysis (Computer) program**, Center for Efficiency and Productivity Analysis, University of New England, 1996
8. Cooper, W.W. and Tone, K. **Measures of inefficiency in DEA and stochastic frontier estimation**, European Journal of Operational Research, 99, 1997, pp. 72-88
9. Doyle, J.R. and Green, R.H. **Efficiency and cross-efficiency in DEA: Derivatives meaning and uses**, Journal of Operational Research Society, 45, 5, 1994, pp. 567-578
10. Fare, R., Grosskopf, S. and Lovell, C.A.K. **The Measurement of Efficiency of Production**, Kluwer-Nijhoff Publ., Boston, MA, 1985
11. Fare, R., Grosskopf, S., Norris, V. and Zhang, Z. **Productivity growth, technical progress, and efficiency change in industrialized countries**, American Economic Review, 84, 1994, pp. 66-82
12. Friedman, L. and Sinuany-Stern, Z. **Scaling units via the canonical correlation analysis in the DEA context**, European Journal of Operational Research, 100, 3, 1997, pp. 629-637
13. Friedman, L. and Sinuany-Stern, Z. **Combining ranking scales and selecting variables in the DEA context: The case of industrial branches**, Computers and Operational Research, 25, 9, 1998, pp. 781-791
14. Ganely, J.A. and Cubbin, S.A. **Public Sector Efficiency Measurement: Applications of Data**

Envelopment Analysis, North-Holland, 1992

15. Hadad, Y., Friedman, L. and Hanani, M.Z. **Measuring efficiency of restaurants using the Data Envelopment Analysis methodology**, Computer Modelling and New Technologies, 11, 4, 2007, pp. 25-36
16. Hadad, Y., Friedman, L. and Israeli, A. **Evaluating hotel advertisements efficiency using DEA**, Journal of Business Economics and Management, 5, 3, 2004, pp. 133-141
17. Hadad, Y., Friedman, L. and Sinuany-Stern, Z. **Ranking methods for units in the DEA context: A case study of fish farms**, Communications in Dependability and Quality Management, 7, 1, 2004, pp. 63-77
18. Hadad, Y., Ben-Yair, A. and Friedman, L. **Comparative efficiency assessment and ranking of public defence authority in Israel**, Technological and Economic Development of Economy, 13, 3, 2009, pp. 27-34
19. Malul, M., Hadad, Y. and Ben-Yair, A. **Measuring and ranking economic and social efficiency of countries**, International Journal of Social Economics, 36, 8, 2009, pp. 832-843
20. Seiford, L.M. **The evolution of the state-of-art (1978-1995)**, Journal of Productivity Analysis, 7, 1996, pp. 99-137
21. Sherman, H.D. **Improving the productivity of service businesses**, Sloan Management Review, 23, 3, 1984, pp. 11-23
22. Silkman, R.H. and Hogan, A.J. **DEA critique and extension**, in: Measuring Efficiency: an Assessment of DEA / R.H. Silkman (ed.), San-Francisco: Jossey-Bass, 1986, pp. 73-104
23. Sinuany-Stern, Z., Mehrez, A. and Barboy, A. **Academic departments' efficiency via DEA**, Computer and O.R., 21, 5, 1994, pp. 543-556
24. Sinuany-Stern, Z. Mehrez, A. and Hadad Y. **An AHP/DEA methodology for ranking Decision-Making Units**, International Transactions in Operational Research, IFORS, 7, 2000, pp. 109-124
25. Sueyoshi, T. **Measuring the industrial performance of Chinese cities by DEA**, Socio-Econ. Plan. Sci., 26, 2, 1992, pp. 75-88.
26. Sueyoshi, T. **Tariff structure of Japanese electric power companies: An empirical analysis using DEA**, European Journal of Operational Research, 118, 2, 1997, pp. 350-374
27. Sueyoshi, T. **DEA non-parametric ranking test and index measurement: Slack-adjusted DEA and an application to Japanese agriculture cooperatives**, Omega, 27, 1999, pp. 315-326
28. Thompson R.G., Langemeier, L.N., Lee, C.-T., Lee, E. and Thrall, R.M. **The role of multiplier bounds in efficiency analysis with application to Kansas farming**, Journal of Econometrics, 46, 2, 1990, pp. 93-108
29. Trout, M.D. **Derivation of the maximum efficiency ratio model from the maximum decisional efficiency principle**, Annals of Operations Research, 73, 1997, pp. 323-338

JOURNAL OF APPLIED QUANTITATIVE METHODS

International Symposium on Stochastic Models in Reliability Engineering, Life Sciences and Operations Management (SMRLO'10)

AESTETICS, USEFULNESS AND PERFORMANCE IN USER-SEARCH – ENGINE INTERACTION[1]

Adi KATZ

PhD, Industrial Engineering and Management Department,
SCE - Shamoon College of Engineering,
Beer-Sheva, Israel

E-mail: adis@sce.ac.il

Abstract: *Issues of visual appeal have become an integral part of designing interactive systems. Interface aesthetics may form users' attitudes towards computer applications and information technology. Aesthetics can affect user satisfaction, and influence their willingness to buy or adopt a system. This study follows previous studies that found that users associate aesthetics with other system attributes, e.g. usability. In this study, we asked whether the well-known phenomenon that beautiful things are perceived as good applies to the perception of the system's usefulness. A controlled laboratory experiment tested the relationships between users' perception of aesthetics, usefulness and user performance in tasks performed by participant using an interactive application that surrogated a search engine. We measured users' perceptions of the search engine before and after they used the system to solve information-seeking tasks, and measured user task performance. As expected, significant correlations were found between perceived aesthetics and perceptions of usability and usefulness prior to actual use of the system. We did not find a relation between perceived aesthetics and usefulness after use; and we did not find an expected effect for aesthetic perceptions neither on perceived usefulness nor on performance. We conclude that there is need for a deeper understanding of aesthetic perceptions; a finer grain perspective of perceived aesthetics that differentiates between aesthetic dimensions may reveal that some aesthetic aspects have greater influence on the relations between aesthetics and usefulness.*

Key words: *usefulness; aesthetics; usability; search engines; human computer interactions; interface design*

1. Introduction

The tension between *function* and *form* has long been at the crossroad of artifact design. While emphasis on function stresses the importance of the artifact's usability and usefulness, accentuating the artifact's form serves more the aesthetic and perhaps the social and emotional needs of designers and users (Tractinsky et al., 2000). Today, more and more researchers and interface designers place emphasis on aspects such as aesthetics or promotion of pleasure, and are involved with seeking for opportunities for positive

experiences like pleasure, fun and excitement. There is a remarkable interest in human computer interactions (HCI) to design positive experiences for the user. Designing a good user experience is important not only when designing systems for play and leisure but also for systems that we use for achieving tasks with a well-defined goal. Search engines are a type of such systems. A search engine is an information retrieval system designed to help find information. Its environment enables us to test the relationships between aesthetics and usefulness, because searching for information is considered a task with a well-defined goal that involves decision-making and cognitive effort.

It is important to design positive experiences with systems that we use, because the emotional system changes the way in which the cognitive system operates: emotions change the way the human mind solves problems, and aesthetics can change our emotional state (Norman, 2004). Aesthetics may form user's attitudes towards the system, may improve (or worsen) their performances, affect their satisfaction, and influence their willingness to buy or adopt the system (Tractinsky, 2004).

The main goal of this study is to test whether previously found relations between perceived aesthetics and usability reflect a more general tendency to associate aesthetics with other system attributes. This study focuses on the potential relations between perceived aesthetics and perceived usefulness. In addition, we test whether aesthetics affect performance and user satisfaction. The context of this study is users interacting with a search engine.

The rest of this paper is structured as follows: In the Theory section, we summarize previous studies related to aesthetics of interactive systems and present our propositions. We then refer to usefulness dimensions that are relevant when users evaluate their interaction with search engines. The Method section describes the experimental participants; the apparatus that we designed for the experiment; the experimental design; manipulations, tasks, procedure and the dependent variables' measurements. In the following section, we present and discuss our results and findings. The last section raises the limitations of the current study, its conclusions, and proposes ideas for future work.

2. Theory

2.1. Aesthetics and Positive Experiences in HCI

MIS and HCI have traditionally ignored matters of aesthetics, and whenever aesthetic issues were discussed in the literature and in HCI textbooks, designers were warned against its potential detrimental effects on performance, comprehension, attention and other task-oriented aspects of the interaction. In that perspective, Skog et al. treated aesthetics and utility in as two conflicting concerns that must be reconciled for creating truly useful ambient information visualizations: Visualizations must strike a balance between aesthetical appeal and usefulness (Skog, et al., 2003). Floris claimed that one has to be aware of the possible opposition of utility and attractiveness. There is need for a sensible choice to be made for the relative strengths of the information bearing and the aesthetic factors - including a 'strength zero' of the latter, if need be (Floris, 2008). Lavie and

Tractinsky (2004) discuss the marginalization treatment that the aesthetics dimension receives in the human-computer interaction literature.

The claim in the mid nineties, however, was that modern design places too much emphasis on aspects of performance but not enough emphasis on aspects such as aesthetics or promotion of pleasure. Lavie & Tractinsky claim that any random perusal of websites would suggest that aesthetic considerations are paramount in designing for the web, and report that a new aesthetic wave is increasingly considering aesthetic aspects in human computer interaction. Issues of visual appeal and aesthetics have become an integral part of interactive system design and of information technology (Lavie & Tractinsky, 2004). According to Tractinsky, aesthetics satisfies basic human needs and aesthetic considerations are becoming increasingly important in our society. Today, more and more research and practical design are involved with seeking for opportunities for positive experiences like pleasure, fun and excitement (Tractinsky, 2004). There is a growing recognition of the role of emotional design of everyday things (Norman, 2004) and of information technology systems. IT users need human computer interactions that are complete and satisfying; they deserve an experience that not only achieves task-oriented goals (like efficiency and effectiveness) but also involves the senses and generates positive affective responses (Venkatesh & Brown, 2001).

User experience (UX), a relatively new realm of research in human computer interface design, emphasizes the users' overall satisfaction and experience with a product or a system. While past activities pretty much focused on avoiding negative experiences and on ways in which information technology should be designed to meet user needs for better task performance in terms of efficiency and effectiveness, designing good experiences for users now occupies the HCI community. As the functionality of new information technology products exceed user's needs, and as the prices of systems decrease, the differentiation between products are in terms of UX enhancing rather than on improving functionality (Norman, 1998). One of the various ways to enhance UX is emphasizing aesthetics.

2.1.1. The Positive Effects of Aesthetics

Recently, findings and theories indicate that human decision-making does not rely only on cognitive processes, but also on the affective state (Tractinsky, 2004; Norman, 2004). The emotional system changes how the cognitive system operates: emotions change the way the human mind solves problems, and aesthetics can change our emotional state (Norman, 2004). Affect changes how well we do cognitive tasks: affect regulates how we solve problems and perform. Negative affect can make it harder to do even easy tasks, while positive affect can make it easier to do difficult tasks (Norman, 2002). Following this idea, this research will test whether aesthetic interfaces affect performance in the context of users interacting with a search engine.

Proposition 1: *Users' aesthetic perceptions of a system have an affect on their performance in the system: users perform better with search engine that they perceive as more beautiful.*

In pleasant, positive situations, people are much more likely to be tolerant of minor difficulties and irrelevancies. Although poor design is never excusable, when people are in a relaxed situation, the pleasant, pleasurable aspects of the design will make them more tolerant of difficulties and problems in the interface (Norman, 2002). Following this logic, we

expected that users would perceive beautiful search engines as useful and satisfying, even when the usefulness of the search engine is low.

Proposition 2: *Users' aesthetic perceptions of a system have an affect on their satisfaction with the system: users are more satisfied with search engine that they perceive as more beautiful.*

2.2. Usefulness, Usability and Aesthetics

Usefulness is the issue of whether the system can be used to achieve some desired goal. *Perceived Usefulness* is the degree to which a person believes that using a particular system would enhance his or her job performance (Davis, 1989, p. 320). People form perceived usefulness judgments in part by cognitively comparing what a system is capable of doing with what they need to be done by their job. TAM2 (The extended technology acceptance model that models how users come to accept and use a technology) theorizes that people use a mental representation for assessing the match between job goals and the consequences of performing the act of using a system as a basis for forming judgments about the use-performance contingency. One key component of the matching process is the user's cognitive judgment of job relevance that exerts a direct effect on perceived usefulness (Venkatesh and Davis, 2000).

Usability is a quality attribute that assesses how easy a user interfaces is to use, and is defined by five quality components: learnability, efficiency, memorability, errors and satisfaction (Nielsen, 1993). *Perceived Usability* is the degree to which a person believes that using a particular system would be free of physical and mental effort (Davis, 1993, p. 477).

2.2.1. Relations of Aesthetics with Usability and Usefulness

It was found that aesthetics are highly correlated with perceptions of the system's usability before (Tractinsky, 1997) and after (Tractinsky et al, 2000) the interaction. Aesthetics may form user's attitudes towards the system, may improve (or worsen) their performances, affect their satisfaction, and influence their willingness to buy or adopt the system (Tractinsky, 2004). Aesthetic impressions may affect how people perceive other attributes of a system, like usability or ease of use (Tractinsky et al., 2000) and perceived goodness of a system (Hassenzahl, 2004). Lavie and Tractinsky (2004) found a relationship between the aesthetics factor and the perceived service quality of a web site, and say that it is possible that aesthetics is the primal factor affecting other perceptions.

Proposition 3: *Aesthetic perceptions of systems are related to usability perceptions. A search engines that is perceived as beautiful is also perceived as usable.*

In the ancient world, judgments of a product's usefulness and beauty were one of the same (Lavie and Tractinsky, 2004). However, correlations between aesthetics and perceived usefulness had not been investigated experimentally. One of the goals of this study is to test whether previously found correlations between perceived aesthetics and usability reflect a more general tendency to associate aesthetics with other system attributes. Perhaps a halo effect may cause carry over of an aesthetic design to perceptions of other design features (Tractinsky et al., 2000). We focus on the potential relation between

perceived aesthetics and perceived usefulness and wish to find whether the well-known phenomenon that beautiful things are perceived as good applies to the perception of system's usefulness.

Proposition 4: *Aesthetic perceptions of systems are related to usefulness perceptions. A search engines that is perceived as beautiful is also perceived as useful.*

2.3. Usefulness of Search Engines

A search engine should allow users to compose their own search queries rather than simply follow pre-specified search paths or hierarchy as in the case of certain catalogues (Chu & Rosenthal, 1996).

2.3.1. Content Relevancy

The main purpose of a search engines is to retrieve the relevant documents for a given request. It is therefore natural that the literature on search engines and retrieval systems has to a large degree concentrated on relevance-oriented questions, i.e., on the relevance and precision of the retrieved results. *Content relevance* is defined as the adequacy of the content of a document in response to the request. Subjective relevance is defined as the usefulness of the document to the user (Bing & Harvold, 1977).

2.3.2. Informative Results

In addition to content relevancy, another aspect of a search engine's usefulness is the degree to which search results are informative. A SERP (Search Engine Results Page) listing contains a list of links to web pages along with a short summary of the pages. Those pages include content that matches the search terms. The usefulness of a web page's description varies on the extent that it conveys helpful information, ranging from descriptions that reveal the answer to the research question (most informative) through descriptions that reveal the content of the web site they represent (informative), to descriptions appearing in gibberish (uninformative). According to Kowalski, there is a likely possibility that there will be items found by the query that are not retrieved by the user for review (Kowalski, 1997). Users will not review items in the SERP listing when the summary of information in the display is sufficient to judge that the item is irrelevant. Usefulness is higher on one hand whenever users are able to avoid accessing into fruitless pages, and on the other hand, when they are able to access into useful pages on the outset. Informative results are in line with Lancaster and Fayen's (1973) form of output dimension that refers to the various formats in which the documents and feedback indicators may be presented to the user and with TAM2's output quality notion, a determinant of perceived usefulness (Venkatesh and Davis, 2000).

In the following section, we describe in detail a laboratory experiment that we conducted to test the propositions. We will also refer to content relevancy and to informative results when we describe the usefulness manipulation.

3. Method

The above propositions were tested in a laboratory experiment. Participants interacted with a computer application that served as a surrogate for a search engine, to find answers for search tasks. Two variables were manipulated: *aesthetics* of the search engine's screen layout, and *usefulness* of the search results. We manipulated aesthetics by allocating subjects to work with a system at a certain level of aesthetics, based on their prior evaluation of the beauty of different screen layouts. Usefulness was manipulated based on two dimensions: the relevancy of the results to the question in task, and the brief summary information conveyed by the site's link in the SERP listings. Below we describe the experimental participants; the participants; the apparatus that we designed for the experiment to surrogated a search engine; the experimental design; the manipulations, tasks, procedure, and the dependent variables' measurements.

3.1. Participants

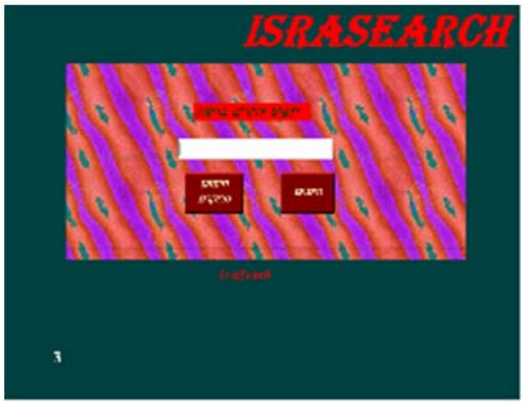
Sixty Israeli undergraduate students from a College of Engineering participated in the experiment, all of them in their third year, and all of them specializing in Information Systems. They received class credit for their participation as part of their "Human Computer Interactions" course. In addition, they were aware of the possibility that the three top performers in the experiment might receive monetary prize. There were 47 males and 17 females, and their ages ranged from 20 to 34 years, with an average age of 26.68. Sixty seven percent of them use search engines frequently (very often or every day) and the rest are familiar with search engines but use them only occasionally.

3.2. Apparatus

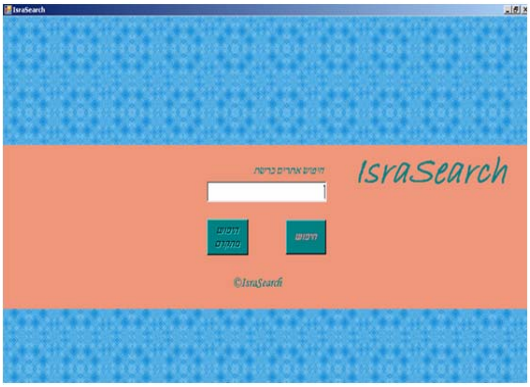
For the research, we built a computer application named IsraSearch, surrogating a search engine. The reason that we did not use a real engine was to ensure experimental control over certain variables that we did not manipulate but might bring potential noise to the experiment (such as the number of links in the SERP listings).

Figure 1 presents the opening page of six interface layouts that were used in the experiment. Israsearch's appearance imitated a real search engine such as Google. There were six different IsraSearch interfaces, identical in their controls and displayed elements. The only difference between them was their aesthetic, in terms of their "skin", i.e., colors of the elements, background textures, font styles, and locations of two captions. The choice of interfaces was based on a pilot with 30 undergraduate students (who did not participate in the main experiment). For the pilot, we designed 32 interface layouts. Each interface included the basic search controls; a textbox for typing search terms and a "search" button, to appear like common search engines. We presented them to the students on a big screen. Afterwards each student sat in front of a personal computer screen, observed the same designs individually, and rated the interfaces

Low on aesthetics



Medium on aesthetics



High on aesthetics

Figure 1. The six IsraSearch interface designs

on a 5 point Likert scale, from "very unattractive" (1) to "very beautiful" (5). Based on these ratings, we chose six designs for the main experiment, of which two were rated as highly aesthetic, two as low in terms of aesthetics, and the remaining two received medium ratings. In Figure 1, we present the six experimental designs arranged in three rows based on their aesthetic ascription in the pilot: low, medium and high, respectively.

The display of SERP (Search Engine Results Page) listings too resembled the format of other search engines, such as the changing colors of visited links, a headline and a short summary describing the web page to be accessed by each link. An example for a SERP listing is presented in Figure 2. The interface and search language were Hebrew.

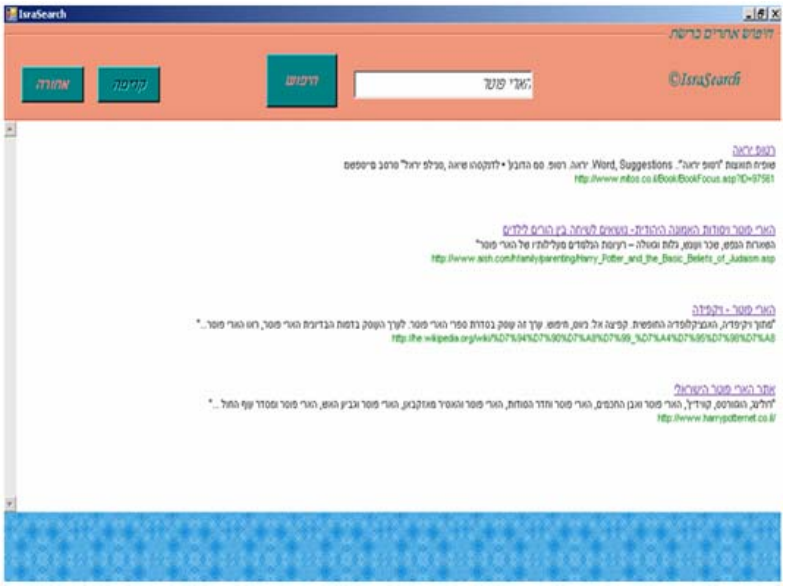


Figure 2. Example of an IsraSearch SERP listing

Being a surrogate for a real search engine, IsraSearch does not have an indexing algorithm. Instead, the pre-selected SERP lists were displayed only if the user's search statement contained at least a minimum set of predefined search terms. The minimum set of predefined search terms was determined for each task in advance, based on the results of a survey we conducted with 10 subjects who did not participate in the experiment. The survey participants were presented with 10 experimental tasks (queries). As each query defines the formal properties that a document must have in order to be retrieved (Bing & Harvold, 1977), we asked them to write a list of appropriate search terms they would type in order to retrieve pages that will help them solve the task. We summed the various terms produces by our participants for each task to a list, and from each list we chose a minimum set of terms that were most frequent (these terms represent the "core" properties that a document must have in order to be retrieved). At the experiment, search results for each query were actually web pages that were selected in advance and were presented to the user whenever he entered the minimum predefined set of search terms. Of course, experimental participants were not informed that the search engine they use is only a surrogate for a real engine.

3.3. Experimental Design

The experiment used a 3 (between) X 3 (within) factorial design. The between groups factor was the aesthetic level of the interface and the within-subjects factor was the usefulness of the search results. Both factors had three levels: low, medium and high. We now explain the manipulations of the aesthetic and the usefulness factor.

3.4. Manipulations

3.4.1. Aesthetics Factor

To obtain three levels of aesthetic interfaces, we randomly allocated subjects to three different experimental conditions. When participants arrived, we first asked each to "blindly" draw a scrap of paper with a system login. Each drawn login was assigned in advance to one of three aesthetic conditions: low, medium or high. Then, each participant sat in front of a personal computer screen, and was exposed to the six IsraSearch interface designs (shown in Figure 2) separately. At each appearance of a design, participants were asked to rate it on a 1-5 Likert scale with regard to three attributes: Aesthetics, Usability and Usefulness. Therefore, each participant made 18 ratings at this phase. We randomized the appearance of designs and rating items. After the rating phase, each subject was assigned to interact with only one of the six screen layouts based on two determinants: 1) the login he raffled at the beginning; and 2) his own ratings of the designs. For example, a participant who drew a login that was assigned to the low aesthetics condition worked with a design that he rated as least aesthetic.

3.4.2. Usefulness Factor

Each participant had to perform 10 search tasks, in each the participant had to answer 10 questions by conducting searches in IsraSearch. To obtain three levels of usefulness, we manipulated two attributes of the search results in an additive way: 1) content relevancy of the results; 2) the degree to which the results were informative. We created a mixture of the two attributes to differentiate between three groups of tasks that ranged from high usefulness to low usefulness. For the 10 searches, the experimental software returned 4 result sets with high usefulness results, 4 result sets with low usefulness results, and 2 sets with medium usefulness results. The ten search tasks are presented in Table 1, grouped in terms of their level of usefulness (high, medium and low). We explain the mixture of content relevancy and the degree to which the results were informative below.

Table 1. The ten search tasks

Usefulness	Task #	Task
high	1	**Which books did Harlen Kuben write?**
	2	**What are isobars?**
	3	**Alexander Mokdon was the son of:**
	4	**What is the height of the first floor at the Eifel tower?**
medium	5	**A poodle dog has several possible sizes. What is the size (height and weight) of the medium poodle?**
	6	**The K1200S engine of a BMW has:**
low	7	**What is the origin of the walnut?**
	8	**Surami is a kind of food that comes from:**
	9	**The turmeric's medical qualities are:**
	10	**The last movie in which Ingrid Bergman played in is:**

3.4.2.1. Content Relevancy

Relevancy has to do with whether the information appearing in a web site is sufficiently specific to answer the user's problem. The manipulation of this dimension is in line with TAM's perceived usefulness construct (Davis, 1989; Venkatesh and Davis, 2000) as previously mentioned, and with previous studies in which users estimate the relevancy of retrieved results (e.g., Shapira et al., 2005; Pan et al. (2007); Coiera and Vickland, 2008).

Content relevancy is defined as the adequacy of the content of a document in response to the request. Users' perception of the content relevancy reflects the usefulness of the document to the user. For a given document in a given situation, users will normally be able to decide whether or not the document is clearly irrelevant, or whether it might be relevant (Bing & Harvold, 1977). Only users can make valid judgments regarding the suitability of information to solve their information need (Kowalski, 1997).

As aforementioned, when a subject entered the minimal set of required search terms, a SERP listing was displayed on the screen. Each SERP listing contained four results (for uniformity and for controlling other sources of variance).

- High usefulness level was achieved by a mixture of search results that had a relatively large proportion of relevant web sites and a relatively low proportion of pages with partial or little relevancy. Highly Relevant pages are ones that contained a full answer to the question asked; partial relevancy pages contained only a portion of the answer, and low relevancy pages are related to the object in question but not to the specific question regarding that object.
- Medium usefulness level was achieved with a mixture of a relatively large proportion of pages with partial relevancy and only a small proportion of relevant or irrelevant pages. Irrelevant pages contained the object referred to in the question but had a very weak relation to that object. In other words, those pages were mainly about other topics and the object in question was only slightly mentioned in them.
- Low usefulness level was achieved by a mixture of web pages that had a relatively large proportion of irrelevant web sites and a relatively low proportion of pages with partial relevancy.

We demonstrate this using task# 5: "A poodle dog has several possible sizes. What is the size (height and weight) of the medium poodle?" A highly relevant page contained a list of all possible poodle sizes; a page with partial relevancy contained information about the height of the medium poodle but no information regarding its weight; a page with low relevancy presented poodles for adoption; and an irrelevant page was an article about a soccer couch with the title "If you want to coach, you need to be a poodle".

3.4.2.2. Web Sites Degree of Informative Results

In addition to content relevancy, the second attribute of usefulness that we manipulated was the degree to which the search results were informative. Resembling real search engines, each SERP listing of IsraSearch contained a list of web pages with titles, a link to each page, and a short summary showing where the search terms have matched content within the page (see Figure 2). The usefulness of a web page's description varied on the extent that it conveyed helpful information, ranging from descriptions that reveal the answer to the research question (highly useful; most informative), through descriptions that contain the content of the web site they represent (medium usefulness; informative), to descriptions appearing in gibberish (low usefulness; uninformative). When tailoring page summaries to achieve different usefulness levels, the mixture of 4 web site links varied in the proportion of informative and uninformative page descriptions: the proportion of informative page descriptions were increased and the proportion of uninformative page description were decreased when advancing from low to high usefulness results. The degree of informative

results was adopted as a usefulness dimension from Kowalski (1997) as described earlier in the theory, when referring to search engine's usefulness.

Table 2 summarizes the manipulation of the two independent variables: aesthetic factor (between) and the usefulness factor (within).

Table 2. Experimental factors (manipulated variables)

Variable	Explanation	Values	Point of Measurement
Aesthetics	Manipulated by assigning each subject to a certain aesthetic condition that matched his system login and his ratings of 6 screen layouts	Low, Medium, High	Pre-experiment
Usefulness	Manipulated by affecting two usefulness dimensions: content relevancy of web sites for each task, and the summary information displayed in the SERP listing.	Low, Medium, High	Pre-experiment

3.5. The Experimental Procedure

Participants worked in computer labs under the experimenters' supervision. Each of them sat separately in front of a personal computer. They first had to type their system login that they blindly drew, as described previously. The experimental session included three stages: 1) layout rating; 2) practice task; and 3) experimental tasks.

1. Layout rating: Participants rated the six IsraSearch interface designs (presented in Figure 1) on three attributes: aesthetics, usability and usefulness. We measured usability perceptions to replicate aforementioned studies that found a relationship between users' perceptions of a system's aesthetics and usability. We used aesthetic ratings before use, to examine its relation with other system attributes perceived by the user (usability and usefulness) and we also used them to manipulate the aesthetic factor as explained earlier in the aesthetics factor section.

2. Practice task: After receiving instructions about the search engine and about the task, they practiced the use of IsraSearch by performing one preliminary search task that we used for practice At this stage, participants already worked with the screen layout that they were assigned to for the experiment (as explained earlier in the aesthetics factor section).

3. Experimental tasks: After the practice stage, participants began the experimental stage, searching the IsraSearch "engine" to answer the 10 questions (see Table 1). The questions were related to various topics, and were deliberately not trivial; that is to say that participants had to use IsraSearch to find full and correct answers. Each question was presented at the bottom of the screen, one at a time, in a random order (different for each participant), along with a set of 4 possible answers, in which only one was correct. To delimit the experiment to a reasonable time range and to raise participant's motivation and arousal, we set a limit of 5 minutes for answering each question, assuming that it is sufficient for finding the answer (in usefulness groups containing the answer in the search results). Each task ended when the subject chose one answer, by clicking on one out of 4 radio buttons and submitting the answer by clicking on a "send answer" button. If the 5-minutes time limit had run out and no answer was chosen (a timer was presented at the bottom of the screen), the task was stopped and the

following task was presented. We informed participants that their goal is to achieve a maximum number of successful answers in a minimum time range.

The experimental task resembles realistic interactions with retrieval systems, in which the user determines the information he needs and creates a search statement. The system processes the search statement, returning potential hits displayed in SERP listings. Resembling real search engines, the summaries of web pages in the SERP listing are descriptions varying from most informative (helpful in revealing the answer to the research question) to uninformative descriptions (appearing in gibberish). Then, the user selects items from the list to review and access. Resembling real search engines, the content of selected web pages varied in relevancy in terms of the adequacy of the content in response to the request.

3.6 Experimental Dependent Measures

We measured five dependent variables: perceptions of usefulness, usability, and aesthetics, user satisfaction and performance. Table 3 summarizes the dependent variables by this order. All dependent measures except for performance used five-point Likert scale items.

At the first stage, previously referred as the *layout rating stage*, Participants rated the six IsraSearch interface designs (presented in Figure 1) on a 1-5 Likert scale with regard to three attributes: aesthetics, usability and usefulness. We measured usability perceptions to replicate aforementioned studies that found a relationship between users' perceptions of a system's aesthetics and usability.

At the *experimental stage*, upon completion of each experimental task, each subject presented with four 5-point Likert-type statements that asked him/her to rate the engine on four attributes: one general usefulness question, two additional usefulness questions reflecting the two usefulness attributes we manipulated (content relevancy and degree of informative results) and the subjective satisfaction with the engine. Each statement was displayed separately, in a window that popped in the middle of the screen. At the end of the experimental stage, upon completion of all experimental tasks, an additional popup window presented again the same four Likert-type statements, this time referring to the search engine generally, that is to say beyond individual tasks.

The system's log recorded various performance measurements during the experimental stage: the number of search iterations, number of visited links (sites), time to complete each task, and the number of successful tasks. Measuring user performance by the *number of search iterations* follows Shapira et al. (2005) who measured user effort by the number of iterations required to perform a task, considering each query submitted as a single iteration. More search iterations for a given task reflects a low usefulness level because the user needs to engage in repeated searches to achieve satisfactory results for accomplishing the task. The same logic of performance efficiency was applied in two additional measures for performance: *number of visited links* and *the time it took to complete each task*. Assuming that more correct answers in a limited time range, exhibit higher performance, user performance was also measured by the overall *number of successful answers*.

4. Results

In this section, we present the results for the experiment. We start with the results of the manipulation checks, and present an examination we conducted to make sure that our participants were able to sense usefulness properly.

4.1. Manipulations Checks

4.1.1. Aesthetics Manipulation Check

As described in the Method section, aesthetics was manipulated by allocating subjects to work with a system at a certain level of aesthetics, based on their prior evaluation of the beauty of different screen layouts. A successful manipulation of aesthetics is one that produces three aesthetic groups whose perceptions of aesthetics are significantly different, each composed of participants that consider the layout they worked with at a compatible aesthetics level. A ddddddone-way analysis of variance (ANOVA) revealed a significant effect of the aesthetic factor: $F(2, 57) = 83.98$, $p < .001$. Mean ratings of IsraSearch's aesthetics were 3.53, SD = 0.125; M = 3.00, SD = 0.108; M = 1.45, SD = 0.115 for high, medium and low aesthetic conditions, respectively. Scheffe post hoc contrasts to test whether the differences between any pair of three conditions were statistically significant revealed a significant difference at 0.001 between the low and the other two conditions, and a difference at the 0.05 between the high and medium conditions. The results indicate that the aesthetics manipulation was successful. Indeed, the high aesthetic group was composed of participants who worked with a design that they ascribed as beautiful, while the medium aesthetic group was composed of participants who worked with a design that they did not consider as beautiful nor as ugly, and the low aesthetic group was composed of participants who worked with a design that they ascribed as ugly.

Table 3. Dependent variables measurements

Variable	Point of Measurement	Item
Usefulness	Rating stage - subjective valuation of usefulness based on the system's screen layout	"What is your evaluation of the system's usefulness?"
	Experimental stage, after each task and after completion of all tasks - three subjective valuation of the system's usefulness	a. "Were the web pages provided by the search engine relevant to the task?" (Relevancy dimension) b. "Was the information conveyed by the page summaries in the SERP listing helpful?" (informative results dimension) c. "Were the search results appropriate?" (General usefulness)
Usability	Rating stage - subjective valuation of the system's usability based on its screen layout	"How easy is it to use the search engine?"
Aesthetics	Rating stage - subjective valuation of the screen layout's aesthetics	"What is your evaluation of the system's aesthetics?"
User Satisfaction	Experimental stage, after each task	"Are you satisfied with the search results for this task?"
	After completion of all tasks	"Are you satisfied with the search results for the various tasks?"
User Performance	Experimental stage, after each task - objective measurements of user's achievement in each task	a. Correctness of the chosen answer (true/false). b. Number of search iterations per task, that is- the number of times a subject stimulates a search by clicking on the search button (after entering search terms). c. Number of visited links (the number of times a subject clicked on the links appearing in the SERP listing d. Time to complete each task

Variable	Point of Measurement	Item
	After completion of all tasks - objective measurements of user's overall achievement	a. Number of successful answers (maximum of 10 correct) b. Overall and average number of search iterations) c. Overall and average number of SERP links clicked d. Overall time to complete all tasks and average time to complete tasks

4.1.2. Usefulness Manipulation Check

As described earlier, we manipulated usefulness by creating for each search task a certain combination of results' relevancy in the SERP listing, and a certain degree for which the results were informative. Four tasks were characterized by result sets with high usefulness, other four were characterized by result sets with low usefulness, and two additional tasks had results sets with medium usefulness.

Using repeated measures, we tested the difference in performance between the three different levels of task usefulness in terms of time to complete the task, the number of search iterations per task, the percentage of correct answers, and the number of visited links. The results are presented in Table 4.

Table 4. Repeated measures for difference in performance between three usefulness groups

Measure	Within Subjects Effect	Means and STD		
		High usefulness	Medium usefulness	Low usefulness
Time (seconds)	F(2,58) =7.16, p < .001	85.51 (6.35)	121.45 (11. 41)	169.97 (8.84)
Number of Search iterations	F(2,58) =28.20, p < .001	7.02 (6.12)	8.27 (8.50)	12.47 (9.55)
Success (% of correct answers)	F(2,58) =130.45, p < .001	87.5% (0.20)	38.3% (0.23)	42.92% (0.21)
Number of visited links	F(2,58) =256.42, p < .001	6.22 (5.50)	13.24 (7.34)	16.15 (7.63)

For the time to complete the task and for the number of search iterations, the results are consistent with the expected pattern for each usefulness level revealing a significant within-subjects effect. For percentage of correct answers, and for the number of visited links, a significant difference was found only between high usefulness tasks and the low and medium usefulness levels. However, we found no significant difference between the medium and the low usefulness levels. This means that the usefulness manipulation did not significantly differentiate between the medium and the low usefulness levels. To deal with this finding, we checked whether certain tasks were misplaced by closely examining performance measures for each task. A following and separate examination of the performance measures for each task in IsraSearch's log revealed that Task 6 (see Table 1), ascribed as medium usefulness, was problematic, having an exceptional amount of search iterations. The fact that this was the only task that required typing keyword in a combination of English and Hebrew, might explain this. In addition, Task 5, ascribed as medium on usefulness, was actually easy, as it took very little time to complete, required only a few search iterations, and only one subject failed to find the right answer. Table 5 presents the means and the standard deviations of time to complete each task (in seconds), number of search iterations, and percentage of success for each task. Based on these results, we decided to omit Task 6, and ascribed Task 5 to the high usefulness task group. A following manipulation check referred to the two remaining usefulness groups: high versus low.

We continued with repeated measures analysis to examine the difference in performance between the two remaining usefulness levels, in terms of time to complete the task, number of search iterations per task, percentage of correct answers, and number of visited links. The results are presented in Table 6.

The results in Table 6 show that subjects perform better when usefulness is higher. Therefore, the new ascription of tasks to two levels of usefulness following the manipulation check is effective.

Table 5. Performance measures for each search task

Useful-ness	Question/Task (abbreviated titles)	Time		Search iterations		Success	
		Means	STD. Dev	Means	STD. Dev	% Correct	STD. Dev
High	1) Harlen Kuben	85.18	75.09	1.57	1.94	90	0.30
	2) Isobars	105.22	143.65	1.17	0.64	81.67	0.39
	3) Mokdon	69.10	49.31	2.77	5.61	83.33	0.38
	4) Eifel tower	82.55	51.32	1.52	1.64	95	0.22
Medium	5) Poodle dog	100.87	163.07	1.28	0.99	98.33	0.13
	6) BMW Engine	142.03	72.01	6.98	8.28	75	0.44
Low	7) Walnut	157.85	63.88	2.18	2.04	66.7	0.25
	8) Surami	187.10	80.50	3.22	2.49	30	0.46
	9) Turmeric	161.53	77.50	3.25	3.05	73.33	0.45
	10) Ingrid Bergman	173.38	160.46	3.82	5.57	61.67	0.49

Table 6. Repeated measures for difference in performance between the two remaining usefulness groups

Measure	Within Subjects Effect	Means and STD	
		High usefulness	Low usefulness
Time (seconds)	F(1,59) =75.08, p < .001	88.58 (52.74)	169.97 (68.50)
Number of search iterations	F(1,59) =75.11, p < .001	1.66 (1.24)	12.47 (9.55)
Success (% of correct answers)	F(1,59) =98.10, p < .001	70.33 (0.16)	42.92 (0.21)
Number of visited links	F(1,59) =245.82, p < .001	1.15 (0.50)	16.15 (7.63)

4.2. Verification of Participants' Usefulness Perceptions

We previously claimed that users are normally able to decide whether or not a document is clearly irrelevant, or whether it might be relevant (Bing & Harvold, 1977), and that only users can make valid judgments regarding the suitability of information to solve their information need (Kowalski, 1997). Therefore, we examined our participants' ability to have a good sense of usefulness, by looking at their usefulness ratings for each search task. As described earlier, after each task, four questions appeared in a pop-up window, referring to that task. The first three questions were about usefulness, while the fourth was about their satisfaction with the search results.

Results of a repeated-measures ANOVA for each usefulness question that popped-up, revealed a significant overall difference between tasks characterized by low versus high usefulness. Table 7 presents repeated measures for the difference in perceived usefulness and satisfaction between usefulness groups.

Table 7. Repeated measures for difference in perceived usefulness and satisfaction between usefulness groups

Measure		Within Subjects Effect	Means and STD	
			High usefulness	Low usefulness
Usefulness	Relevancy dimension	F(1,59) =237.53 p < .001	4.07 (0.65)	2.57 (0.73)
	Informative results dimension	F (1, 59) =221.29 p < .001	3.93 (0.67)	2.56 (0.80)
	Usefulness - general item	F (1, 59) =173.12 p < .001	4.09 (0.65)	2.61 (0.76)
User satisfaction		F (1, 59) =164.88 p < .001	4.16 (0.68)	2.56 (0.81)

For the first question, "Were the web pages provided by the search engine relevant to the task?" (Relevancy dimension) high usefulness tasks were rated as having significantly more relevant web pages than low usefulness tasks. Participants were able to sense usefulness properly, in other words, they were able to distinguish between results characterized by a high relevancy of content and results that were low on content relevancy.

For the question "Was the information conveyed by the page summaries in the SERP listing helpful?" (Informative results dimension), high usefulness tasks were rated as having significantly more helpful SERP listing than low usefulness tasks. Participants were able to sense usefulness in terms of the usefulness of page summaries in the SERP listings.

For the question, "Were the search results appropriate?" (a general usefulness-item), high usefulness tasks were rated as having significantly more appropriate results than low usefulness tasks. Participants were able to sense the appropriateness of the results to the task they were conducting.

Results of a repeated-measure ANOVA for satisfaction revealed a significant overall difference between tasks characterized by low versus high usefulness. For the forth question, "Are you satisfied with the search results for this task?", high usefulness tasks results were rated as more satisfying than low usefulness tasks results. Participants were satisfies when search results were informative and relevant to their task.

4.3. Propositions 1-2

4.3.1. Proposition 1

To test Proposition 1, that aesthetics of search engines will affect performance, a test of a between-subject affect for aesthetics on all performance measures was conducted using MANOVA. Table 8 shows descriptive statistics for each performance measure and results of F-tests of the aesthetic effect. For all performance measures, there was no effect of the aesthetics factor. The user's aesthetic perception of a search engine did not affect his performance in the task of information searching.

Table 8. Performance measurements in each aesthetic group

Performance measure	Aesthetic group*	Mean and Std.	F (df)
Overall number of successful answers	High	5.88 (1.22)	
	Medium	6.22 (1.62)	0.466 (2,57)
	Low	5.80 (1.58)	
Overall time to complete all tasks (m-sec)	High	1185.94 (343.44)	
	Medium	1223.96 (334.23)	0.197 (2,57)
	Low	1153.95 (416.71)	
Average time to complete tasks (m-sec)	High	127.31 (51.67)	
	Medium	135.06 (47.72)	0.905 (2,57)
	Low	115.91 (40.59)	
Overall number of search iterations	High	27.23 (15.18)	
	Medium	26.69 (14.56)	0.213 (2,57)
	Low	29.40 (12.50)	
Average number of search iterations	High	2.72 (1.52)	
	Medium	2.67 (1.46)	0.213 (2,57)
	Low	2.94 (1.25)	
Overall number of SERP links clicked	High	23.29 (9.07)	
	Medium	24.48 (7.90)	0.789 (2,57)
	Low	20.95 (10.79)	
Average number of SERP links clicked	High	2.33 (0.91)	
	Medium	2.45 (0.79)	0.789 (2,57)
	Low	2.09 (1.08)	

**Number of subjects is 17, 23 and 20, for high, medium and low aesthetic groups, respectively*

4.3.2. Proposition 2

To test Proposition 2, that aesthetics of search engines will affect user satisfaction, we tested the between-subject affect for aesthetics on post-satisfaction using ANOVA. Table 9 shows that we did not find an effect for the aesthetics factor on satisfaction as expected in Proposition 2. In other words, user's aesthetic perception of a search engine did not affect the degree of satisfaction with it.

Table 9. Post-satisfaction in each aesthetic group

Aesthetic group	Mean and Std.	F (df)
High	3.35, SD = 0. 20	
Medium	3.59, SD = 0.18	1.911 (2,56)
Low	3.10, SD = 0.17	

4.4. Propositions 3-4 and Additional Intercorrelations

4.4.1. Proposition 3

We measured perceptions of the search engine before (layout rating stage), and after (popup questions at the end of the experiment) the actual use of IsraSearch. We measured perceived usefulness, perceived usability, perceived aesthetics, and user satisfaction. Intercorrelations among the perceived aspects both before and after the experiment are presented in Table 10.

Table 10: Pearson correlation matrix of pre- and post experimental perceived measures

	Pre-Aesthetics	Pre-Usefulness	Pre-Usability	Post-Usefulness1	Post-Usefulness2	Post-Usefulness3	Post-Satisfaction
Pre-Aesthetics	-	0.610**	0.402**	0.093	0.063	0.145	0.243*
Pre-Usefulness		-	0.432**	0.203	0.150	0.108	0.205
Pre-Usability			-	0.142	0.183	-0.103	0.005
Post-Usefulness1				-	0.634**	0.523**	0.710**
Post-Usefulness2					-	0.527**	0.537**
Post-Usefulness3						-	0.598**
Post-Satisfaction							-

*** Correlation is significant at the 0.01 level (1-tailed), N=60*
Post-usefulness dimensions: 1 - relevancy; 2 - informative results; 3 - general usefulness

As proposition-3 predicted, pre-use perceptions of IsraSearch's aesthetics and their perceived usability are significantly correlated. This follows previous studies by e.g. Kurosu & Kashimora (1995) and Tractinsky et al. (2000), who found that users associate aesthetics with usability.

4.4.2. Proposition 4

Pre-use perceptions of IsraSearch's aesthetics and their perceived usefulness are significantly correlated as expected in proposition 4, and are high as the correlations between aesthetics and perceived usability obtained by Kurosu & Kashimora (1995) and Tractinsky et al. (2000).

While correlation between pre-experimental perception of aesthetics and post-experimental perceived usability were significant (0.5) in the study of Tractinsky et al. (2000), correlations between pre-experimental perceived aesthetics and post-usefulness perceptions were diminished at the end of the experiment (as shown in table 10). System layouts that were considered as aesthetic before use were not perceived as more useful after using the system. We will refer to this result in the study limitation section.

As expected in Propositions 3-4, pre-use perceptions of IsraSearch's aesthetics and pre perceptions of usability and usefulness are significantly correlated, suggesting that a halo effect causes carry over of aesthetics on other perceptions of a search engine. Prior to actual use, search engines that are perceived as beautiful are also perceived as more usable and as more useful.

4.4.3. Additional Correlations

We originally formulated two propositions (3-4) which refer to correlations between aesthetic perceptions and other users' perceptions of the search engine, and found additional correlations in our results. Table 10 shows that pre perceptions of usability and usefulness are significantly correlated. This relation was not central in the current research, but is not surprising; TAM (Technology Acceptance Theory) argues that perceived usefulness is influenced by perceived ease of use. The easier a system is to use, the more useful it can be (Venkatesh and Davis, 2000). The three post-use perceptions of usefulness are significantly inter-correlated. In addition, they are all correlated with overall satisfaction with the system with remarkably high correlation between the relevancy dimension and satisfaction. Search engines that return highly relevant web pages satisfy their users. An interesting finding is that while post-use satisfaction is uncorrelated with pre-use perceptions of usability and usefulness, it is significantly correlated with pre-use perception of aesthetics. Users were satisfied with search engines with layouts that were aesthetically pleasing but

were not necessarily satisfied with search engines with a layout that seemed usable or useful for searching information before use.

5. Discussion and Conclusions

5.1. Limitations

The relatively high correlation between pre-experimental aesthetics and pre-usefulness measure was not found for pre-experimental aesthetics and post-usefulness perceptions. System layouts that were considered as aesthetic before use were not perceived as more useful after using the system. A limitation of the current study is that aesthetic perceptions were not measured after using IsraSearch. In future studies, it would be reasonable to examine the correlations of post-experimental aesthetics with post-usefulness and post-satisfaction because aesthetic perceptions may change during and after the actual interaction with a system.

The idea that emotions change the way the human mind solves problems and aesthetics can change our emotional state (Norman, 2004), lead us to expect that aesthetics perceptions will affect usefulness perceptions, user satisfaction and performance - but we did not find this effect. An explanation for the lack of aesthetic effect is that we measured aesthetics by a single construct (see Table 3) .This measurement may not be enough to test the effect of aesthetics by a general aesthetics construct. Aesthetic perceptions are more complex and there may be certain aesthetic dimensions that are more influential than others on other system perceptions and even on performance. For example, alignment and grouping are important for rapid performance (Parush et al., 1998) but not all attractive features of graphic design improve performance (Shneiderman, 2004).

5.2. Conclusions

Norman's idea that emotions change the way the human mind solves problems and aesthetics can change our emotional state (Norman, 2004), lead to the expectation that aesthetics perceptions will affect usefulness perceptions, user satisfaction and performance - but the results show no such effect. Perhaps the aesthetic design did not have a significant effect on the user's emotional state, or it did not affect their emotional state to a point where it affects performance. Norman claims that emotions last for a relatively short periods – even minutes. It will not be unreasonable to think that users in the high aesthetic group felt good for receiving a beautiful interface, and users in the low aesthetic group felt bad for receiving an ugly interface. However, these emotions where very short and relatively weak in the context of a laboratory experiment, allowing them to quickly move their focus to the experimental demands, leaving their feelings behind.

We conclude that it is importance to drill down when investigating the notion of perceived aesthetics. Different system layouts may arouse different aesthetic dimensions. For example, a background of electric wires or chips would give a modern, professional or sophisticated look, while leaves and butterflies would give a very different feel of pleasure, harmony and beauty. Some aesthetic dimensions may influence some perceptions of the system while others may not. For example, Lavie and Tractinsky (2004) found that classical aesthetic dimensions are more closely related to perceived usability than expressive aesthetic dimensions. If the system's "skin" is judged first and creates a halo effect, then system designers should design to arouse the "right" aesthetic impressions. In other words, the

image of the system reflected by its design should fit its purpose. A search engine that is perceived as beautiful, artistic and skillfully designed can at the same time involve elements that are considered as old fashion, and therefore may be perceived as relatively low on usefulness.

We followed previous studies that researched aesthetics of interactive systems, which manipulated aesthetics in terms of changing colors of elements, background texture, font's style, and the location of captions. Subjects were able to express their aesthetic taste and state that one screen is beautiful while another is ugly, but the reasons for these statements may be most important. However, it is necessary to understand why users state that a layout is beautiful and find out whether certain dimensions of aesthetics influence other system perceptions such as usefulness. The results of this research and the results of the studies conducted by Lavie and Tractinsky (2004), show that there is a need to have a better comprehension of aesthetic perceptions of interactive systems. When Lavie and Tractinsky (2004) delved deeper towards a better understanding of the various aesthetic dimensions, their results shed new light on the already known usability-aesthetic relation: perceived usability was correlated substantially higher with the classic aesthetic dimension than with the expressive aesthetic dimension. Perhaps a deeper understanding of the various aesthetic dimensions may similarly reveal whether some dimensions have greater influence on the associations between aesthetics and usefulness. Finer grain perspectives of perceived aesthetics can also follow Hermeren's (1988) distinction between five types of aesthetic qualities: emotion, behavior, gestalt, taste and reaction.

In addition, perhaps different aesthetic dimensions are more influential on the user's general perception of a system, depending on the various contexts of use (such as user' goals and tasks, application genres, etc.). Future research that will view perceived aesthetics in a finer grain resolution and that will take the context of use into account may find that specific aesthetic dimensions have an impact on the perception of a system's usefulness. An interesting research we suggest may test which aesthetic dimensions are more influential on different cognitive processes. Perhaps some aesthetic dimensions are more influential when user's cognitive processes are characterized by automatic behavior (e.g. "freely" browsing the internet with no specific goal), and others are more important when cognitive processes are characterized by controlled behavior (e.g. searching for specific information to accomplish a task with a well-defined goal). Another research route to examine is the possibility that different dimensions dominate aesthetic perceptions of different types of end users (children versus adults, etc.). There are many fascinating paths to follow in understanding the notion of aesthetic perceptions of interactive systems, and its influence on user's attitudes and behavior with those systems.

References

1. Bing, J. and Harvold, T. **Legal Decision and Information Systems**, Oslo: Publications of the Norwegian Research Center for Computers and Law, 1977
2. Chu, H., and Rosenthal, M. **Search engines for the World Wide Web: A comparative study and evaluation methodology**, in Proceedings of the ASIS 1996 Annual Conference, October, 33, pp. 127-35, http://www.asis.org/annual-96/ElectronicProceedings/chu.html (August 19, 2003)

3. Coiera, E.W. and Vickland, V. **Is relevance relevant? User relevance ratings may not predict the impact of internet search on decision outcomes**, Journal of the American Medical Informatics Association 15, (April 2008), pp. 542-545
4. Davis, F.D. **Perceived usefulness, perceived ease of use, and user acceptance of information technology**, MIS Quarterly, 13, 3, 1989, pp. 319-339
5. Davis, F.D. **User acceptance of information technology: System characteristics, user perceptions and behavioral impacts**, International Journal on Man-Machine Studies, 38, 1993, pp. 475-487
6. Floris, van N.-L. **Aesthetics versus utility in electronic imaging**, in Proceedings of SPIE, the International Society for Optical Engineering
7. Hassenzahl, M. **The interplay of beauty, goodness and usability in interactive products**, Human-Computer Interaction, 19, 4, 2004, pp. 319-349
8. Hermeren, G. **The variety of aesthetic qualities**, in: Mitias, M.H. (ed.), „Aesthetic Quality and Aesthetic Experience", Amsterdam: Rodopi, 1988, pp. 11-23
9. Kowalski, G**. Information system evaluation**, in „Information Retrieval Systems: Theory and Implementation", 1st ed. Norwell, MA: Kluwer Academic Publishers, 1997
10. Lancaster, F.W. and Fayen, E.G. **Information Retrieval On-Line**, Los-Angeles, CA: Melville Publishing Co., 1973, Chapter 6
11. Lavie, T. and Tractinsky, N. **Assessing dimensions of perceived visual aesthetics of web sites**, Human-Computer Studies, 60, 2004, pp. 269-298
12. Nielsen, J. **Usability Engineering**, AP Professional, 1993
13. Norman, D.A. **The invisible computer: Why good products can fail, the personal computer is so complex, and information appliances are the solution**, Cambridge, MA: MIT Press, 1998
14. Norman, D.A. **Emotion & design: Attractive things work better**, Interactions of ACM, 9, 4, 2002, pp. 36-42
15. Norman, D.A. **Emotional design: Why we love (or hate) everyday things**, New-York: Basic Books, 2004
16. Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G. and Granka, L. **In Google we trust: Users' decisions on rank, position, and relevance**, Journal of Computer-Mediated Communication, 12, 3, 2007, pp. 801-823
17. Parush, A., Nadir, R., and Shtub, A. **Evaluating the layout of graphical user interface screens: Validation of the numeric computerized model**, International Journal of Human-Computer Interaction, 10, 4, 1998, pp. 343-360
18. Shapira. B, Taieb-Maimon, M. and Nemeth, Y. **Subjective and objective evaluation of interactive and automatic query expansion**, Online Information Review, 29, 4, 2005, pp. 374-390
19. Shneiderman, B. **Designing for fun: How can we design user interfaces to be more fun?**, Interactions, 11, 5, 2004, pp. 48-50
20. Skog, T., Ljungblad, S. and Holmquist, L. **Between aesthetics and utility: Designing ambient information visualiza-tions**, IEEE Symposium on Information Visualization, INFOVIS-2003, pp. 233-240
21. Tractinsky, N. **Aesthetics and apparent usability: Empirically assessing cultural and methodological issues**, CHI-97 Conference Proceedings, Atlanta: 22-27 March, 1997, ACM, New-York, pp. 115-122
22. Tractinsky, N., Katz, A.S. and Ikar, D. **What is beautiful is usable**, Interacting with Computers, 13, 2, 2000, pp. 127-145
23. Tractinsky, N. **Towards the study of aesthetics in information technology**, 25th Annual International Conference on Information Systems, Washington, DC, December 12-15, 2004, pp. 771-780

24. Venkatesh, V. and Brown, S.A. **A longitudinal investigation of personal computers in homes: Adoption determinants and emerging challenges**, MIS Quarterly, 25, 1, 2001, pp. 71-102

25. Venkatesh, V. and Davis, F.D. **A theoretical extension of the technology acceptance model: Four longitudinal field studies**, Management Science, 46, 2, 2000, pp. 186-204

[1] **Acknowledgement**

The author extends her sincere thanks to Liron Rotem for her contributions in designing IsraSearch, planning of the experimental tasks, and running the pilot and the laboratory experiment.

JOURNAL OF APPLIED QUANTITATIVE METHODS

International Symposium on Stochastic Models in Reliability Engineering, Life Sciences and Operations Management (SMRLO'10)

ONLINE HOEFFDING BOUND ALGORITHM FOR SEGMENTING TIME SERIES STREAM DATA[1]

Dima ALBERG
PhD Candidate, Department of Information Systems Engineering,
Ben-Gurion University of the Negev,
Beer-Sheva, Israel

E-mail: alberg@bgu.ac.il

Avner BEN-YAIR
PhD, Center for Reliability and Risk Management,
SCE - Shamoon College of Engineering,
Beer-Sheva, Israel

E-mail: avner2740@gmail.com

Abstract: *In this paper we introduce the ISW (Interval Sliding Window) algorithm, which is applicable to numerical time series data streams and uses as input the combined Hoeffding bound confidence level parameter rather than the maximum error threshold. The proposed algorithm has two advantages: first, it allows performance comparisons across different time series data streams without changing the algorithm settings, and second, it does not require preprocessing the original time series data stream in order to determine heuristically the reasonable error value. The proposed algorithm was implemented in two modes: off line and online. Finally, an empirical evaluation was performed on two types of time series data: stationary (normally distributed data) and non stationary (financial data).*

Key words: *data stream; time series; linear approximation; segmentation; Hoeffding bound; SWAB (Sliding Window Bottom Up) algorithm; ISW (Interval Sliding Window) algorithm*

1. Introduction

Time series data streams are ubiquitous in finance, meteorology and engineering. They are an application area of growing importance in the data stream mining research. For example, sensors generate one million samples every three minutes [13], therefore one of the primary purposes of data stream research boils down to fast and reliable time series data streams segmentation or dimensionality reduction techniques. These techniques are used in many areas of data stream mining as: frequent patterns finding, structural changes and

concept drifts detection [4], time series classification and prediction [13], time series similarities searching [8], [15], etc. The main principle of segmentation algorithms concludes in reducing the time series dimensionality by dividing the time axis into intervals behaving approximately according to a simple model. A good time series data stream segmentation algorithm has on-line, fast, accurate and comparable with other algorithms structure. For example the Sliding Window algorithm [8] on the one hand is online, very fast and relatively simple for using in online segmentation applications but on the other hand, it sometimes gives poor accuracy and does not allow to perform online multivariate segmentation.

The segmentation problem can be defined in following way: first, given a time series data stream to produce the best representation such that the maximum error for any segment does not exceed some user specified confidence level error threshold. It is important to add, that using a combined relative parameter such as Hoeffding bound [5] confidence level will allow to evaluate an online multivariate segmentation and second, to construct a user friendly segmentation application which will evaluate and compare the proposed online segmentation algorithms in real time. As we shall see in later sections, the state-of-the-art segmentation algorithms do not meet all these requirements.

The rest of the paper is organized as follows. In Section 2, we provide a literature review of three state-of-the-art online piecewise linear segmentation algorithms. In Section 3, we provide a methodology for improving the existing state-of-the-art online segmentation algorithms. The proposed methodology based on Hoeffding bound error estimation, which uses a relative probability parameter instead of maximum error nominal parameter and allows performing online multivariate segmentation. Section 4 briefly demonstrates a real-time segmentation application. Finally, in Section 5 and 6 we provide brief and meaningful empirical comparison of the proposed algorithms and suggest final conclusions.

2. Related Studies

Several high level representations of time series have been proposed in the research literature, including Fourier Transforms [8], Wavelets [1], Symbolic Mappings [3], [17] and Piecewise Linear Approximation (PLA) [20], [1], [4], [6], [5], [7], [10], [11], [14], [16], [18], [19], [21]. In this work, our attention will confine to PLA, perhaps the most frequently used representation in continuous time series data streams. Obviously, all piecewise linear segmentation algorithms can also be classified as batch or online [20]. The problem discussed by Shatkay and Zdonnik, [17], Keogh et al., [9], Biffet and Kirkby [1] is actually how to build online, adaptive, fast and accurate algorithm for piecewise linear segmentation of time series data stream, because on the one hand, the main problem of online Sliding Window algorithm [2], [7] concerns in its poor accuracy [18], [21] and its inability to look ahead. On the other hand the offline accurate Bottom Up [9] algorithm is impractical or may even be unfeasible in a data mining context, where the data are in the order of terabytes or arrive in continuous streams. This problem is very important because for scalability purposes the proposed piecewise linear segmentation algorithm needs to capture the online nature of sliding windows and yet retain the superiority of Bottom Up algorithm.

Keogh et al. [9] introduced new online Sliding Window Bottom Up (SWAB) algorithm which scales linearly with the size of the dataset, requires only constant space, produces high quality approximations of the initial time series data, and can be seen as operating on a continuum between the two extremes of Sliding Windows and Bottom-Up. The authors have shown that the most popular Sliding Window approach generally produces

very poor results, and that while the second most popular approach, Top-Down, can produce reasonable results, it does not scale well with massive time series stream data. The main problem with the Sliding Windows algorithm is its inability to look ahead, lacking the global view of its offline (batch) counterparts. The Bottom-Up and the Top-Down [7] approaches produce better results, but are offline and require the scanning of the entire data set. For example, the SWAB [9] algorithm has three nominal input parameters, which need to be defined carefully by the user in order to obtain an accurate segmentation model. Often the user obstructs to determine for the value of the maximal error threshold, because the data has very noisy non-stationary behavior. Therefore, in aim to produce an accurate segmentation model, the user needs to perform the preprocessing of the obtained data or to perform a time consuming experiment design. Second, the inner loop of the SWAB algorithm simply invokes the Bottom-Up algorithm each time. This results in some computation redundancy and increases the computational complexity of algorithm. Second, the performance of the Sliding Window and SWAB algorithms depends on the value of maximal error. As maximal error goes to zero the Sliding Window and SWAB algorithms have the same performance, since they would produce multiple short segments with no error. At the opposite end, as the maximal error becomes very large, the algorithms once again will all have the same performance, since they will simply approximate a data stream with a single best-fit line.

However, most works along these research lines that we know of [1], [19] and [9] recommend to test the relative performance for some "reasonable value" of maximal error, a value that achieves a good tradeoff between compression and fidelity. Because this "reasonable value" is subjective and dependent on the data mining application and the data itself, they did the following. First, they chose a "reasonable value" of maximal error for each dataset and then bracketed it with 6 values separated by powers of two. The lowest of these values tends to produce an over-fragmented approximation, and the highest tends to produce a very coarse approximation. Second, they chose performance in the mid-range of the 6 values which by their opinion should be considered most important. Obviously, the maximal error calculation routine proposed by and [8], [9] and [15] are heuristic, requires multi pass computational efforts and have no rigorous guarantees of performance.

We therefore, introduce an improved algorithm combining the online nature of the sliding window algorithm and time series data stream, decreasing the number of input parameters and decreasing computational redundancy and complexity. We call the proposed algorithm *ISW (Interval Sliding Window) algorithm*.

3. Methodology

3.1. The ISW Segmentation Algorithm

The proposed ISW algorithm derives the maximal error by defining appropriate confidence level (e.g. 95%) and using Hoeffding bound one pass calculation. In fact, a similar approach was used in the VFDT decision tree induction algorithm introduced in Hulten and Domingos [13]. Suppose we have segment A with range R_A and n_A observations, and segment B with range R_B and n_B observations, which belong to a sliding window S of time series T. Assume that $\bar{x}_A$ and $\bar{x}_B$ are the sample means of segments A and B, respectively. The new merged segment AB has range R_{AB}, $(n_A + n_B)$

observations and sample mean $\bar{x}_{AB}$ equals $c_A \bar{x}_A + c_B \bar{x}_B$, where c_A and c_B equal to $\frac{n_A}{n_A + n_B}$ and $\frac{n_B}{n_A + n_B}$, respectively.

Proposition 1: The Hoeffding bound states that with confidence level of δ the true mean of the merged segment AB lies in the interval $\bar{x}_{AB} \pm \varepsilon_{AB}$ where:

$$\varepsilon_{AB} = R_{AB} \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2(n_A + n_B)}} \tag{1}$$

Proof of Proposition 1: Let $X_1, X_2, ..., X_n$ be independent random variables. Assume that each x_i is bounded, that is $P(X_i \in R = [a_i, b_i]) = 1$. Let $\bar{x}_{AB} = \frac{1}{n_A + n_B} \sum_{i=1}^{n_A + n_B} x_i$, with expected value $E(X_{AB})$. Then according to Hoeffding inequality theorem [5] for any $\varepsilon > 0$

$$P\left[\bar{x}_{AB} - E[X_{AB}] > \varepsilon\right] \le e^{-2 \cdot \frac{(n_A + n_B)^2 \varepsilon^2}{R_{AB}^2}} < \delta .$$

From this theorem we can derive absolute error ε_{AB} with confidence level of δ.

According to Motwani and Raghavan [12] the Hoeffding bound ε_{AB} represented in (1) is independent of the distribution generating the examples. This bound is applicable to all situations where observations are independent and generated by a stationary distribution. Important to note, that Hoeffding bound is addictive, its error is absolute and it does not require calculation of expected means of two merged segments. It is easy to show, that when the confidence level $\delta = 1$, the Hoeffding error value is equal to zero meaning that the observed and the segmented time series data streams are the same.

The DASWI algorithm works in the following way: each time a new observation arrives the algorithm calculates the Hoeffding bound using (1) and a user defined confidence level δ then in case the new calculated error is greater than the previously calculated Hoeffding bound, the algorithm starts a new sliding window, otherwise it continues with the current sliding window. This incremental technique on the one hand is more sensitive to data stream concept drift changes and on the other hand allows to create relatively large segments when the data stream is stable and therefore to decrease significantly the running time of the proposed algorithm. The pseudocode for the ISW algorithm is shown in Figure 1.

Input:	Data stream, DS Confidence Level, δ Sliding Window size, SW
Output:	ISW Segmented data stream.

```
anchor = 1;
While not finished segmenting time series
    i = 2;
--- Bound and Error Calculation
While Model_error < Hoeffding_Bound
   Model_Error(Segment[anchor: anchor + i ])
```

```
Hoeffding_Bound(Segment[anchor], δ)
 i ++;
 New_Segment = Create_Segment(T[anchor: anchor + (i-1)]
 Segment = Merge(Segment, New_Segment)
 anchor = anchor + i;
```

Figure 1. **The ISW algorithm pseudocode**

Example 1. The following numerical example briefly explains main calculation procedure of the ISW algorithm. Suppose that the current sliding window segment includes four observations: 1, 2, 3 and 4 (Table 1). Now, a new, fifth observation arrives and the ISW algorithm checks whether to start a new segment or to continue updating the previous one. The current segment range equals 0.8 (4.8-4.0), the number of observations is 4 and with the user specified confidence level of 95% the value of Hoeffding bound equals 0.06. Now, with the aid of the new, fifth observation we will recalculate the linear interpolation model error. The new model error equals 0.025 and it is lower than the calculated Hoeffding bound therefore the algorithm increases the current segment.

Table 1. Observations for ISW numerical example

n	1	2	3	4	5 (new)
Obs. Value	4.1	4	4.4	4.8	4.7
Segment	1	1	1	1	2

4. Experimental Results

In aim to compare accuracies of the proposed algorithms to the traditional algorithms SW [8] and SWAB [9] the following validation experiment was performed. First, three time series data streams used in [9] were selected (ECG, Space Shuttle and Radio Waves), after that the mean square error accuracies of algorithms SW and SWAB were evaluated with zero error threshold value. Second, on the basis of evaluated accuracies the appropriate confidence levels of the proposed algorithms were retrieved. Finally, Table 2 organizes the obtained results and clearly demonstrates that our proposed methods don't inferior to traditional SW and SWAB algorithms reported accuracies.

Table 2. Comparison between SW and SWAB algorithms

Dataset	**ECG**	**Space Shuttle**	**Radio Waves**
ISW	95%	90%	93%
ISWAB	95%	92.5%	95%

Our experimental study is aimed at estimating the accuracy and comparing the performance of the proposed algorithms. The first part was focused on stationary time series data streams (TSDS) and the second one was focused on non stationary data streams. The stationary data stream was generated from two synthetically distributed normally distributed time series ND25 and ND100 whereas the non-stationary data streams was obtained from two Israel's daily financial indexes TA25 and TA100 [20] . These finance indexes behavior strongly depends on time and therefore they demonstrate non stationary behavior. The few descriptive statistics for the four selected time series is shown in Table 3.

Table 3. Descriptive statistics

TSDS	N	Max	Min	Avg.	St.Dev.
ND25	1,228	1,693.82	-29.21	748.06	244.32
ND100	1,228	1,426.14	-22.79	756.70	225.00
TA25	1,228	1,237.13	333.90	748.06	244.32
TA100	1,228	1,189.04	341.04	756.70	225.00

The ND25 time series is similar to TA25 because their averages, standard deviations and lengths are equal. Same thing is right regarding to TA100. Figure 2 demonstrates the ISW algorithm evaluation on the four collected time series. The blue columns point out the non stationary data as financial indexes and red stationary data e.g. normal processes ND25 and ND100. The most obvious result is that ISW produces more accurate results[2] on stationary data (red columns) when the user specified confidence level is greater than 80%. In case of non stationary data streams the ISW algorithm produces stable but less accurate results. This stable quality pattern results from the ISW algorithm ability to detect mean concept drifts in time series data stream behavior. The remaining error will be caused by other non stationary elements, e.g. non stationary variance or/and non-stationary auto covariance.

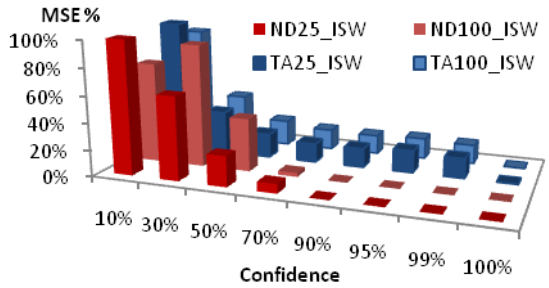


Figure 2. The ISW algorithm accuracy

Figure 2 demonstrates the performance results of ISW algorithm. Actually, this figure Y axis shows the number of created segments when it is obvious that a large number of segments increases the evaluation time of algorithm and vice versa. As previously mentioned the ISW algorithm produces good accuracy results for stationary data, i.e. red columns accuracies range from 0% to 10% when confidence level is greater than 80%.

5. Conclusions

This study has highlighted a number of limitations in existing state-of-the-art online piecewise linear segmentation approaches: Sliding Window (SW) and SWAB. First, the new relative parameter of confidence level was used instead of nominal input parameter of maximal error threshold. This parameter has two advantages: first is that the user does not need to preprocess the original time series data stream in aim to determine the reasonable maximum error value and second is that the proposed technique allows to perform cross comparisons between different time series data streams. Also, the implementation of new real time application was performed. Finally, we have performed an empirical comparison of the proposed time series segmentation algorithms on two types of time series data: stationary (normally distributed data) and non-stationary (financial data).

References

1. Biffet, A. and Kirkby, R. **Data stream mining - A practical approach**, COSI, available at http://www.cs.waikato.ac.nz /~abifet /MOA, 2009, pp 127-141
2. Chan, K. and Fu, W. **Efficient time series matching by wavelets**, proceedings of the 15th IEEE International Conference on Data Engineering, 1999
3. Das, G., Lin, K., Mannila, H., Renganathan, G. and Smyth, P. **Rule discovery from time series**, proceedings of the 3rd International Conference of Knowledge Discovery and Data Mining, 1998, pp. 16-22
4. Ge, X. and Smyth P. **Segmental Semi-Markov models for endpoint detection in plasma etching**, IEEE Transactions on Semiconductor Engineering, 2001.
5. Hoeffding, W. **Probability inequalities for sums of bounded random variables**, Journal of the American Statistical Association, 58(301), 1963, pp. 13-30
6. Hunter, J. and McIntosh, N. **Knowledge-based event detection in complex time series data**, in: Artificial Intelligence in Medicine, Springer, 1999, pp. 271-280.
7. Junker, H., Amft, O., Lukowicz, P. and Tröster, G. **Gesture spotting with body-worn inertial sensors to detect user activities**, in: Source Pattern Recognition, Elsevier, 2008, pp. 2010-2024
8. Keogh, E., Chakrabarti, K., Pazzani, M. and Mehrotra. **Dimensionality reduction for fast similarity search in large time series databases**, Journal of Knowledge and Information Systems, Proceedings of the 22th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 2000
9. Keogh, E., Chu, S., Hart, D. and Pazzani, M. **Segmenting time series: A survey and novel approach**, in: Data Mining in Time Series Databases, World Scientific Publishing Company, 2004, pp. 1-21
10. Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., and Allan, J. **Mining of concurrent text and time series**, Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, 2000, pp. 37-44
11. Li, C., Yu, P. and Castelli, V. **MALM: A framework for mining sequence database at multiple abstraction levels**, Proceedings of the 9th International Conference on Information and Knowledge Management, 1998, pp. 267-272
12. Motwani, R. and Raghavan, P. **Randomized algorithms**, ACM Computing Surveys, 1996, pp. 33-37
13. Hulten, P. and Domingos, G. **A general framework for mining massive data streams**, Journal of Computational and Graphical Statistics, 12(4), 2003, pp. 945-949
14. Osaki, R., Shimada, M. and Uehara, K. **Extraction of primitive motion for human motion recognition**, The 2nd International Conference on Discovery Science, 1999, pp. 351-352
15. Park, S., Kim, S.W. and Chu, W.W. **Segment-based approach for subsequence searches in sequence databases**, Proceedings of the 16th ACM Symposium on Applied Computing, 2001
16. Park, S., Lee, D. and Chu, W.W. **Fast retrieval of similar subsequences in long sequence databases**, Proceedings of the 3rd IEEE Knowledge and Data Engineering Exchange Workshop, 1999
17. Perng, C., Wang, H., Zhang, S. and Parker, S. **Landmarks: A new model for similarity-based pattern querying in time series databases**, Proceedings of 16th International Conference on Data Engineering, 2000
18. Qu, Y., Wang, C. and Wang, S. **Supporting fast search in time series for movement patterns in multiples scales**, Proceedings of the 7th International Conference on Information and Knowledge Management, 1998
19. Shatkay, H. and Zdonik, S. **Approximate queries and representations for large data**

sequences, Proceedings of the 12th IEEE International Conference on Data Engineering, 1996, pp. 546-553

20. Vullings, H.J.L.M., Verhaegen, M.H.G. and Verbruggen, H.B. **ECG segmentation using time-warping**, Proceedings of the 2nd International Symposium on Intelligent Data Analysis, 1997
21. Wang, C. and Wang, S. **Supporting content-based searches on time series via approximation**, Proceedings of the 12th International Conference on Scientific and Statistical Database Management, 2000
22. www.finance.yahoo.com

[1] **Reproducible Results Statement:**
In the interests of competitive scientific inquiry, all datasets and code used in this work are available, together with a spreadsheet detailing the original results, by emailing the first author.

[2] Each experimental accuracy result (i.e. a chart column) is defined by aid of mean square error measure and after that normalized by dividing by the accuracy of the worst algorithm on that experiment.

JOURNAL OF APPLIED QUANTITATIVE METHODS

International Symposium on Stochastic Models in Reliability Engineering, Life Sciences and Operations Management (SMRLO'10)

RESOURCE REALLOCATION MODELS FOR DETERMINISTIC NETWORK CONSTRUCTION PROJECTS

Avner BEN-YAIR
PhD, Center for Reliability and Risk Management,
SCE - Shamoon College of Engineering,
Beer-Sheva, Israel

E-mail: avner2740@gmail.com

Nitzan SWID

PhD Candidate, Industrial Engineering and Management Department,
Ariel University Center (AUC) of Samaria,
Ariel, Israel

E-mail: nitzan01@bezeqint.net

Dimitri GOLENKO-GINZBURG
Prof, Department of Industrial Engineering and Management (Emeritus),
Ben-Gurion University of the Negev, Beer-Sheva, Israel
& Department of Industrial Engineering and Management,
Ariel University Center (AUC) of Samaria,
Ariel, Israel

E-mail: dimitri@bgu.ac.il

Olga GRECHKO

PhD Candidate, Department of Industrial Engineering and Management,
Ben-Gurion University of the Negev,
Beer-Sheva, Israel

E-mail: olga.grechko@yahoo.com

Abstract: *Hierarchical budget reallocation models for a portfolio of construction network projects with deterministic activity durations are considered. Optimal reallocation models both at the company level and at the project level are developed.*

Key words: *network construction project; hierarchical budget reallocation models; time-cost optimization problems; R&D network projects*

1. Introduction

It can be well-recognized nowadays that the broad variety of network projects is subdivided into two classes:

- R&D projects which are carried out under random disturbances, and
- Construction projects to be identified by an essentially lower level of indeterminacy.

R&D projects are characterized by the following properties [1]:

a) the level of indeterminacy is extremely high, moreover, random parameters are implemented in the nature of the project's elements, e.g., project's activities are of random duration and an essential part of R&D projects are of random branching type. In the latter case the direction of realizing the project is of random nature;

b) R&D projects are usually realized by means of networks of acyclic type;

c) monitoring R&D projects is carried out through probability control, i.e., by implementing chance constraint models;

d) R&D projects are facilitated by both cost- and non-cost detailed resources.

Construction projects are usually characterized by:

a) operations (activities) of deterministic nature and corresponding networks of deterministic structure. The level of indeterminacy is, thus, caused not by the network, but by disorder of the project's functioning. This results in generally higher levels of determinacy;

b) networks with both acyclic and cyclic graphs. Construction network projects comprise more complicated logical relations, than R&D networks;

c) standard resource units (i.e., in the form of construction teams). Individual resources are less frequent, than for the case of R&D projects;

d) monitoring by means of scheduling network activities. Probability control is practically not used;

e) control actions are usually determined by optimal cost reallocation models at the company's or project levels.

In the paper under consideration hierarchical budget reallocation models for a portfolio of construction network projects with deterministic activity durations are suggested. Optimal reallocation models both at the company level and at the project level are developed.

2. Notation

Let us introduce the following terms:

$G_k(N, A)$ - construction network projects, $1 \le k \le n$;

n - number of projects;

C - budget value assigned to carry out all projects (to be optimized);

C_{kt} - budget assigned to $G_k(N, A)$ at moment t;

$(i, j)_k$ - activity (i, j) entering project $G_k(N, A)$;

$c_{ijk}^{(r)}$ - budget value which can be assigned to $(i, j)_k$ to operate and realize the latter ($1 \le r \le n_{ijk}$, both $c_{ijk}^{(r)}$ and $n_{ijk}^{(r)}$ are deterministic and pregiven);

n_{ijk} - number of possible durations of activity $(i, j)_k$ by means of assigned budget $c_{ijk}^{(r)}$;

$t_{ijk}^{(r)}$ - deterministic duration of activity $(i,j)_k$ by means of assigned budget $c_{ijk}^{(r)}$ (pregiven). Note that relation $c_{ijk}^{(r_1)} < c_{ijk}^{(r_2)} \Rightarrow t_{ijk}^{(r_2)} < t_{ijk}^{(r_1)}$ holds;

S_{ijk} - moment activity $(i,j)_k$ starts to be operated (to be determined);

F_{ijk} - moment activity $(i,j)_k$ is finished (to be determined, depends on assigned budget $c_{ijk}^{(r)}$; note that relation $F_{ijk} = S_{ijk} + t_{ijk}^{(r)}$ holds);

C_t - total non-realized budget at moment t; $0 \le t$; $C_0 = C$;

F_k - actual moment project $G_k(N,A)$ is accomplished (to be determined);

D_k - the due date for project $G_k(N,A)$ (pregiven);

$G_{kt}(N,A)$ - a non-realized part of project $G_k(N,A)$ at moment t;

$T_{cr}\left\{G_{kt}(N,A)/c_{ijk}^{(r)}\right\}$ - critical path length of project $G_{kt}(N,A)$ subject to assigned budget values $c_{ijk}^{(r)}$.

3. The Problem

The optimization problem is as follows:

At any moment $t \ge 0$ determine both C_t and C_{kt}, $1 \le k \le n$, as well as values $c_{ijk}^{(r)}$ for all non-accomplished at moment t activities $(i,j)_k$, to minimize objective

$$\underset{\{C_{kt}, c_{ijk}^{(r)}\}}{Min}\ C_t \tag{1}$$

subject to

$$\sum_{k=1}^{n} C_{kt} = C, \tag{2}$$

$$\sum_{\{i,j\}_k^t} c_{ijk}^{(r)} \le C_{kt}, \tag{3}$$

$$\underset{r}{Min}\ c_{ijk}^{(r)} \le c_{ijk}^{(r)} \le \underset{r}{Max}\ c_{ijk}^{(r)}, \tag{4}$$

$$t + T_{cr}\left\{G_{kt}(N,A)/c_{ijk}^{(r)}\right\} \le F_k,\ 1 \le k \le n, \tag{5}$$

where $\{i,j\}_k^t$ denotes the set of activities entering $G_{kt}(N,A)$.

Note that since we deal only with budget values C_{kt}, and taking into account that all projects under consideration can be regarded as independent ones, problem (1-5) can be simplified and transformed to the case of one project only. Cancel index k of the project and formulate the amended problem as follows:

Given for each activity (i,j) a set of n_{ij} couples $\left\{c_{ij}^{(r)}, t_{ij}^{(r)}\right\}$, $1 \le r \le n_{ij}$, where each couple denotes the possible assigned budget value with the corresponding activity duration, together with the pregiven due date D of a deterministic network project $G(N,A)$, determine budget values c_{ij} as well as the total budget $C = \sum_{\{i,j\} \in G(N,A)} c_{ij}$, which results in minimizing value C

$$\underset{\{c_{ij}\}\subset G(N,A)}{Min}\ C = C^* \tag{6}$$

subject to

$$\sum_{(i,j)\subset G(N,A)} c_{ij} \le C^* \tag{7}$$

and

$$T_{cr}\left\{G(N,A)/c_{ij}\right\} \le D\,. \tag{8}$$

It can be well-recognized that determining C for each project $G_k(N,A)$ independently, results in providing objectives (1-2) using relation (3).

To solve problem (6-8), we require solving an auxiliary problem (AP).

4. Auxiliary Problem (AP)

Problem AP [2] may be considered as follows:

Given a deterministic PERT graph $G(N,A)$ together with pregiven functions $t_{ij}^{(r)}$ and $c_{ij}^{(r)}$,

$t_{ij}^{(r)} = f_{ij}\left(c_{ij}^{(r)}\right),\ 1 \le r \le n_{ij}$,

number n_{ij} pregiven for each activity $(i,j)\subset G(N,A)$, determine:

• the minimal total project direct costs C,

$Min\ C$, and (9)

• the optimal assigned budget values c_{ij}^{opt}, subject to

$$T_{cr}\left\{t_{ij} = f_{ij}\left(c_{ij}^{opt}\right)\right\} \le D\,, \tag{10}$$

$$\sum_{\{i,j\}} c_{ij}^{opt} = C\,, \tag{11}$$

$$c_{ij\,\min} \le c_{ij}^{opt} \le c_{ij\,\max}\,, \tag{12}$$

where D stands for a pregiven due date.

Problem AP is solved in [2] by means of heuristic methods based on transferring the possible budget ΔC from non-critical activities (which have practically no influence on the project's critical path duration) to critical activities either belonging to the critical path or being very close to the latter.

The corresponding algorithm is described in [2].

5. Problem's (6-8) Solution

The step-wise procedure of solving problem (6-8) is as follows:

Step 1. Determine the minimal budget value C_1,

$$C_1 = \sum_{(i,j)\subset G(N,A)} \left(\min_r c_{ij}^{(r)}\right), \tag{13}$$

which by no means can be diminished - otherwise problem (6-8) has no solution.

Step 2. Determine the maximal budget value C_2

$$C_2 = \sum_{(i,j)\subset G(N,A)} \left(\max_r c_{ij}^{(r)}\right), \tag{14}$$

which by no means should be increased - otherwise solving problem (6-8) results in redundant budget spending.

Step 3. For both cases considered in Steps 1-2, calculate critical path lengths for graph $G(N,A)$:

$$t_{ij}(C_1) = \min_r t_{ij}^{(r)}, \ \ (i,j)\subset G(N,A), \tag{15}$$

$$t_{ij}(C_2) = \max_r t_{ij}^{(r)}, \ \ (i,j)\subset G(N,A), \tag{16}$$

Call henceforth T_1 the critical path length of $G(N,A)$ in case (15) and T_2 - in case (16). It can be well-recognized that when $D < T_1$ problem (6-8) has no solution. Taking into account [1,3] the obvious relation

$$c^{(r_1)}(i,j) > c^{(r_2)}(i,j) \Rightarrow t^{(r_1)}(i,j) < t^{(r_2)}(i,j) \tag{17}$$

for any $(i,j)\subset G(N,A)$, a conclusion can be drawn that relation $D \geq T_2$ should result in $C^* \leq C_2$.

Assume, further [1,3], that the critical path length T_{cr} of any project $G(N,A)$, designated henceforth as $T_{cr}(C)$, depends linearly on budget C assigned to that project; in other words,

$$\frac{T_{cr}(C') - T_{cr}(C'')}{C'' - C'} \approx const \tag{18}$$

holds.

Step 4. Calculate, by means of (18), the preliminary (non-minimal!) value C corresponding to the due date D. Using

$$\frac{T_{cr}(C_1) - T_{cr}(C_2)}{C_2 - C_1} \approx \frac{T_{cr}(C_1) - D}{C - C_1},$$

we finally obtain

$$C = C_1 + \frac{(T_{cr}(C_1) - D)\cdot(C_2 - C_1)}{T_{cr}(C_1) - T_{cr}(C_2)}. \tag{19}$$

Step 5. Solve subsidiary Problem A:

Given budget C assigned to project $G(N,A)$, determine the minimal critical path length $T_{\min}^{(\nu)}$ by means of redistributing C among activities $(i,j)\subset G(N,A)$. Let ν be the number of the current iteration.

Step 6. Compare values $T_{\min}^{(\nu)}$ and D. If $T_{\min}^{(\nu)} > D$, go to 7. Otherwise apply the next step.

Step 7. Set $T_{\min}^{(\nu)} \Rightarrow T_{cr}(C_1)$, $C \Rightarrow C_1$, $\nu + 1 = \nu$. Go to Step 4.

Step 8. Compare $T_{\min}^{(\nu)}$ with $T_{\min}^{(\nu-1)}$. If $\left|T_{\min}^{(\nu)} - T_{\min}^{(\nu-1)}\right| < \varepsilon$, where ε stands for the problem's accuracy, apply the next step. Otherwise go to Step 11.

Step 9. Budget value C referring to value $T_{\min}^{(\nu)}$ at Step 5, is considered to be the optimal (minimal) value C^* with local budget values $\left\{c_{ij}^{(r)}\right\}$ obtained in the course of solving Problem A at Step 5.

Step 10. The problem's solution terminates.

Step 11. Set $T_{\min}^{(\nu)} \Rightarrow T_2$, $C \Rightarrow C_1$, $\nu + 1 = \nu$. Go to Step 4.

The solution of the global problem (1-4) can be obtained by summarizing values C_{kt} calculated independently for each k at any moment $t \geq 0$. If $t > 0$ we take into account the updated graph $G_t(N, A)$ instead of $G(N, A)$ (for a single project) and determine value C_t^* instead of value C^*. After the algorithm's (6-8) termination all optimal values C_{kt}^* are summarized in order to obtain the updated total value C_t^*.

6. Conclusions

1. The newly developed algorithm is easy in usage and effective in practice. Its implementation requires mostly no more than 3÷5 iteration.

2. The algorithm has been widely used both for medium- and large-scale projects with the number of activities exceeding 50÷100. In all cases the algorithm performed well and the number of iterations did not exceed 5.

3. The algorithm can be realized on the basis of classical algorithms which are widely used in network planning and are described in many textbooks on project management.

4. The model suggested in this paper is open for various modifications: e.g., instead of purely deterministic values defining the closeness of activities to the critical area and, thus, the level of their influence on the project's duration, other terms may be implemented. However, those modifications are not essential from the principal point of view.

5. In our opinion, the developed research can be widely used for network construction projects of acyclic type.

References

1. Golenko-Ginzburg, D. **Stochastic Network Models for Managing R&D Projects**, Nauchnaya Kniga Publishers, Voronez, Russian Federation, 2010 (in Russian)
2. Greenberg, D., Golenko-Ginzburg, D. and Ben-Yair, A. **Time-cost optimization problem for deterministic PERT net-works**, Communications in Dependability and Quality Management, 10(2), 2007, pp. 80-87
3. Ben-Yair, A. **Harmonization models in strategic management and safety engineering**, Ph.D. Thesis, Ben-Gurion University of the Negev, Beer-Sheva, Israel, 2004

JOURNAL OF APPLIED QUANTITATIVE METHODS

International Symposium on Stochastic Models in Reliability Engineering, Life Sciences and Operations Management (SMRLO'10)

IMPLEMENTING BETA-DISTRIBUTION IN PROJECT MANAGEMENT

Doron GREENBERG

PhD, Department of Economics and Business Administration, Faculty of Social Science, Ariel University Center (AUC) of Samaria, Ariel, Israel

E-mail: dorongreen2@gmail.com

Dimitri GOLENKO-GINZBURG

Prof, Department of Industrial Engineering and Management (Emeritus), Ben-Gurion University of the Negev, Beer-Sheva, Israel
& Department of Industrial Engineering and Management, Ariel University Center (AUC) of Samaria, Ariel, Israel

E-mail: dimitri@bgu.ac.il

Abstract: *A research is undertaken to justify the use of beta-distribution p.d.f. for man-machine type activities under random disturbances. The case of using one processor, i.e., a single resource unit, is examined. It can be proven theoretically that under certain realistic assumptions the random activity – time distribution satisfies the beta p.d.f.*

Changing more or less the implemented assumptions, we may alter to a certain extent the structure of the p.d.f. At the same time, its essential features (e.g. asymmetry, unimodality, etc.) remain unchanged.

The outlined above research can be applied to semi-automated activities, where the presence of man-machine influence under random disturbances is, indeed, very essential. Those activities are likely to be considered in organization systems (e.g. in project management), but not in fully automated plants.

Key words: *random activity duration; time – activity beta-distribution; operating by means of a single processor; convergence to a beta-distribution "family"*

1. Introduction

In PERT analysis [1-24, etc.] the activity-time distribution is assumed to be a beta-distribution, and the mean value and variance of the activity time are estimated on the basis of the "optimistic", "most likely" and "pessimistic" completion times, which are subjectively determined by an analyst. The creators of PERT [3, 17] worked out the basic concepts of PERT analysis, and suggested the estimates of the mean and variance values

$$\mu = \frac{1}{6}(a + 4m + b), \tag{1}$$

$$\sigma^2 = \frac{1}{36}(b - a)^2, \tag{2}$$

subject to the assumption that the probability density function (p.d.f.) of the activity time is

$$f_y(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{(y-a)^{\alpha-1}(b-y)^{\beta-1}}{(b-a)^{\alpha+\beta-1}}, \quad a < y < b, \quad \alpha, \beta > 0. \tag{3}$$

Here a is the optimistic time, b - the pessimistic time, and m stands for the most likely (modal) time.

Since in PERT applications parameters a and b of p.d.f. (3) are either known or subjectively determined, we can always transform the density function to a standard form,

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1, \quad \alpha, \beta > 0, \tag{4}$$

where $x = \frac{y-a}{b-a}$ has the following parameters:

$$\mu_x = \frac{\mu_y - a}{b-a}, \quad \sigma_x = \frac{\sigma_y}{b-a}, \quad m_x = \frac{m_y - a}{b-a}. \tag{5}$$

Let $\alpha - 1 = p$, $\beta - 1 = q$. Then p.d.f. (4) becomes

$$f(x) = \frac{\Gamma(p+q+2)}{\Gamma(p+1)\Gamma(q+1)} x^p(1-x)^q, \quad 0 < x < 1, \quad p, q > -1, \tag{6}$$

with the mean, variance and mode as follows:

$$\mu_x = \frac{p+1}{p+q+2}, \tag{7}$$

$$\sigma_x^2 = \frac{(p+1)(q+1)}{p+q+2}, \tag{8}$$

$$m_x = \frac{p}{p+q}. \tag{9}$$

From (6) and (9) it can be obtained

$$f(x) = \frac{\Gamma(p+q+2)}{\Gamma(p+1)\Gamma(q+1)} x^p(1-x)^{p(1/m_x - 1)}. \tag{10}$$

Thus, value m_x, being obtained from the analyst's subjective knowledge, indicates the density function. On the basis of statistical analysis and some other intuitive arguments, the creators of PERT assumed that $p + q \cong 4$. It is from that assertion that estimates (1) and (2) were finally obtained, according to (6-9).

Although the basic concepts of PERT analysis have been worked out many years ago [3, 17], they are open till now to considerable criticism. Numerous attempts have been made to improve the main PERT assumptions for calculating the mean μ_x and variance

σ_x^2 of the activity-time on the basis of the analyst's subjective estimates. In recent years, a very sharp discussion [7, 10, 14, and 21] has taken place in order to raise the level of theoretical justifications for estimates (1) and (2).

Grubbs [12] pointed out the lack of theoretical justification and the unavoidable defects of the PERT statements, since estimates (1) and (2) are, indeed, "rough" and cannot be obtained from (3) on the basis of values a, m and b determined by the analyst. Moder [18-19] noted that there is a tendency to choose the most likely activity – time m much closer to the optimistic value a than to the pessimistic one, b, since the latter is usually difficult to determine and thus is taken conservatively large. Moreover, it is shown [8] that value m, being subjectively determined, has approximately one and the same relative location point in $[a,b]$ for different activities. This provides an opportunity to simplify the PERT analysis at the expense of some additional assumptions. McCrimmon and Ryavec [16], Lukaszewicz [15] and Welsh [22] examined various errors introduced by the PERT assumptions, and came to the conclusion that these errors may be as great as 33%. Murray [20] and Donaldson [4] suggested some modifications of the PERT analysis, but the main contradictions nevertheless remained. Farnum and Stanton [6] presented an interesting improvement of estimates (1) and (2) for cases when the modal value m is close to the upper or lower limits of the distribution. This modification, however, makes the distribution law rather uncertain, and causes substantial difficulties to simulate the activity network.

In this paper, a research will be undertaken to develop some theoretical justifications for using the beta-distribution p.d.f.

2. The Operation's Description

We will consider a man-machine operation which is carried out by one processor, i.e., by one resource unit. The processor may be a machine, a proving ground, a department in a design office, etc.

Assume that the operation starts to be processed at a pregiven moment T_0. The completion moment F of the operation is a random value with distribution range $[T_1, T_2]$. Moment T_1 is the operation's completion moment on condition that the operation will be processed without breaks and without delays, i.e., value T_1 is a pregiven deterministic value. Assume, further, that the interval $[T_0, T_1]$ is subdivided into n equal elementary periods with length $(T_1 - T_0)/n$. If within the first elementary period $[T_0, T_0 + (T_1 - T_0)/n]$ a break occurs, it causes a delay of length $\Delta = (T_2 - T_1)/n$. The operation stops to be processed within the period of delay in order to undertake necessary refinements, and later on proceeds functioning with the finishing time of the first elementary period
$T_0 + (T_1 - T_0)/n + (T_2 - T_1)/n = T_0 + (T_2 - T_0)/n$.

It is assumed that there cannot be more than one break in each elementary period. The probability of a break at the very beginning of the operation is set to be p. However, in the course of carrying out the operation, the latter possesses certain features of self-adaptivity, as follows:

- the occurrence of a break within a certain elementary period results in increasing the probability of a new break at the next period by value η, and
- on the contrary, the absence of a break within a certain period decreases the probability of a new break within the next period, practically by the same value.

3. The Concept of Self-Adaptivity

The probabilistic self-adaptivity can be formalized as follows:

Denote A_i^k the event of occurrence of a break within the $(i+1)$-th elementary period, on condition, that within the i preceding elementary periods k breaks occurred, $1 \le k \le i \le n$. It is assumed that relation

$$P\left(A_i^k\right) = \frac{p + k \cdot \eta}{1 + i \cdot \eta} \tag{11}$$

holds. Note that (11) is, indeed, a realistic assumption.

Relation (11) enables obtaining an important assertion. Let $P\left(A_i^0\right)$ be the probability of the occurrence of a break within the $(i+1)$-th period on condition, that there have been no breaks at all as yet. Since

$$P\left(A_i^0\right) = \frac{p}{1 + i \cdot p}, \tag{12}$$

it can be well-recognized that relation

$$\frac{P\left(A_i^{k+1}\right) - P\left(A_i^k\right)}{P\left(A_i^0\right)} = \frac{\eta}{p} \tag{13}$$

holds. Thus, an assertion can be formulated as follows:

Assertion. *Self-adaptivity (11) results in a probability law for delays with a constant ratio (13) for a single delay.*

4. Calculating the Activity-Time Distribution

Let us calculate the probability $P_{m,n}$ of obtaining m delays within n elementary periods, i.e., the probability of completing the operation at the moment

$$F = T_1 + m \cdot \Delta = T_1 + \frac{m}{n}\left(T_2 - T_1\right).$$

The number of sequences of n elements with m delays within the period $\left[T_0, F\right]$ is equal C_n^m, while the probability of each such sequence equals

$$\frac{\left[\prod_{i=0}^{m-1}\left(p + i\eta\right)\right]\left[\prod_{i=0}^{n-m-1}\left(1 - \eta + i\eta\right)\right]}{\prod_{i=0}^{n-1}\left(1 + i\eta\right)}. \tag{14}$$

Relation (14) stems from the fact that if breaks occurred within h periods and did not occur within k periods, the probability of the occurrence of the delay at the next period is equal

$$\frac{p+h\eta}{1+(k+h)\eta}\ , \tag{15}$$

while the probability of the delay's non-appearance at the next period satisfies

$$\frac{1-\eta+k\eta}{1+(k+h)\eta}\ . \tag{16}$$

Using (14-16), we finally obtain

$$P_{m,n} = C_n^m \frac{\left[\prod_{i=0}^{m-1}(p+i\eta)\right]\left[\prod_{i=0}^{n-m-1}(1-\eta+k\eta)\right]}{\prod_{i=0}^{n-1}(1+i\eta)}\ . \tag{17}$$

Note that $\eta=0$, i.e., the absence of self-adaptivity, results in a regular binomial distribution.

Let us now obtain the limit value $P_{m,n}$ on condition that $n\to\infty$. From relation (17) we obtain

$$\frac{P_{m+1,n}}{P_{m,n}} = \frac{n-m}{m+1}\frac{p+m\eta}{1-p+(n-m-1)\eta}\ . \tag{18}$$

Denoting $\frac{p}{\eta}=\alpha$, $\frac{p}{\eta}\left(\frac{1}{p}-1\right)=\beta$, we obtain

$$\frac{P_{m+1,n}-P_{m,n}}{P_{m,n}} = \frac{(\alpha-1)n+(2-\alpha-\beta)m-\beta+1}{(m+1)(\beta+n-m-1)} = \frac{(\alpha-1)+(2-\alpha-\beta)\frac{m}{n}+\frac{1-\beta}{n}}{n\frac{m+1}{n}\left(1-\frac{m+1}{n}+\frac{\beta}{n}\right)}\ .$$

Denoting $m/n=x$, $(m+1)/n=x+\Delta x$, $P_{m,n}=y$, $P_{m+1,n}=y+\Delta y$, via convergence $n\to\infty$ or $\Delta x\to 0$ and, later on, by means of integration, we finally obtain

$$y = Cx^{\alpha-1}(1-x)^{\beta-1}\ . \tag{19}$$

It can be well-recognized that the p.d.f. of random value $\xi=\lim_{n\to\infty}\frac{m}{n}$ satisfies

$$p_\xi(x) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}\ , \tag{20}$$

where $B(\alpha,\beta)$ represents the Euler's function. Thus, relation (20) practically coincides with (10).

Thus, ξ is a random value with the beta-distribution activity–time p.d.f. By transforming $x=(y-a)/(b-a)$, we obtain the well-known p.d.f. (3).

5. Conclusions

The following conclusions can be drawn from the study:

1. Under certain realistic assumptions we have proven theoretically that the activity-time distribution satisfies the beta-distribution with p.d.f. (3) being used in PERT analysis.
2. Changing more or less the implemented assumptions, we may alter to a certain extent the structure of the p.d.f. At the same time, its essential features (e.g. asymmetry, unimodality, etc.) remain unchanged.
3. The outlined above research can be applied to semi-automated activities, where the presence of man-machine influence under random disturbances is, indeed, very essential. Those activities are likely to be considered in organization systems (e.g., in project management), but not in fully automated plants.

References

1. Battersby, A. **Network Analysis for Planning and Scheduling**, 3rd edition, MacMillan: London, 1970
2. Berny, J. **A new distribution function for risk analysis**, J. Oper. Res. Soc., 40(12), 1989, pp. 1121-1127
3. Clark, C.E. **The PERT model for the distribution of an activity**, Opns. Res., 10, 1962, pp. 405-406
4. Donaldson, W.A. **The estimation of the mean and variance of PERT activity time**, Opns. Res., 13, 1965, pp. 382-385
5. Elmaghraby, S.E. **Activity Networks: Project Planning and Control by Network Models**, Wiley: New-York, 1977
6. Farnum, N.R. and Stanton, L.W. **Some results concerning the estimation of beta distribution parameters in PERT**, J. Oper. Res. Soc., 38, 1987, pp. 287-290
7. Gallagher, C. **A note on PERT assumptions**, Mgmt. Sci., 33, 1987, pp. 1360-1362
8. Golenko-Ginzburg, D. **Statistical Models in Network Planning and Control**, Nauka: Moscow (in Russian), 1966
9. Golenko-Ginzburg, D. **On the distribution of activity time in PERT**, J. Oper. Res. Soc., 39(8), 1988, pp. 767-771
10. Golenko-Ginzburg, D. **PERT assumptions revisited**, Omega, 17(4), 1989, pp. 393-396
11. Gonik, A. **Planning and controlling multilevel stochastic projects**, Ph.D. Thesis, Ben-Gurion University of the Negev, Beer-Sheva, Israel, 1995
12. Grubbs, F.E. **Attempts to validate certain PERT statistics or "picking on PERT"**, Opns. Res., 10, 1962, pp. 912-915
13. Kelley, J.E. Jr. **Critical path planning and scheduling: Mathematical basis**, Opns. Res., 9(3), 1961, pp. 296-320
14. Littlefield, T.K. Jr. and Randolph, P.H. **Another note on PERT times**, Mgmt. Sci., 33, 1987, pp. 1357-1359
15. Lukaszevicz, J. **On the estimation of errors introduced by standard assumptions concerning the distribution of activity duration PERT calculations**, Opns. Res., 13, 1965, pp. 326-327
16. MacCrimmon, K.R. and Ryaveck, C.A. **An analytical study of the PERT assumptions**, Opns. Res., 12, 1964, pp. 16-37
17. Malcolm, D., Roseboom, J., Clark, C. and Fazar, W. **Application of a technique for research and development program evaluation**, Opns. Res., 7, 1959, pp. 646-669

18. Moder, J.J., Phillips, C.R. and Davis, E.W. **Project Management with CPM and PERT and Precedence Diagramming**, Van-Nostrand Reinhold Co., Inc.: New-York, 1983
19. Moder, J.J. and Cecil, R.P. **Project Management with CPM and PERT**, Van-Nostrand Reinhold Co., Inc.: New-York, 1970
20. Murray, J.E. **Consideration of PERT Assumptions**, Conduction Corporation, Ann. Arbor: Michigan, 1962
21. Sasieni, M.W. **A note on PERT times**, Mgmt. Sci., 32, 1986, pp. 1652-1653
22. Welsh, D. **Errors introduced by a PERT assumption**, Opns. Res., 13, 1965, pp. 141-143
23. Williams, T.M. **Practical use of distributions in network analysis**, J. Oper. Res. Soc., 43(3), 1992, pp. 265-270
24. Williams, T.M. **What are PERT estimates?** J. Oper. Res. Soc., 46(12), 1995, pp. 1498-1504

JOURNAL OF APPLIED QUANTITATIVE METHODS

International Symposium on Stochastic Models in Reliability Engineering, Life Sciences and Operations Management (SMRLO'10)

FUZZY PROBABILISTIC MODELS FOR STRUCTURAL SERVICEABILITY[1]

Milan HOLICKÝ
Prof, Klokner Institute,
Czech Technical University in Prague,
Prague, Czech Republic

E-mail: Milan.Holicky@klok.cvut.cz

Abstract: *Structural serviceability is of uttermost importance for the overall performance of many common structures. As a rule, both the load effects (serviceability indicators due to loading) and admissible constraints (ensuring required structural performance) are random variables of considerable scatter and significant vagueness. Common experience indicates that a structure does not loose its ability to comply with specified performance requirements abruptly at a distinct point of the serviceability indicator, but gradually within a certain transition interval. Fuzzy-probabilistic methods are therefore employed to analyze the structural serviceability.*
As an example, serviceability limit states of water retaining structures with respect to cracking are investigated in detail. Fuzzy probabilistic models are proposed to derive theoretical models for the limiting crack width. It is shown that the fuzzy probabilistic distribution of serviceability requirements may be used similarly as classical distribution function to specify the characteristic value of limiting crack width, to analyze reliability of crack width and to optimize structural design to achieve the minimum total costs.

Key words: *fuzzy probabilistic models; probability; serviceability; cracks width; optimization*

1. Introduction

Structural performance is becoming a fundamental concept of advanced engineering design in construction. However, performance requirements (including serviceability, safety, security, comfort, functionality) of buildings and engineering works are often affected by various uncertainties that can hardly be entirely described by traditional probabilistic models. As a rule, transformation of human desires, particularly those describing occupancy comfort and aesthetic aspects, to performance (user) requirements

often results in an indistinct or imprecise definition of the technical criteria for relevant performance indicators (for example the limiting defection or crack width).

Thus, in addition to natural randomness of basic variables, performance requirements may be considerably affected by vagueness in the definition of technical criteria. Two types of uncertainty of performance requirements are therefore identified here: randomness, handled by commonly used methods of the theory of probability, and, fuzziness, described by basic tools of the recently developed theory of fuzzy sets (Brown and Yao 1983). Similarly as in the previous studies (Holický 2006), the fundamental condition of structural performance, $S \le R$, between an action effect S and a relevant performance requirement R, is considered assuming the randomness of S and both the randomness and fuzziness of R. In this study the performance resistance R is analysed in detail.

An illustrative example of continuous vibration in offices is used throughout the paper to clarify general concepts. In this example, it is shown that it is impossible to identify a distinct value of an appropriate indicator (root mean square value of acceleration) that would separate a satisfactory from an unsatisfactory performance (Holický et al. 2009). Typically, a broad transition region is observed, where the building is gradually losing its ability to perform adequately and where the degree of damage (inadequate performance or malfunction) gradually increases.

2. Fuzzy Probabilistic Models of Performance Requirements

Fuzziness due to vagueness and imprecision in the definition of performance requirement R is described by the membership function $v_R(x)$ indicating the degree of the membership of a structure in a fuzzy set of damaged (unserviceable) structures (Holický 2006); here x denotes a generic point of a relevant performance indicator (a deflection or a root mean square of acceleration) considered when assessing structural performance. Common experience indicates that a structure is loosing its ability to comply with specified requirements gradually within a certain transition interval $\langle r_1, r_2 \rangle$.

The membership function $v_R(x)$ describes the degree of structural damage (lack of functionality). If the rate $dv_R(x)/dx$ of the "performance damage" in the interval $\langle r_1, r_2 \rangle$ is constant (a conceivable assumption), then the membership function $v_R(x)$ has a piecewise linear form as shown in Figure 1. It should be emphasized that $v_R(x)$ describes the non-random (deterministic) part of uncertainty in the requirement R related to economic and other consequences of inadequate performance. The randomness of R at any damage level $v = v_R(x)$ may be described by the probability density function $\varphi_R(x|v)$ (see Figure 1), for which a normal distribution having the constant coefficient of variation $V_R = 0.10$ is considered in the following.

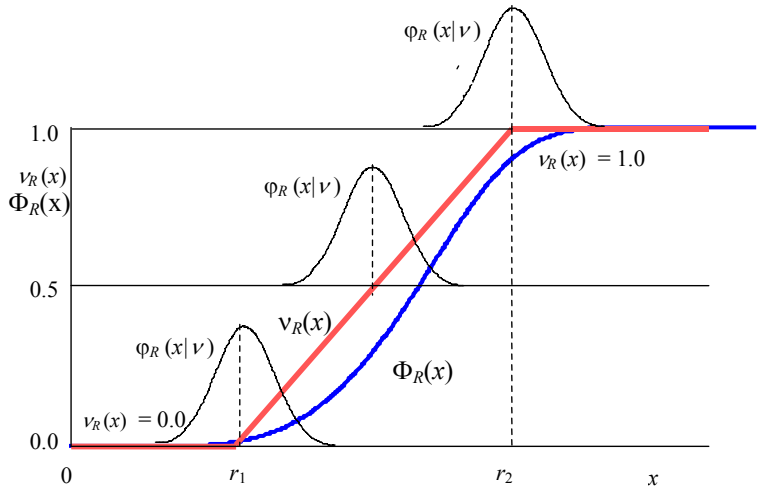


Figure 1. The fuzzy probabilistic model of the performance requirement R

The transition region $\langle r_1, r_2 \rangle$, where the structure is gradually losing its ability to perform adequately and its damage increases, may be rather broad, depending on the nature of the performance requirement. For common serviceability requirements (deflections) the upper limit r_2 may be a multiple of the lower limit r_1 (for example, $r_2 = 2 \cdot r_1$).

The fuzzy probabilistic measures of structural performance is defined as the damage function $\Phi_R(x)$ being the weighted average of damage probabilities reduced by the corresponding damage level (Holický 2006)

$$\Phi_R(x) = \frac{1}{N} \int_0^1 \nu \left(\int_{-\infty}^{x} \varphi_R(x' \mid \nu) \mathrm{d}x' \right) \mathrm{d}\nu , \tag{1}$$

where N denotes a factor normalizing the damage function $\Phi_R(x)$ to the conventional interval $\langle 0,1 \rangle$ (see Figure 1) and x' is a generic point of x. The density of the damage $\varphi_R(x)$ follows from (1) as

$$\varphi_R(x) = \frac{1}{N} \int_0^1 \nu \ \varphi_R(x \mid \nu) \, \mathrm{d}\nu . \tag{2}$$

The damage function $\Phi_R(x)$ and density function $\varphi_R(x)$ defined by equation (1) and (2) may be considered as generalized distribution functions of the performance requirements R that can be used similarly as classical probabilistic functions.

3. Fuzzy Probability of Performance Failure

The damage function $\Phi_R(x)$ defined by equation (1) may be used similarly as the classical distribution function of structural resistance. If the action effect S of a structural member is well-known and its probability density function $\varphi_S(x)$ is available, the fuzzy probability of performance failure π_f may be assessed as

$$\pi_f = \int_{-\infty}^{\infty} \varphi_S(x)\Phi_R(x)\mathrm{d}x\,. \tag{3}$$

The damage function $\Phi_R(x)$ defined by equation (1) and the fuzzy probability of performance failure π defined by equation (3) enable the formulation of various design criteria in terms of relevant randomness and fuzziness parameters. In addition, fuzzy probabilistic optimization can be used to specify the optimum structural design and appropriate fuzzy reliability level. However, adequate data for the specification of the fuzziness parameters r_1, r_2, the membership function $v_R(x)$ and its coefficient of variation V_R (describing the requirement R) and the probability density $\varphi_S(x)$ of the load effect S are needed.

4. The Characteristic Value of Performance Requirement

The characteristic value r_K of the performance requirement R can be determined as a specified fractile of the damage function $\Phi_R(x)$

$$\pi_k = \Phi_R(r_k)\,. \tag{4}$$

Here π_k is the fuzzy probability of not achieving the characteristic value r_K. It may differ from the commonly accepted value $\pi_k = 0.05$ in the case of classical definition of probability. Previous studies (Holický 2006) based on the fuzzy probabilistic optimization indicate that the characteristic value of serviceability requirements (limiting value in design) corresponding to the probability $\pi_k = 0.05$ may not lead to the optimum reliability level.

5. The Limiting Crack Width for Water Retaining Structures

Water retaining structures are usually designed on the basis of crack width requirements. The limiting values are commonly within the interval from $r_1 = 0.05\,mm$ to $r_2 = 0.2\,mm$ (Holický et al. 2009). Considering these values as the deterministic lower and upper bounds of the transition region, the membership function $v_R(x)$, the damage function $\Phi_R(x)$ and the density function $\varphi_R(x)$ defined by equations (1) and (2) are shown in Figure 2. It should be mentioned that the transition region might be slightly shifted if the values $r_1 = 0.05\,mm$ and $r_2 = 0.2\,mm$ are considered as fractiles of the admissible crack width. Figure 2 also indicates the characteristic value of the limiting crack width $r_k = 0.082\,mm$ corresponding to the conventionally accepted probability $\pi_k = 0.05$ of not achieving r_k, if the probability $\pi_k = 0.20$, then the characteristic value is $r_k = 0.115\,mm$.

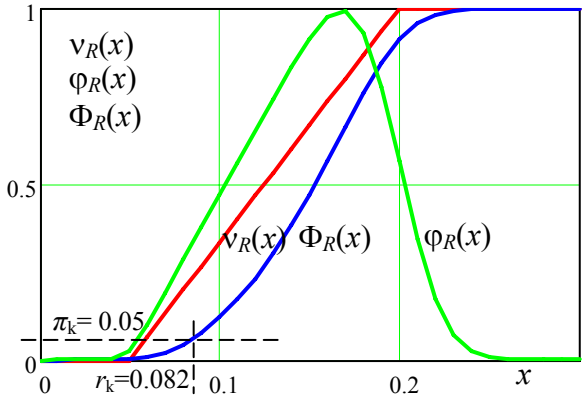


Figure 2. The membership function $v_R(x)$, the damage function $\Phi_R(x)$ and the damage density function $\varphi_R(x)$ for the transition region from $r_1 = 0.05\,mm$ to $r_2 = 0.2\,mm$ and $V_R = 0.10$

It follows from Figure 2 that the characteristic value $r_k = 0.082\,mm$ is relatively close to the lower bound of transition region $a = 0.05\,mm$. However, as indicated above, in the case of serviceability requirements the probability $p = 0.05$ may not be the optimum value used to define the characteristic serviceability resistance, for example the limiting crack width.

Note that damage density may be well approximated by the Beta distribution having the mean $0.15\,mm$, standard deviation $0.039\,mm$, the lower bound $0.02\,mm$ and the upper bound $0.238\,mm$.

6. Fundamental Concepts in Eurocodes

6.1. Crack Width

Verification of cracking is mostly based on semi empirical formulae supported by experimental evidence, experience and structural detailing (EN 1992-1-1 (2004), Narayanan and Beeby (2005)). A number of different approaches leading to considerably diverse results may be found in literature and codes of practice ((EN 1992-1-1 (2004), EN 1992-1-3 (2006), Narayanan and Beeby (2005)). The following probabilistic study is based on the concepts accepted in Eurocodes. Basic relationship for the assessment of crack width w is written in the form of simple compatibility condition (EN 1992-1-1 (2004), Narayanan and Beeby (2005))

$$w_m = S_{rm}\varepsilon_m, \tag{5}$$

where w_m is the mean crack width, S_{rm} the mean crack spacing and ε_m the mean strain in between the two adjacent cracks. The mean crack spacing S_{rm} can be assessed using a semi empirical formula (EN 1992-1-1 (2004), Narayanan and Beeby (2005))

$$S_{rm} = 2c + 0.25 \cdot k_1 \cdot k_2 \cdot \phi / \rho_{eff} \,, \tag{6}$$

where c denotes concrete cover, k_1 is a coefficient taking into account bond properties of the reinforcement (a value 0.8 for high bond and 1.6 for smooth bars), k_2 is a coefficient depending on the form of stress distribution (a value 0.5 for bending, 1.0 for pure tension), ϕ is the bar diameter and ρ_{eff} the effective reinforcement ratio $A_s / A_{c,eff}$. Here A_s is the reinforcement area and $A_{c,eff}$ is the effective concrete area surrounding the reinforcing bars. Detailed instructions on how to determine the area $A_{c,eff}$ are provided in EN 1992-1-1 (2004). Note that $A_{c,eff}$ is usually smaller than the concrete area A_c considered normally for the reinforcement ratio of flexural or compressive members, and, consequently, the effective reinforcement ratio ρ_{eff} may be greater than the commonly used reinforcement ratio ρ.

The mean strain ε_m for reinforced concrete members (non prestressed) may be calculated from the expression (EN 1992-1-1 (2004), Narayanan and Beeby (2005))

$$\varepsilon_m = \varepsilon_{sm} - \varepsilon_{cm} = \frac{\sigma_s - k_t f_{ct,eff}(1 + \alpha_e \rho_{eff}) / \rho_{eff}}{E_s} \geq 0.6 \frac{\sigma_s}{E_s} \,, \tag{7}$$

where ε_{sm} is the mean strain in reinforcing bars, ε_{cm} the mean strain in surrounding concrete, σ_s is the stress in tension reinforcement at the crack section, k_t is a factor dependent on the duration of the load (0.6 for short term loading, 0.4 for long term loading), $f_{ct,eff}$ is the mean of the tensile strength of the concrete, effective at the time when the cracks may first develop ($f_{ct,eff} \leq f_{ctm}$), and α_e is the ratio modulus E_s / E_{cm}.

6.2. Design Condition

To verify the mean crack width, w_m is multiplied by the factor $\beta_w (= 1.7)$ and compared with the limiting crack width w_d. Thus, it is required that

$$w_k \approx \beta_w w_m < w_{\lim} \,. \tag{8}$$

It is assumed that the product $w_k \approx \beta_w w_m$ is called the characteristic value of the crack width, which is supposed to be equal to the upper 5% fractile of the crack width w. The required value $w_{\lim}$ is considered as a deterministic quantity (for water retaining structures up to $0.2\,mm$).

6.3. Load Combinations

The quasi-permanent combinations of actions are usually considered in design verification of crack width as follows (EN 1990 (2002)):

$$G_k + L_k + \psi Q_k \,. \tag{9}$$

Here G_k denotes the characteristic value of the permanent load G, L_k stands for the characteristic value of the liquid load L (considered similarly as the permanent load, $\mu_L = L_k$), Q_k is the characteristic value of the variable load Q, ψ is the combination factor for the variable load Q. In some cases (for example in case of a wall of water retaining structures) the liquid load L can be considered only (effect of other loads are negligible). In the design verification of ultimate limit states the partial factors for all actions should be considered as recommended in relevant codes, for the liquid load L the partial factor should be considered as $\gamma = 1.2$ as recommended in EN 1992-3 (2006).

7. Probabilistic Formulation

7.1. The Limit State Function

Random behavior of crack width w can be analyzed using equations (5), (6) and (7), where all input quantities are considered as random variables. Equation (5) can be thus written as

$$w = S_r \varepsilon . \tag{10}$$

Here w denotes the crack width, S_r is the crack spacing and ε is the strain as random variable. The crack spacing S_r is assumed to be described by equation (6), the strain ε by equation (7) assuming that all input quantities are considered as random variables having the means equal to nominal values. In equation (7) the lower bound $0.6 \cdot \sigma_s / E_s$ is not considered in the following reliability analysis.

The theoretical model for the strain ε is partly based on experimental observation. Its uncertainty is taken into account by a factor θ expressing model uncertainty. The limit state function g may be then written in a simple form

$$g = w_{\lim} - \theta w . \tag{11}$$

Here the random crack width is given by equation (10), (6) and (7). The model uncertainty θ is introduced as an additional random variable (having the mean equal to unity and the coefficient of variation 10%). In the following analysis the limiting crack width $w_{\lim}$ is considered as a fuzzy random serviceability resistance R analysed above. It is defined by general equation (1) and described by the damage function (2) or the damage density function (3). A particular form of these functions relevant to the foreseen example of water retaining structures is shown in Figure 2.

7.2. Theoretical Models of Basic Variables

All the quantities entering equations (6), (7) and (11) including the model uncertainty θ are in general random variables. Some of them are, however, approximated by deterministic values (those having relatively small variability). Theoretical models (including the type of distribution and their parameters) of all variables used in the following reliability analysis are indicated in Table 1.

Table 1. Theoretical models of basic variables

Name	Symbol X	Unit	Distribution	Char. Value X_k	The mean μ_X	St. dev. σ_X
Width	b	m	Det	1.00	1.00	0
Cover	c	m	Gamma	0.04	0.04	0.01
Reinf. diam.	ϕ	m	Det	0.012 to 0.03	0.012 to 0.03	0
Tensile strength	f_{ct}	MPa	LN	2.9	2.9	0.55
Steel mod.	E_s	GPa	Det	200	200	0
Concrete mod.	E_c	GPa	Det	33	33	0
Creep coeffic.	φ	-	Det	2	2	0
Coefficient	k_1	-	Det	0.8	0.8	0
Coefficient	k_2	-	Det	1	1	0
Coefficient	k_t	-	Det	0.4	0.4	0
Limiting width	w_{lim}	m	Beta*)	0.0000823	0.00015	0.000039
Pressure	L_k	MPa	N	0.07	0.07	0.0035
Diameter	D	m	Det	28	28	0
Action uncer.	θ_E	-	LN	1.00	1.00	0.10

*) Parameters of the Beta distribution are derived from the above fuzzy probabilistic analysis of the limiting crack widths considering the lower limit value of the transition region $0.05\,mm$ and the upper limit $0.2\,mm$. The lower bound of Beta distribution $a = 0.02\,mm$ and the upper bound $b = 0.238\,mm$

The following notations are used in Table 1: Normal - for normal distribution, Gamma - for gamma distribution, LN - for log-normal distribution, Det - for deterministic value. Note that the model uncertainty θ is supposed to cover uncertainties in some variables that are indicated as deterministic quantities.

It follows from Table 1 that the limiting crack width $w_{\lim}$ is approximated by Beta distribution indicated in the above general analysis of fuzzy random properties of the serviceability resistance R.

8. Reliability Analysis

8.1. An Example of Water Reservoir

As an example of probabilistic design for cracking a cylindrical water reservoir with diameter $D = 28\,m$, height $7\,m$ (the maximum water pressure $L_k = 70\,kN/m^2$) and wall thickness $0.25\,m$ is considered (Holický et al. 2009). Crack width is analyzed in the wall only under pure tension due to water pressure. The maximum characteristic force in the wall is thus

$$N_s = D \cdot L_k / 2 = 980\,kN\,. \qquad (12)$$

The basic reinforcement area $A_0 = 0.0027\,m^2$ in the wall is determined considering the ultimate limit state of tensile capacity of the wall using the partial load factor $\gamma = 1.2$, thus the design force in the wall is $N_d = \gamma N_s = 1{,}176\,kN$.

It is common that the basic reinforcement A_0 must be increased to an acceptable value A in order to control cracking. For the data given in Table 1 the enhancement factors given by ratio A/A_0 follow from general equations (5) to (7). For the deterministic design for crack width control according to EN 1992-1-1 (2003) the enhancement of the deterministic crack limit $w_{\lim} = 0.20$ is more than a factor of 2, and for the crack limit $w_{\lim} = 0.05$ it is

more than 5, depending on the steel diameter. In the following analysis these outcomes of the deterministic calculation are compared with results of probabilistic analysis.

8.2 Probabilistic Analysis

Crack width of the reservoir wall exposed to pure tension is analyzed considering the limit state function (11) and theoretical models of basic variables given in Table 1. Various diameters of the reinforcing bars ϕ (from 12 to $30\,mm$) are considered. The limiting crack width $w_{\lim}$ is generally described by the Beta distribution. Note, however, that this approximation is derived from deterministic limiting crack widths $w_{\lim} = 0.05\,mm$ and $w_{\lim} = 0.20\,mm$. Figure 3 shows the variation of the failure probability with increasing area A/A_0 within the broad range from 1 to 5.

It follows from Figure 3 that without substantial enhancement of the reinforcement area the crack width would exceed the required limiting width $w_{\lim}$ with a very high probability. The basic reinforcement area A_0 should be increased approximately by the factor of 2 to comply with the required crack width. Figure 3 also indicates the fuzzy probability of failure $\pi_f = 0.05$ accepted in EN 1992-1-1 (2004) for verification of serviceability limit states including crack widths.

Desired enhancement of the reinforcement depends obviously on the reinforcing bars´ diameter ϕ. Figure 3 shows that for $\phi = 12\,mm$ the reinforcement ratio A/A_0 should be about 2.3, for $\phi = 30\,mm$ the reinforcement ratio A/A_0 should be about 3.2. This finding induces a crucial question concerning the required reliability level. In some cases the reliability may be decreased (target failure probability increased), in other cases (for example in case of a vital reservoir) it may be increased (target failure probability decreased). It appears that the methods of probabilistic optimization may provide a valuable guidance.

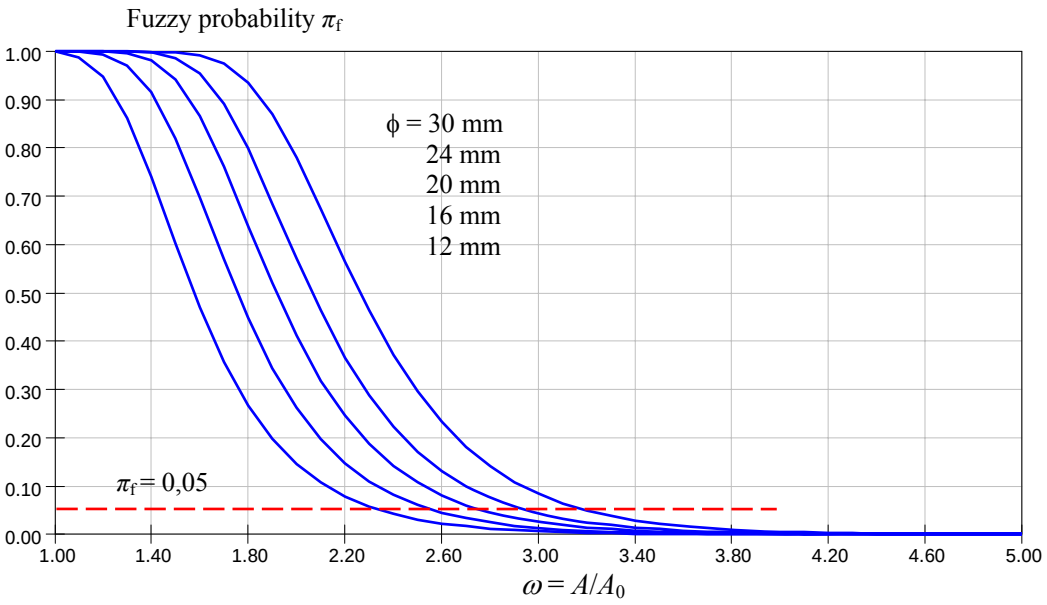


Figure 3. Variation of the probability of failure with the reinforcement area ratio $\omega = A/A_0$ for selected reinforcement diameter ϕ

9. Probabilistic Optimization

Probabilistic optimization may be effectively used to specify the optimum value of some basic variables (decisive parameters) and the target reliability of a structure. In some cases of the design of a concrete structure for cracking the objective function may be written in a simple form as the total cost

$$C_{tot}(\omega) = C_0 + C_1\omega + C_f\pi_f(\omega), \tag{13}$$

where $C_{tot}(\omega)$ denotes the total cost,

C_0 - the initial cost,

C_1 - the margin cost per unit of the decisive parameter ω,

C_f - the discounted cost serviceability failure,

$\pi_f(\omega)$ - fuzzy probability of failure,

ω - the decision parameter.

Here the initial cost C_0 is assumed to be independent of parameter ω. The product $C_1\omega$ denotes the additional cost due to an increase of parameter ω and $C_f\pi_f(\omega)$ the expected malfunctioning cost. The discounted cost of serviceability failure C_f takes into account the time when the crack width w exceeds the limit value $w_{\lim}$. The probability of failure $\pi_f(\omega)$ is considered as a function of the parameter ω. Instead of the total cost $C_{tot}(\omega)$ given by equation (13) the normalized $\kappa_{tot}(\omega)$ may be considered

$$\kappa_{tot}(\omega) = [C_{tot}(\omega) - C_0]/C_1 = \omega + \pi_f(\omega)C_f/C_1 . \tag{14}$$

It follows from the first derivative of $\kappa_{tot}(\omega)$ that the necessary condition for the optimum parameter ω_{opt} can be written as

$$\partial P_f(\omega)/\partial\omega = -C_1/C_f . \tag{15}$$

In the design of a concrete structure for cracking the reinforcement area A is optimized. A generic value of A may be introduced through the reinforcement ratio $\omega = A/A_0$, where the basic value A_0 is given by the ultimate limit state design. The optimum value $\omega_{opt} = A_{opt}/A_0$ can be assessed from the minimum of the standardized cost $\kappa_{tot}(\omega)$ given by equation (14) or directly from the necessary condition (15).

Figure 4 shows the variation of the reliability index β and the total standardized costs $\kappa_{tot}(\omega)$ given by equation (14) with the reinforcement area ratio $\omega = A_s/A_{s0}$. It follows from Figure 4 that the optimum parameter ω_{opt} increases with the cost ratio C_f/C_1; for $C_f/C_1 = 1$, $\omega_{opt} = 1.0$, for $C_f/C_1 = 1{,}000$, $\omega_{opt} = 4.3$. Thus, in general, the reinforcement area A needs to be substantially increased to reach the minimum total cost $\kappa_{tot}(\omega)$.

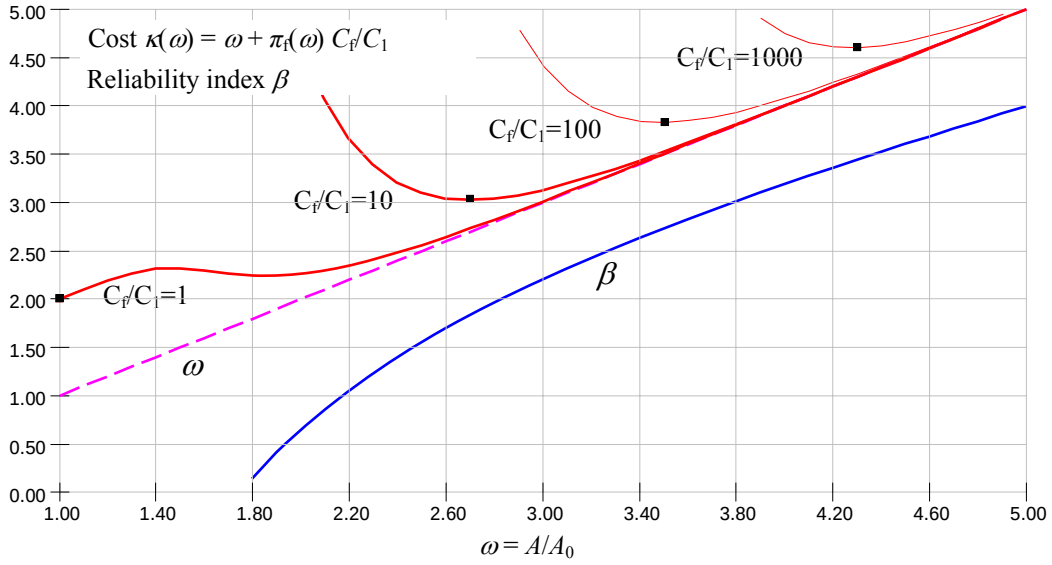


Figure 4. Variation of the total normalized cost $\kappa(\omega)$ and reliability index β with the reinforcement ratio ω for the reinforcement diameter $\phi = 16\,mm$

Figure 4 also indicates that the reliability index β is within a broad interval from 0 to 3.5. Thus, for high costs of serviceability failure ($C_f/C_1 = 1{,}000$) the optimum reliability levels reach the commonly recommended levels for the ultimate limit states. Obviously, an appropriate reliability level depends on the cost ratio C_f/C_1, which has to be assessed taking into account specific conditions of a particular structure.

10. Concluding Remarks

(1) Performance requirements are commonly specified by quantities of a random and vague nature.

(2) The damage function and damage density functions are basic fuzzy probabilistic tools for the description of random and vague serviceability requirements.

(3) The characteristic value or serviceability requirements may be defined as a fractile of the fuzzy probabilistic distribution, the probability $p = 0.05$ may not be the optimum value.

(4) The fuzzy probabilistic distribution of serviceability requirements may be used similarly as a classical distribution function to analyze reliability level and to optimize structural design.

(5) For water retaining structures the basic reinforcement area A (given by ultimate limit states) needs to be substantially increased to reach the limiting crack width.

(6) The optimum parameter $\omega_{opt} = A_{opt}/A_0$ increases with the cost ratio C_f/C_1 of the malfunctioning cost C_f to the cost per unit of the reinforcement area C_1 (for $C_f/C_1 = 1{,}000$, the optimum ω_{opt} can be about 4.0).

(7) The optimum reliability index can be expected within the broad interval from 0 to 3.5 depending on the cost ratio C_f/C_1.

(8) Further research should be focused also on the assessment of economic consequences of a serviceability failure when the crack width exceeds the limiting value.

References

1. Brown, C.B. and Yao, J.T.P. **Fuzzy sets and structural engineering**, Journal of Structural Engineering, 109(5), 1983, pp. 1211-1225
2. EN 1990 **Eurocode: Basis of structural design**, April 2002
3. EN 1992-1-1 **Eurocode 2: Design of concrete structures, General rules and rules for building**, December 2004
4. EN 1992-3 **Eurocode 2: Design of concrete structures, Part 3: Liquid retaining and containment structures**, June 2006
5. Holický, M. **Fuzzy probabilistic models in structural reliability**, Maintenance and Reliability, 2(30), 2006, pp. 11-13
6. Holický, M., Retief, J. and Wium, J. **Probabilistic design for cracking of concrete structures**, In: Van Gelder, Proske & Vrijling (eds.), Proceedings of the 7th International Probabilistic Workshop, 2009, Delft
7. Narayanan, R.S. and Beeby, A. **Designers Guide to EN 1992-1-1 and EN 1992-1-2**, Thomas Telford, 2005

[1] **Acknowledgement**
This study is a part of the project GACR 103/09/0693 „Safety and Risk Assessment of Technical Systems" supported by The Grant Agency of the Czech Republic.

JOURNAL OF APPLIED QUANTITATIVE METHODS

International Symposium on Stochastic Models in Reliability Engineering, Life Sciences and Operations Management (SMRLO'10)

PERINATAL ASSISTANCE NETWORK PLANNING VIA SIMULATION

Paola FACCHIN
Prof, Epidemiology and Community Medicine Unit,
Department of Paediatrics,
University of Padova,
Padova, Italy

E-mail: epi@pediatria.unipd.it

Anna FERRANTE
PhD, Epidemiology and Community Medicine Unit,
Department of Paediatrics,
University of Padova,
Padova, Italy

E-mail: epi@pediatria.unipd.it

Elena RIZZATO
PhD, Epidemiology and Community Medicine Unit,
Department of Paediatrics,
University of Padova,
Padova, Italy

E-mail: epi@pediatria.unipd.it

Giorgio ROMANIN-JACUR
Prof, Department of Management and Engineering,
University of Padova,
Vicenza, Italy

E-mail: romjac@dei.unipd.it

Laura SALMASO
PhD, Epidemiology and Community Medicine Unit,
Department of Paediatrics,
University of Padova,
Padova, Italy

E-mail: epi@pediatria.unipd.it

Abstract: *Consider a geographical region where population is distributed in health districts, and there exists a neonatal care network, which includes birth centres able to supply assistance at three levels, respectively basic assistance, mild pathology care and intensive care. Each mother-to-be is admitted to a facility where the assistance level corresponds to the expected newborn conditions; newborn transfers from a lower to a higher-level facility are affected if conditions worsen. Each district has a known probabilistic demand for each care level previously mentioned and each facility is characterized by its capacity, i.e., the amount of patients simultaneously admittable there. A simulation model describing mothers and newborns movements from districts to birth centres and among centres has been built up, with the aim of revealing inadequacies of the assistance network and of obtaining useful suggestions about network resizing to improve service quality and reduce trouble due to distance. The model has been applied to Veneto region in North-East Italy but its use may be extended to other similar situations.*

Key words: *perinatal assistance network; decision support; discrete stochastic simulation*

1. Introduction

In recent years numerous clinical and technological advances have brought about dramatic improvements in neonatal care, enabling newborns that are severely premature or born in critical conditions to be kept alive. For this small group of newborns few specialized facilities, able to supply neonatal intensive care, are activated to serve large catchment areas. Alongside this restrict group of patients, there is a large group of newborns coming into the world in physiological or mildly pathological conditions, who can be cared for at less specialized facilities and by less expert staff. It is more convenient to distribute such facilities on the territory so to serve smaller catchment areas, and to contain the distances between the facilities and the population residential areas they serve. To satisfy all the above requirements demands the creation of a birth assistance network with facilities of different levels, moreover transfers needed from lower level facilities to higher level facilities for patients whose initial conditions deteriorate shall be provided. Three-tiered perinatal care networks have been planned in Europe and in USA ever since the '80s (Brann et al. 1980; Le Roy et al. 2006; Van Reempts et al. 2007). The three levels correspond to basic neonatal care on the first level, intermediate care for neonatal diseases on the second and neonatal intensive care on the third. Each perinatal care network has to be accurately planned in logistic terms, providing suitably sized and connected facilities for ensuring the maximum efficiency of each level of care.

In this paper we simulate movements of patients from the territory towards the assistance network and inside the network. We assume that mothers-to-be live in health districts dotted all over a given territory; every mother-to-be is admitted to a birth centre serving the level corresponding to the needs of her pregnancy; therefore each district has a certain demand for each of the three perinatal care levels previously mentioned. Within the considered territory there is a service network that includes facilities for each level of neonatal care, and a characteristic of the network lies in that for every higher level facility also lower levels facilities are provided in the same place. Each facility is characterized by a capacity, i.e., the amount of patients who can be simultaneously admitted: more precisely the capacity of the third level of care is given by the number of ventilated incubators

available and the capacity of the second level by the number of incubators; the capacity related to all the three care levels combined emerges from the numbers of beds available for the mothers. As seen above, every mother-to-be should be admitted to a facility on a level appropriate to the foreseen newborn conditions; if places are available, she is admitted to the closest facility with respect to her district, otherwise to another facility chosen according to increasing distance order. At birth the newborn may present unexpected complications and require urgent transfer to a centre providing a higher level of care, at the same location or elsewhere, depending on the availability of levels of care and vacant places.

The aim of the simulation lies in i) checking the patients distribution among the existing facilities, ii) checking the assistance network ability of supplying a suitable service, in terms of admitting all patients sufficiently close to their home, and iii) possibly suggesting convenient adaptations in case of inadequate service.

In health care literature, and particularly for what concerns health care services planning, the problem of facilities location-allocation was discussed in (Toregas et al. 1971; Branas et al. 2000; Takinawa et al. 2006; Sahin et al. 2007; Mitropulos et al. 2006; Ratick et al. 2008;) with an optimization approach. The problem of perinatal facility planning was studied by (Galvão et al. 2002; Boffey et al. 2003; Galvão et al. 2006). All above papers solve the planning problem by means of optimization but the solution is given as the average, in other words the mean of admitted patients is considered. In this paper the detailed movement of newborns is obtained revealing all peak conditions. The simulation results report: patients admitted far from home, patients admitted out of the region, transfers among hospitals because of missing places. From the results useful suggestions may be obtained about suitable network resizing in order to improve service quality and reduce trouble due to distance and related costs.

2. Birth Assistance and Birth Centres

Pregnancy may take a physiological or a pathological course. Complications in pregnancy may include gestational diabetes, infections, foetal malformations, and so on. Such complications may cause problems during delivery and/or in the neonatal period, so a higher level of care should preferably be chosen in advance for such cases (Le Roy et al. 2006; Eberhard et al. 2008; Mayfield et al. 1990).

Birth may take a physiological or a pathological course too. Some complications, such as those due to foetal-pelvic disproportion, placental deformity, etc., may be foreseen, so a higher level of care can be planned. Others, e.g. foetal distress, haemorrhage, premature birth, etc., may be unexpected and require prompt referral to a higher level of care.

Mothers are assisted during delivery by a gynaecological-obstetric team and newborns by a paediatric-neonatology team. The care provided is generally classified on three levels, i.e. neonatal basic care, intermediate care for neonatal diseases and neonatal intensive care; such a classification is widely accepted in Europe and the USA, as reported in (Brann et al. 1980; Le Roy et al. 2006; Van Reempts et al. 2007; Zeitlin et al. 2004; Committee of fetus and newborn. 2004).

The main characteristics of each level of care are recalled below:

- the first level of care is provided in the case of uncomplicated pregnancies and is characterized by the following capabilities:

- assistance for the mother: continuous specialist assistance can be provided;
- assistance for the newborn: continuous neonatological assistance can be provided, a neonatological unit can be used for primary resuscitation;

- the second level of care is provided in the event of pregnancies and births at low risk and/or involving mild prematurity (pregnancies shorter than 30 weeks or a birth weight below 1.5 kg), and is characterized by the following capabilities:

- assistance for the mother: continuous specialist assistance can be provided, specialist monitoring equipment may be used during labour;
- assistance for the newborn: continuous neonatological assistance can be provided, a neonatological unit can be used for resuscitation, and incubators may be used;

- the third level of care is provided in the case of higher-risk pregnancies and/or severely premature births (before 28 weeks of gestation, or birth weights below 1.0 kg), and is characterized by the following capabilities:

- **assistance for the mother: continuous specialist assistance, with anaesthesiological, intensive care and other specialist consultants available; specialist monitoring apparatus can be used during labour;**
- **assistance for the newborn: continuous neonatological assistance, intensive care with artificial ventilation, and specialist consultants available; a neonatal intensive care ward is used.**

Each birth centre may be associated with various levels of care. Most centres provide only basic care, while a few specialist centres provide both first and second levels of care, and only a very few highly-specialized centres provide all levels of care.

Clearly, every pregnant woman should generally be admitted to a birth centre where the level of care is appropriate for any expected birth complications; in other words, a "level requirement" may be defined

3. The Problem

Let us consider a birth centre and all the levels of care it can provide, bearing in mind that there are three levels of neonatal care and, for each of these, the capacity in terms of births per year can be established from the distribution of the duration of the users' stay. More precisely, the capacity related to all care levels combined emerges from the number of beds available for mothers in the obstetric ward. In addition, the capacity of the third level of care is limited by the number of artificially-ventilated incubators available and the capacity of the second level by the number of incubators, while for the first level of care there are no such structural limitations. The site and size of the birth centres are fixed a priori. Recommendations for adjusting the size of a given centre may emerge from the solution of the model.

Let us consider a health district, i.e. a homogeneous built-up area that can be seen as a point in relation to distances. For each district and each level of care, we define the corresponding demand in terms of births per year (this demand can be extrapolated from the statistics of previous years).

As seen in the previous section, a pregnant woman should be admitted to a centre and occupy a place on a level appropriate to her needs. Let us consider a situation where she has occupied a place where first or second-level care is provided but her newborn

presents unexpected complications and requires urgent transfer to a centre providing a higher level of care, at the same location or elsewhere, depending on the availability of levels of care and vacant places. In such a case, the newborn is transferred while the mother remains at the centre where she gave birth. Lower-level birth centres may thus become a source of further requests for the admission of newborn, adding to the above-mentioned requests coming directly from the urban nuclei.

4. The Model

Admission requests rise in a random independent way for every level and for every district, according to statistics extrapolated from past data. Every request is dispatched to a facility of the corresponding level, more precisely the closest one with available places both for the mother and the newborn. If the newborn conditions worsen then he/she is transferred to the closest higher-level facility, without moving the mother who remains in the first admission facility. The ratio of newborns needing transfer between different levels is statistically known from the past. The length of stay distributions are shown to be gamma functions with parameters obtained from past data.

All distributions parameters are obtained by elaborating data from patient discharge papers, which are compiled in correspondence to every patient exit from hospital and therefore for every newborn.

The model evidences the saturation of all facilities and therefore their bed occupancy ratio. Moreover all network malfunctions may be revealed, more precisely the amount of patients that are admitted far from home and the amount of newborns that are transferred to another facility because of missing places.

4.1. Model Implementation

The model is quite simple, therefore it can be implemented using many simulation languages. Here it was implemented in MicroSaint Student, which was at disposition, but it may be easily translated in any language similar to SIMAN-Arena. MicroSaint implementation has a graphical representation that is based on four basic elements: tasks, describing activities, arrows, representing activity sequences, rhombuses, representing decisions and striped rectangles, representing queues.

Our model implementation is reported in Figure 1, where task 1 is employed to start simulation, tasks 2-4 report all parameters utilized by the model, more precisely demand rate for each level and each district, all facility capacities and all distances (both district to birth centre and centre to centre). Tasks 5-7 are demand generators for the three levels; the arrival rate is the sum of all districts arrival rates while the district is probabilistically stated. Tasks 8-10 decide the dispatching of mothers and newborns and the possible newborns transfers. Task 11 represents mothers and newborns admitted to

facilities out of the region, tasks 12-13 manage the dispatching of transferred newborns. Task 14-16 are necessary for functional aims and, finally, tasks 17-19 respectively describe mothers admissions and second and third level newborns admissions.

Note that patients' movements from all districts towards all facilities and among facilities have been described by means of a small model.

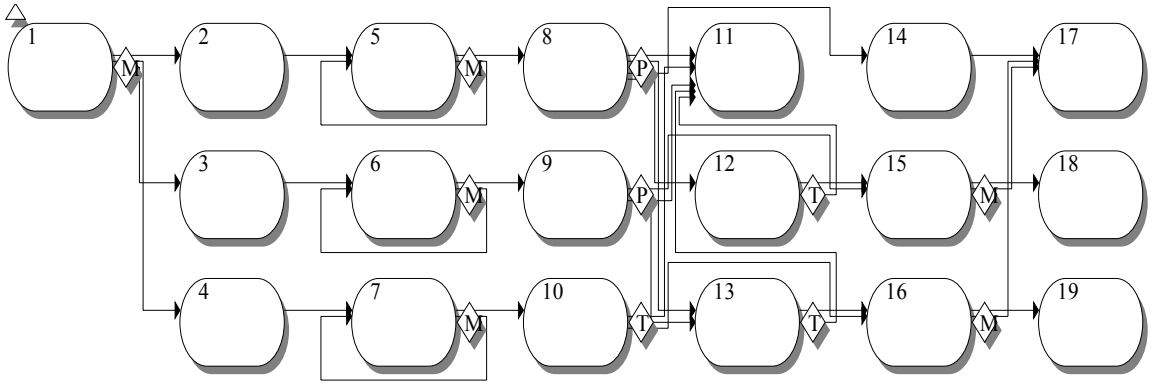


Figure 1. The model

4.2. Model Application and Results

The model was applied to the Veneto Region in North-East Italy, a region with a population of about 4,700,000 and about 43,000 newborns per year. In the region there exists 52 health districts and 41 birth centres, all of which have both first facilities, for a total of 632 places, and second level facilities, with a total of 245 places, while only 9 of them have the third level facilities with a total of 57 places. The length of stay is a gamma distribution with a mean of respectively 3.6 days for the first level, 6.5 days for the second level, 17.6 days for the third level. All the above data have been obtained from the regional statistical office.

The simulation reveals a lot of interesting results about the patients' movements and beds saturation. The saturation of beds for the mothers is 58.5%, the saturation of second level places is 61% and the saturation of third level places is over 99%. No first level newborn needs to be admitted out of the region, but the 21% are not admitted to the first choice birth centre and the 59.6% are constrained to be admitted to a facility over 15 Km far from home: that evidences that the first level assistance is over dimensioned but badly distributed on the territory; the same happens for the second level, for which only the 1,5% needs to be admitted out of the region. 47% patients cannot go to the closest centre and 25% are admitted to a facility 25 Km far from home. For what concerns the third level, we see that 64% are admitted out of the region and that evidences that the third level assistance is absolutely insufficient. If an improvement is to be obtained that shall be performed by means of an increase of places for the third level; a mild reduction of beds for the mothers and of second level places may be accepted; the first action requires an expansion of existing units, that means an increase of both ventilated incubators and specialized personnel; the second action may be simple. Obviously new distributions of birth centres on the territory would be desirable and may be easily simulated but that is much more complex to obtain in practice.

5. Conclusions

We have built up and implemented a simulation model describing in detail the movements of mothers-to-be and newborns among health districts and assistance facilities, where assistance facilities are classified according to three different levels corresponding to the severity of newborn conditions. The aim of the model lies in checking system effectiveness and efficiency in providing adequate care to patients close enough to their

home. Simulation results permit to reveal assistance lacks and to suggest suitable correcting actions. The model has been implemented on a personal computer and applied to Veneto region but it can be easily applied to other Italian or foreign regions.

References

1. Boffey, B., Yates, D. and Galvão, R.D. **An algorithm to locate perinatal facilities in the municipality of Rio de Janeiro**, Journal of the Operational Research Society, 54, 2003, pp. 21-31
2. Branas, C.C., MacKenzie, E.J. and ReVelle, C.S. **A trauma resource allocation model for ambulances and hospitals**, Health Services Research, 35(2), 2000, pp. 489-507
3. Brann, A.W., Hall, R.T., Harper, R.G., Maisels, J., Poland, R.L., Rhodes, P.G. et al. **Level II neonatal units**, Paediatrics, 66(5), 1980, pp. 810-811
4. Committee of fetus and newborn. **Levels of neonatal care**, Paediatrics, 114, 2004, pp. 1341-1347
5. Galvão, R.D., Espejo, L.G.A. and Boffey, B. **A hierarchical model for the location of perinatal facilities in the municipality of Rio de Janeiro**, European Journal of Operational Research, 138, 2002, pp. 495-517
6. Galvão, R.D., Espejo, L.G.A. and Boffey, B. **Practical aspects associated with location planning for maternal and perinatal assistance**, Annals of Operational Research, 143, 2006, pp. 31-44
7. Eberhard, A., Wenzlaff, P., Lack, N., Misselwitz, B., Kaiser, A. and Bartels, D.B. **Federal admission criteria for levels of perinatal care: Definition, interpretation and first conclusions**, Zeitschift fur Geburtshilfe und Neonatologie, 212(3), 2008, pp. 100-108 (in German)
8. Le Roy, C., Carayol, M., Zeitlin, J., Breart, G. and Goffinet, F. **Level of perinatal care of the maternity unit and rate of cesarean in low risk nulliparas, Obstetrics & Gynecology**, 107(6), 2006, pp. 1269-1277
9. Mayfield, J.A., Rosenblatt, R.A., Boldwin, L.M., Chu, J. and Logerfo, J.P. **The relation of obstetrical volume and nursery level to perinatal mortality**, American Journal of Public Health, 80(7), 1990, pp. 819-823
10. Mitropoulos, P., Mitropoulos, I., Giannikos, I. and Sissouras, A. **A biobjective model for the locational planning of hospitals and health centers**, Health Care and Management Science, 9, 2006, pp. 171-179
11. Ratick, S., Osleeb, J.P. and Hozumi, D. **Application and extension of the Moore and ReVelle hierarchical maximal covering model**, Socio-Economic Planning Sciences, 30, 2008, pp. 1-10
12. Sahin, G., Sural, H. and Meral, S. **Locational analysis for regionalization of Turkish red crescent blood services**, Computer & Operations Research, 34, 2007, pp. 692-704
13. Tanikawa, T., Ohba, H., Terashita, T., Uesugu, M., Jiang, G., Ogasawara, K. and Sakurai, T. **Model analysis for optimal allocation of pediatric emergency center**, AMIA Annual Symposium Proceeding, 2006, p.1115
14. Toregas, C., Swain, R., ReVelle, C. and Bergman, L. **The location of emergency service facilities**, Operations Research, 19(6), 1971, pp. 1363-1373
15. Van Reempts, P., Gortner, L., Milligan, D. et al. **Characteristics of neonatal units that care for very preterm infants in Europe: Result from the Mosaic study**, Paediatrics, 120(4), 2007, pp. 815-825
16. Zeitlin, J., Papiernik, E. and Breart, G. **Regionalization of perinatal care in Europe**, Seminars in Neonatology, The Europet group, 9, 2004, pp. 99-110

JOURNAL OF APPLIED QUANTITATIVE METHODS

International Symposium on Stochastic Models in Reliability Engineering, Life Sciences and Operations Management (SMRLO'10)

SEMI-MARKOV RELIABILITY MODEL OF THE COLD STANDBY SYSTEM

Franciszek GRABSKI
Prof, Department of Mathematics and Physics,
Polish Naval University,
Gdynia, Poland

E-mail: franciszekgr@onet.eu

Abstract: *The semi-Markov reliability model of the cold standby system with renewal is presented in the paper. The model is some modification of the model that was considered by Barlow & Proshan (1965), Brodi & Pogosian (1978). To describe the reliability evolution of the system, we construct a semi-Markov process by defining the states and the renewal kernel of that one. In our model the time to failure of the system is represented by a random variable that denotes the first passage time from the given state to the subset of states. Appropriate theorems from the semi-Markov processes theory allow us to calculate the reliability function and mean time to failure. As calculating an exact reliability function of the system by using Laplace transform is often complicated we apply a theorem which deals with perturbed semi-Markov processes to obtain an approximate reliability function of the system.*

Key words: *semi-Markov process; perturbed process; reliability model; renewal standby system*

1. Description and Assumptions

We assume that the system consists of one operating series subsystem (unit), an identical stand-by subsystem and a switch (see Figure 1):

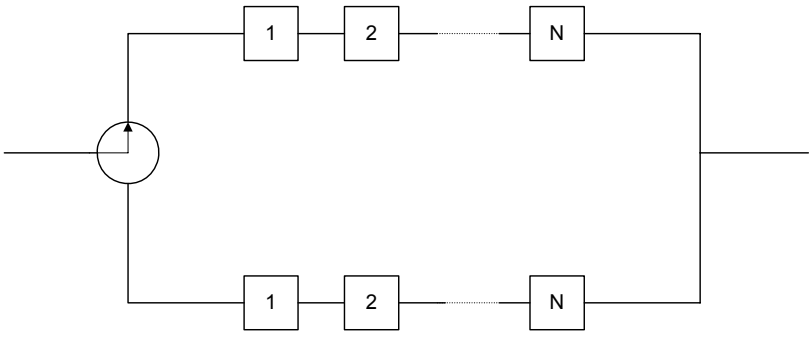


Figure 1. Diagram of the system

Each subsystem consists of N components. We assume that time to failure of those elements are represented by non-negative mutually independent random variables ζ_k, $k=1,...,N$, with distributions given by probability density functions $f_k(x)$, $x \geq 0$, $k=1,...,N$. When the operating subsystem fails, the spare is put in motion by the switch immediately. The failed subsystem is renewed. There is a single repair facility. A renewal time is a random variable having distribution depending on a failed component. We suppose that the lengths of repair periods of units are represented by identical copies of non-negative random variables γ_k, $k=1,...,N$, which have cumulative distribution functions $H_k(x)=P(\gamma_k \leq x)$, $x \geq 0$. The failure of the system occurs when the operating subsystem fails and the subsystem that has sooner failed in not still renewed or when the operating subsystem fails and the switch also fails. Let U be a random variable having binary distribution

$$b(k) = P(U = k) = a^k(1-a)^{1-k}, k = 0, 1,\ 0 < a < 1,$$

where $U=0$, if a switch is failed at the moment of the operating unit failure, and $U=1$, if the switch works at that moment. We suppose that the whole failed system is replaced by the new identical one. The replacing time is a non negative random variable η with CDF

$$K(x) = P(\eta \leqslant x),\ x \geqslant 0.$$

Moreover, we assume that all random variables mentioned above are independent.

2. Construction of Semi-Markov Reliability Model

To describe the reliability evolution of the system, we have to define the states and the renewal kernel. We introduce the following states:

0 – failure of the system;

k – renewal of the failed subsystem after a failure of k -th, $k=1,...,N$, component and the work of a spare unit

$N+1$ – both an operating unit and a spare are "up".

The scheme shown in Figure 2 presents functioning of the system. Let $0=\tau_0^*,\tau_1^*,\tau_2^*$- denote the instants of the states changes, and $\{Y(t): t \geq 0\}$ be a random process with the state space $S=\{0,1,...,N,N+1\}$, which keeps constant values on the half-intervals $[\tau_n^*, \tau_{n+1}^*), 0,1,...$, and is right-hand continuous. This process is not a semi-Markov one, as no memory property is satisfied for any instants of the state changes of that one.

Let us construct a new random process in a following way. Let $0=\tau_0$ and $\tau_1,\tau_2,...$ denote instants of the subsystem failures or instants of the whole system renewal.

The random process $\{X(t): t \geq 0\}$ defined by equation

$$X(0)=0, \quad X(t)=Y(\tau_n) \quad \text{for} \quad t \in [\tau_n,\ \tau_{n+1}) \tag{1}$$

is the semi-Markov one.

To have a semi-Markov process as a model we have to define its initial distribution and all elements of its kernel. Recall that the semi-Markov kernel is the matrix of transition probabilities of the Markov renewal process

$$\boldsymbol{Q}(t) = [Q_{ij}(t) : \ i, j \in S], \tag{2}$$

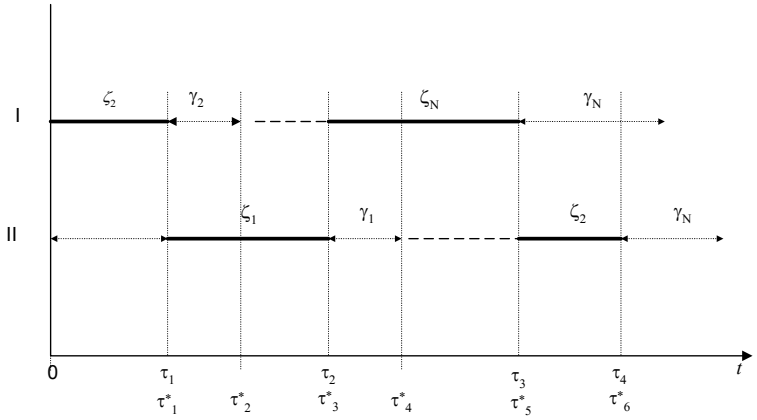


Figure 2. Reliability evolution of the standby system

where

$$Q_{ij}(t) = P(\tau_{n+1} - \tau_n \leqslant t, \ X(\tau_{n+1}) = j \,|\, X(\tau_n) = i), \quad t \geqslant 0. \tag{3}$$

From the definition of semi-Markov process it follows that the sequence $\{X(\tau_n) : n = 0,1,...\}$ is a homo-geneous Markov chain with transition probabilities

$$p_{ij} = P(X(\tau_{n+1}) = j \,|\, X(\tau_n) = i) = \lim_{t\to\infty} Q_{ij}(t). \tag{4}$$

The function

$$G_i(t) = P(\tau_{n+1} - \tau_n \ \leqslant t \,|\, X(\tau_n) = i) = \sum_{j\in S} Q_{ij}(t) \tag{5}$$

is a cumulative probability distribution of a random variable T_i that is called a waiting time of the state i. The waiting time T_i is the time spent in state i when the successor state is unknown. The function

$$F_{ij}(t) = P(\tau_{n+1} - \tau_n \ \leqslant t \,|\, X(\tau_n) = i, \ X(\tau_{n+1}) = j) = \frac{Q_{ij}(t)}{p_{ij}} \tag{6}$$

is a cumulative probability distribution of a random variable T_{ij} that is called a holding time of a state i, if the next state will be j. From here we have

$$Q_{ij}(t) = p_{ij}F_{ij}(t). \tag{7}$$

It follows from that a semi-Markov process with a discrete state space can be defined by the transition matrix of the embedded Markov chain: $P=\left[p_{ij} : i, j \in S\right]$ and a matrix of CDF of holding times: $F(t)=\left[F_{ij}(t) : i, j \in S\right]$.

In this case semi-Markov kernel has a form

$$\boldsymbol{Q}(t) = \begin{bmatrix} 0 & 0 & \cdots & 0 & Q_{0N+1}(t) \\ Q_{1\,0}(t) & Q_{1\,1}(t) & \cdots & Q_{1\,N}(t) & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Q_{N\,0}(t) & Q_{N\,1}(t) & \cdots & Q_{N\,N}(t) & 0 \\ Q_{N+1\,0}(t) & Q_{N+1\,1}(t) & \cdots & Q_{N+1N}(t) & 0 \end{bmatrix}. \tag{8}$$

The semi-Markov $\{X(t): t \geq 0\}$ will be defined if we define all elements of the matrix $\boldsymbol{Q}(t)$.

For $j = 1,...,N$ we obtain

$$Q_{N+1\,j}(t) = P(X(\tau_{n+1}) = j,\ \tau_{n+1} - \tau_n \leqslant t \,|\, X(\tau_n) = N+1) =$$

$$= P(A, \zeta_j \leqslant t, \zeta_i > \zeta_j \ \text{ for } \ i \neq j) = a \underbrace{\int\int \ldots \int}_{D_{N+1\,j}} dF_1(x_1) dF(x_2) \cdots dF_N(x_N),$$

where

$$D_{N+1\,j} = \{(x_1, x_2, \ldots, x_N) :\ 0 \leqslant x_j \leqslant t, x_i > x_j,\ i \neq j\}.$$

Using Fubini theorem we obtain

$$Q_{N+1\,j}(t) = a \int_0^t \prod_{i \neq j}^{N} [1 - F_i(x)] f_j(x) dx. \tag{9}$$

For $j = 0$ we have

$$Q_{N+1\,0}(t) = P(X(\tau_{n+1}) = 0,\ \tau_{n+1} - \tau_n \leqslant t \,|\, X(\tau_n) = N+1) =$$

$$= P(A,\ \min(\zeta_1, \ldots, \zeta_N) \leq t) = (1-a)(1 - \prod_{i=1}^{n}(1 - F_i(t))). \tag{10}$$

For $i, j = 1,...,N$ we get

$$Q_{ij}(t) = P(A, \zeta_j \leqslant t, \zeta_k > \zeta_j \text{ for } j \neq k, k = 1, \ldots, N, \gamma_i < \zeta_j).$$

The same way we obtain

$$Q_{ij}(t) = a \int_0^t H_i(x) \prod_{k \neq j}^{N} [1 - F_k(x)] f_j(x) dx. \tag{11}$$

For $i = 1,...,N$ and $j = 0$ we have

$$Q_{i0}(t) = P(\min(\zeta_1, \ldots, \zeta_N) \leqslant t,\ \gamma_i > \min(\zeta_1, \ldots, \zeta_N)) +$$
$$+ P(A',\ \min(\zeta_1, \ldots, \zeta_N) \leqslant t,\ \gamma_i < \min(\zeta_1, \ldots, \zeta_N)) =$$
$$= F(t) - a \int_0^t H_i(x) dF(x), \tag{12}$$

where

$$F(x) = P(\min(\zeta_1, \ldots, \zeta_N) \leqslant t) = 1 - \prod_{k=1}^{N} [1 - F_k(x)]. \tag{13}$$

From the assumption it follows that

$$Q_{0\,N+1}(t) = K(t). \tag{14}$$

All elements of the kernel $Q(t)$ have been defined, hence the semi-Markov process $\{X(t): t \geq 0\}$ describing reliability of the renewal cold standby system has been constructed.

3. Exponential Time to Failure of Elements

Assuming the exponential time to failure of elements we obtain a special case of the model. Suppose that random variables ζ_k, $k = 1,...,N$ are exponentially distributed with parameters λ_k, $k = 1,...,N$, correspondingly. Hence

$$f_k(x) = \lambda_k e^{-\lambda_k x}, \quad x \geqslant 0.$$

Because of the no memory property of the exponential distribution, the assumption concerning of the whole subsystem renewal can be substituted by the assumption concerning failed element renewal.

In this case we obtain

$$Q_{N+1\,j}(t) = a\int_0^t \prod_{i\neq j}^{N} [e^{-\lambda_i x}]\lambda_j e^{-\lambda_j x} dx = a\frac{\lambda_j}{\Lambda}(1 - e^{-\Lambda t}), \quad t \geqslant 0 \tag{15}$$

for $j = 1,...,N$, where

$$\Lambda = \lambda_1 + \ldots + \lambda_N.$$

For $j = 0$ we obtain

$$Q_{N+1\,0}(t) = (1-a)(1 - e^{-\Lambda t}), \quad t \geqslant 0. \tag{16}$$

For $i, j = 1,...,N$

$$Q_{ij}(t) = a\lambda_j \int_0^t H_i(x) e^{-\Lambda x} dx. \tag{17}$$

For $j = 0$

$$Q_{i0}(t) = 1 - e^{-\Lambda t} - a\Lambda \int_0^t H_i(x) e^{-\Lambda x} dx. \tag{18}$$

4. Approximate Model

For simplicity we consider an approximate model. We can assume that the renewal time of the subsystem is a random variable γ having CDF

$$H(x) = \sum_{i=1}^{N} q_i H_i(x) \quad \text{where} \quad q_k = \frac{E(\gamma_k)}{\sum_{i=1}^{N} E(\gamma_i)}. \tag{19}$$

This way we obtain 3-state semi-Markov process with kernel

$$\boldsymbol{Q}(t) = \begin{bmatrix} 0 & 0 & Q_{02}(t) \\ Q_{10}(t) & Q_{11}(t) & 0 \\ Q_{20}(t) & Q_{21}(t) & 0 \end{bmatrix}, \tag{20}$$

where

$$Q_{02}(t) = K(t), \tag{21}$$

$$Q_{10}(t) = F(t) - a\int_0^t H(x)dF(x), \quad Q_{11}(t) = a\int_0^t H(x)dF(x), \tag{22}$$

$$Q_{20}(t) = (1-a)F(t), \quad Q_{21}(t) = aF(t).$$

Assume that, the initial state is 2. It means that an initial distribution is

$$\boldsymbol{p}(0) = [\,0 \quad 0 \quad 1\,]. \tag{23}$$

Hence, the semi-Markov model has been constructed.

5. Reliability Characteristics

A value of a random variable

$$\Delta_A = \min\{n \in \mathbb{N}:\ X(\tau_n) \in A\} \tag{24}$$

denotes a discrete time (a number of state changes) of *a first arrival* at the set of states $A \subset S$ of the embedded Markov chain, $\{X(\tau_n):\ n \in \mathbb{N}_0\}$.

$$\Theta_A = \tau_{\Delta_A} \tag{25}$$

denotes a first passage time to the subset A or the time of a first arrival at the set of states $A \subset S$ of the semi-Markov process $\{X(t): t \geq 0\}$. A function

$$\Phi_{iA}(t) = P(\Theta_A \leqslant t \,|\, X(0) = i),\ t \geqslant 0 \tag{26}$$

is the Cumulative Distribution Function (CDF) of a random variable Θ_{iA} denoting the first passage time from the state $i \in A'$ to a subset A or the exit time of $\{X(t): t \geq 0\}$ from the subset A' with the initial state i. We will present some theorems concerning distributions and parameters of the random variables Θ_{iA} which are conclusions from theorems presented by Koroluk & Turbin (1976), Silvestrov (1980), Grabski (2002).

THEOREM 1

For the regular semi-Markov processes such that,

$$f_{iA} = P(\Delta_A < \infty | X(0) = i) = 1, \quad i \in A', \tag{27}$$

distributions $\Phi_{iA}(t) \quad i \in A'$ *are proper and they are the unique solutions of the equations system*

$$\Phi_{iA}(t) = \sum_{j \in A} Q_{ij}(t) + \sum_{k \in S}\int_0^t \Phi_{kA}(t-x)dQ_{ik}(x), \quad i \in A'. \tag{28}$$

Applying a Laplace-Stieltjes (L-S) transformation for the system of integral equations we obtain the linear system of equations for (L-S) transforms

$$\tilde{\phi}_{iA}(s) = \sum_{j \in A} \tilde{q}_{ij}(s) + \sum_{k \in A'} \tilde{q}_{ik}(s)\tilde{\phi}_{kA}(s), \tag{29}$$

where

$$\tilde{\phi}_{iA}(s) = \int_0^\infty e^{-st} d\Phi_{iA}(t), \tag{30}$$

are L-S transforms of the unknown CDF of the random variables Θ_{iA}, $i \in A'$, and

$$\tilde{\phi}_{iA}(s) = \int_0^\infty e^{-st} d\Phi_{iA}(t), \tag{31}$$

are L-S transforms of the given functions $Q_{ij}(t),\ i,j \in S$. That linear system of equations is equivalent to the matrix equation

$$(\boldsymbol{I} - \tilde{\boldsymbol{q}}_{A'}(s))\,\tilde{\boldsymbol{\phi}}_{A'}(s) = \tilde{\boldsymbol{b}}(s), \tag{32}$$

where

$$\boldsymbol{I} = [\delta_{ij} :\ i,j \in A'] \tag{33}$$

is the unit matrix,

$$\tilde{\boldsymbol{q}}_{A'}(s) = [\tilde{q}_{ij}(s) :\ i,j \in A'] \tag{34}$$

is the square sub-matrix of the L-S transforms of the matrix $\tilde{\boldsymbol{q}}(s)$, while

$$\tilde{\boldsymbol{\phi}}_{A'}(s) = [\tilde{\phi}_{iA}(s) :\ i \in A']^T, \quad \tilde{\boldsymbol{b}}(s) = [\sum_{j\in A} \tilde{q}_{ij}(s) :\ i \in A']^T \tag{35}$$

are one column matrices of the corresponding L-S transforms.

The linear system of equations (29) for the L-S transforms allows us to obtain the linear system of equations for the moments of random variables $\Theta_{iA},\ i \in A'$.

THEOREM 2

If

- *assumptions of theorem 1 are satisfied,*
- $\bigvee\limits_{d>0}\ \bigwedge\limits_{i,j\in S}\ 0 < E(T_{ij}^2) \leqslant d,$
- $\bigwedge\limits_{i\in A}\ \mu_{iA}^2 = \sum\limits_{n=1}^{\infty} n^2 f_{iA}(n) < \infty,$

then there exist expectations $E(\Theta_{iA}),\ \ i \in A'$ *and second moments* $E(\Theta_{iA}^2), i \in A'$ *and they are unique solutions of the linear systems equations, which have following matrix forms*

$$(\boldsymbol{I} - \boldsymbol{P}_{A'})\overline{\boldsymbol{\Theta}}_{A'} = \overline{\boldsymbol{T}}_{A'}, \tag{36}$$

where

$$\boldsymbol{P}_{A'} = [p_{ij} :\ i,j \in A'],\ \overline{\boldsymbol{\Theta}}_{A'} = [E(\Theta_{iA}) :\ i \in A']^T,\ \overline{\boldsymbol{T}}_{A'} = [E(T_i) :\ i \in A']$$

$$(\boldsymbol{I} - \boldsymbol{P}_{A'})\overline{\boldsymbol{\Theta}^2}_{A'} = \boldsymbol{B}_A, \tag{37}$$

where

$$\boldsymbol{P}_{A'} = [p_{ij} : i,j \in A'], \quad \overline{\boldsymbol{\Theta}^2}_{A'} = [E(\Theta_{iA}^2) : i \in A']^T,$$

$$\boldsymbol{B}_A = [b_{iA} : i \in A']^T, \quad b_{iA} = E(T_i^2) + 2\sum_{k\in A'} p_{ik}E(T_{ik})E(\Theta_{kA}),$$

and $\boldsymbol{I}$ *is the unit matrix.*

In our case the random variable Θ_{iA}, that denotes the first passage time from the state $i = 2$ to the subset $A = \{0\}$ represents the time to failure of the system in our model. The function

$$R(t) = P(\Theta_{20} > t) = 1 - \Phi_{20}(t), \quad t \geqslant 0 \tag{38}$$

is the reliability function of the considered cold standby system with repair.

In this case the system of linear equations (29) for the Laplace-Stieltjes transforms with the unknown functions $\tilde{\phi}_{i0}(s), \quad t \geqslant 0, \quad i = 1,2$ is

$$\tilde{\phi}_{10}(s) = \tilde{q}_{10}(s) + \tilde{\phi}_{10}(s)\,\tilde{q}_{11}(s),\ \tilde{\phi}_{20}(s) = \tilde{q}_{20}(s) + \tilde{\phi}_{10}(s)\tilde{q}_{21}(s). \tag{39}$$

Hence

$$\tilde{\phi}_{10}(s) = \frac{\tilde{q}_{10}(s)}{1-\tilde{q}_{11}(s)}, \quad \tilde{\phi}_{20}(s) = \tilde{q}_{20}(s) + \frac{\tilde{q}_{21}(s)\tilde{q}_{10}(s)}{1-\tilde{q}_{11}(s)}. \tag{40}$$

Consequently, we obtain the Laplace transform of the reliability function

$$\tilde{R}(s) = \frac{1-\tilde{\phi}_{20}(s)}{s}. \tag{41}$$

The transition matrix of the embedded Markov chain of the semi-Markov process $\{X(t): t \geq 0\}$ is

$$\boldsymbol{P} = \begin{bmatrix} 0 & 0 & 1 \\ p_{10} & p_{11} & 0 \\ p_{20} & p_{21} & 0 \end{bmatrix}, \tag{42}$$

where

$$p_{10} = 1 - p_{11}, \quad p_{11} = P(U = 1, \gamma < \zeta) = a \int_0^\infty H(x) dF(x),$$

$$p_{20} = 1 - a, \quad p_{21} = P(U = 1) = a.$$

The CDF of the waiting times $T_i, \quad i = 0, 1, 2$ are

$$G_0(t) = K(t), \quad G_1(t) = F(t), \quad G_2(t) = F(t).$$

Hence

$$E(T_0) = E(\eta), \quad E(T_1) = E(\zeta), \quad E(T_3) = E(\zeta). \tag{43}$$

In this case equation (37) takes the form of

$$\begin{bmatrix} 1-p_{11} & 0 \\ -a & 1 \end{bmatrix} \begin{bmatrix} E(\Theta_{10}) \\ E(\Theta_{20}) \end{bmatrix} = \begin{bmatrix} E(\zeta) \\ E(\zeta) \end{bmatrix}. \tag{44}$$

The solution of it is:

$$E(\Theta_{10}) = \frac{E(\zeta)}{1-p_{11}}, \quad E(\Theta_{20}) = E(\zeta) + \frac{a\,E(\zeta)}{1-p_{11}}. \tag{45}$$

6. An Approximate Reliability Function

In this case calculating an exact reliability function of the system by means of Laplace transform is a complicated matter. Finding an approximate reliability function of that system is possible by using results from the theory of semi-Markov processes perturbations. The perturbed semi-Markov processes are defined in different ways by different authors. We introduce Pavlov and Ushakov (1978) concept of the perturbed semi-Markov process presented by I.B. Gertsbakh (1984).

Let $A' = S - A$ be a finite subset of states and A be at least countable subset of S. Suppose $\{X(t): t \geq 0\}$ is SM process with the state space $S = A \cup A'$ and the kernel $\boldsymbol{Q}(t) = [Q_{ij}(t) : i, j \in S]$, the elements of which have the form $Q_{ij}(t) = p_{ij}F_{ij}(t)$.

Assume that

$$\varepsilon_i = \sum_{j \in A} p_{ij} \tag{46}$$

and

$$p_{ij}^0 = \frac{p_{ij}}{1-\varepsilon_i}, \quad i, j \in A'. \tag{47}$$

Let us notice that $\sum_{j\in A'} p_{ij}^0 = 1$.

A semi-Markov process $\{X(t): t \geq 0\}$ with the discrete state space S defined by the renewal kernel $\boldsymbol{Q}(t) = [p_{ij}F_{ij}(t) : i,j \in S]$, is called the perturbed process with respect to SM process $\{X^0(t) : t \geqslant 0\}$ with the state space A' defined by the kernel $\boldsymbol{Q}^0(t) = [p_{ij}^0 F_{ij}(t) : i,j \in A']$.

We are going to present our version of theorem proved by I.B. Gertsbakh (1984).

The number

$$m_i^0 = \int_0^\infty [1 - G_i^0(t)]dt, \quad i \in A', \tag{48}$$

where

$$m_i^0 = \int_0^\infty [1 - G_i^0(t)]dt, \quad i \in A', \tag{49}$$

is the expected value of the waiting time in state i for the process $\{X^0(t) : t \geqslant 0\}$. Denote the stationary distribution of the embedded Markov chain in SM process $\{X^0(t) : \ t \geqslant 0\}$ by $\pi^0 = [\pi_i^0 : \ i \in A']$. Let

$$\varepsilon = \sum_{i\in A'} \pi_i^0 \varepsilon_i, \quad m^0 = \sum_{i\in A} \pi_i^0 m_i^0. \tag{50}$$

We are interested in the limiting distribution of the random variable $\Theta_{iA} = \inf\{t : X(t) \in A \,|\, X(0) = i\}, \quad i \in A'$, that denotes the first passage time from the state $i \in A'$ to the subset A.

THEOREM 3

If the embedded Markov chain defined by the matrix of transition probabilities $\boldsymbol{P} = [\, p_{ij} : i,j \in S]$ *satisfies the following conditions*

$$f_{iA} = P(\Delta_A < \infty | X(0) = i) = 1, \quad i \in A', \tag{51}$$

$$\bigvee_{c>0} \bigwedge_{i,j\in S} \ 0 < E(T_{ij}) \leqslant c, \tag{52}$$

$$\bigwedge_{i\in A} \mu_{iA} = \sum_{n=1}^{\infty} n f_{iA}(n) < \infty, \tag{53}$$

then

$$\lim_{\varepsilon\to 0} P(\varepsilon\Theta_{iA} \ > \ x) = e^{-\frac{x}{m^0}}, \tag{54}$$

where $\pi^0 = [\pi_i : i \in A']$ *is the unique solution of the linear system of equations*

$$\pi^0 = \pi^0 \boldsymbol{P}^0, \quad \pi^0 \mathbf{1} = 1. \tag{55}$$

From that theorem it follows that for small ε we get the following approximating formula

$$P(\Theta_{iA} > t) \ \approx \ \exp\left[- \frac{\sum_{i\in A'} \pi_i^0 \varepsilon_i}{\sum_{i\in A'} \pi_i^0 m_i^0} \, t \right], \quad t \geqslant 0. \tag{56}$$

The considered SM process $\{X(t):\ t \geqslant 0\}$ with the state space $S = \{0, 1, 2\}$ we can assume to be the perturbed process with respect to the SM process $\{X^0(t):\ t \geqslant 0\}$ with the state space $A' = \{1, 2\}$ and the kernel

$$\boldsymbol{Q}^0(t) = \begin{bmatrix} Q_{11}^0(t) & 0 \\ Q_{21}^0(t) & 0 \end{bmatrix}, \tag{57}$$

where

$$Q_{11}^0(t) = p_{11}^0 F_{11}(t), \quad Q_{21}^0(t) = p_{21}^0 F_{21}(t).$$

Because $A = \{0\}$ and

$$\varepsilon_1 = p_{10} = Q_{10}(\infty) = 1 - a\int_0^\infty G(x)f(x)dx,$$

then

$$p_{11}^0 = \frac{p_{11}}{1-\varepsilon_1} = 1.$$

From

$$\frac{Q_{11}(t)}{p_{11}} = F_{11}(t),$$

we get

$$Q_{11}^0(t) = F_{11}(t) = \frac{\int_0^t G(x)f(x)dx}{\int_0^\infty G(x)f(x)dx}.$$

Notice, that $\varepsilon_2 = p_{20} = 1 - a$. Hence $p_{21}^0 = \frac{p_{21}}{1-\varepsilon_2} = 1$. Finally we obtain

$$Q_{21}^0(t) = F_{21}(t) = F(t).$$

The transition matrix of the embedded Markov chain of SM process $\{X^0(t):\ t \geqslant 0\}$ is

$$\boldsymbol{P}^0 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}. \tag{58}$$

From the system of equations

$$[\pi_1^0, \pi_2^0]\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} = [\pi_1^0, \pi_2^0], \tag{59}$$

$$\pi_1^0 + \pi_2^0 = 1.$$

we get $\boldsymbol{\pi}^0 = [1,\ 0]$. It follows from the theorem 3 that for a small ε

$$P(\Theta_{iA} > t) \approx \exp\left[-\frac{\sum_{i\in A'}\pi_i^0\varepsilon_i}{\sum_{i\in A'}\pi_i^0 m_i^0}\,t\right], \quad t \geqslant 0, \tag{60}$$

where

$$m_i^0 = \int_0^\infty [1 - G_i^0(t)]dt,\ i \in A',\ G_i^0(t) = \sum_{j\in A'} Q_{ij}^0(t),\ \varepsilon = \sum_{i\in A'}\pi_i^0\varepsilon_i, \tag{61}$$

and

$$m^0 = \sum_{i \in A} \pi_i^0 m_i^0. \tag{62}$$

Therefore we have

$$\varepsilon = \varepsilon_1 = 1 - a \int_0^\infty H(x) f(x) dx,$$

$$m^0 = m_1^0 = \frac{\int_0^\infty x H(x) f(x) dx}{\int_0^\infty H(x) f(x) dx}.$$

For ε close to 0 we obtain the approximate reliability function of the system

$$R(t) = P(\Theta_{iA} > t) = P(\varepsilon \Theta_{iA} > \varepsilon t) \approx \exp\left[-\frac{\varepsilon}{m^0} t\right], \quad t \geqslant 0.$$

From a shape of the parameter ε it follows that we can apply this formula only if the number $P(\gamma \geqslant \zeta)$, denoting probability of a component failure during a period of an earlier failed component, is small.

Finally we obtain an approximate relation

$$R(t) = P\{\Theta_{20} > t\} \approx \exp\left[-\frac{c(1-ac)}{m_*} t\right], \tag{63}$$

where

$$c = \int_0^\infty H(x) f(x) dx = P(\gamma < \zeta),$$

$$m_* = \int_0^\infty x H(x) f(x) dx.$$

7. Conclusions

- The expectation $E(\Theta_{20})$ denoting the mean time to failure of the considered cold standby system is

  $$E(\Theta_{20}) = E(\varsigma) + \frac{a\, E(\varsigma)}{1 - p_{11}},$$

  where

  $p_{11} = a \int_0^\infty H(x) dF(x)$.

- The cold standby determines the increase of the mean time to failure

  $$1 + \frac{a}{1 - p_{11}}$$

  times.

- The approximate reliability function of the system is exponential with a parameter

  $$\Lambda = \frac{c(1-ac)}{m_*},$$

  where

$$c = P(\gamma < \zeta) = \int_0^\infty H(x) f(x) dx,$$

$$m_* = \int_0^\infty x H(x) f(x) dx.$$

References

1. Barlow, R.E. and Proshan, F. **Mathematical Theory of Reliability**, Wiley: New-York, London, Sydney, 1965
2. Brodi, S.M. and Pogosian, I.A. **Embedded Stochastic Processes in the Queue Theory**, Naukova Dumka, Kiev, 1978 (in Russian)
3. Gertsbakh, I.B. **Asymptotic methods in reliability theory: A review**, Adv. Appl. Prob., 16, 1984, pp. 147-175
4. Grabski, F.G. **Semi-Markov model of reliability and operation**, PAN IBS, Operation Research, 30, Warszawa, 2002, pp. 161 (in Polish)
5. Grabski, F.G. **Applications of Semi-Markov processes in safety and reliability analysis**, SSARS 2009, Gdańsk, 2009, pp. 94
6. Koroluk, W.S. and Turbin, A.F. **Semi-Markov Processes and Their Applications**, Naukova Dumka, Kiev, 1976 (in Russian)
7. Pavlov, I.V. and Ushakov, I.A. **The asymptotic distribution of the time until a semi-Markov process gets out of the kernel**, Engineering Cybernetics, 20 (3), 1978
8. Silvestrov, D.S. **Semi-Markov Processes with Discrete State Space**, Sovietskoe Radio, Moscow, 1980 (in Russian)

JOURNAL OF APPLIED QUANTITATIVE METHODS

International Symposium on Stochastic Models in Reliability Engineering, Life Sciences and Operations Management (SMRLO'10)

ANALYTICAL AND NUMERICAL STUDIES OF PERTURBED RENEWAL EQUATIONS WITH MULTIVARIATE NON-POLYNOMIAL PERTURBATIONS

Ying NI
PhD Candidate, Division of Applied Mathematics,
School of Education, Culture and Communication,
Mälardalen University,
Västerås, Sweden

E-mail: ying.ni@mdh.se

Abstract: *The object of study is a model of nonlinearly perturbed continuous-time renewal equation with multivariate non-polynomial perturbations. The characteristics of the distribution generating the renewal equation are assumed to have expansions in a perturbation parameter with respect to a non-polynomial asymptotic. Exponential asymptotics for such a model as well as their applications are given. Numerical studies are performed to gain insights into the asymptotical results.*

Key words: *perturbed renewal equation; nonlinear perturbation; non-polynomial perturbation; perturbed risk process; ruin probability*

1. Introduction

This paper deals with nonlinearly perturbed renewal equations with a new type of non-polynomial perturbations. That is, some characteristics of the distribution generating the perturbed renewal equation, namely the defect and moments, can be expanded in the perturbation parameter ε up to some order α with respect to the following non-standard non-polynomial asymptotic scale,

$$\{\varphi_{\vec{n}}(\varepsilon) = \varepsilon^{\vec{n}\cdot\vec{\omega}}, \vec{n} \in \mathbf{N}_0^k\}, \text{as } \varepsilon \to 0, \tag{1}$$

where $\mathbf{N}_0$ is the set of non-negative integers, $\mathbf{N}_0^k \equiv \mathbf{N}_0 \times \cdots \times \mathbf{N}_0, 1 \le k < \infty$ with the product being taken k times, and $\vec{\omega}$ is a parameter vector of dimension k. In (1), $\vec{n} \cdot \vec{\omega}$ denotes the dot product of vector $\vec{n}$ and $\vec{\omega}$, and by the definition of asymptotic scale, the gauge functions $\varphi_{\vec{n}}(\varepsilon)$ are ordered by index $\vec{n}$ in such way that the later function in

the sequence is always o-function of the previous one. Further, we assume that the parameter vector $\vec{\omega} = (\omega_1, \omega_2, \ldots, \omega_k)$ has the following properties: **(i)** $1 = \omega_1 < \omega_2 < \ldots < \omega_k$; **(ii)** the components are linearly independent over the field Q of rational numbers, i.e., ω_i / ω_j is an irrational number for any $i \neq j, i, j = 1, \ldots, k$. Note that it follows from **(i)** and **(ii)** that $\omega_2, \ldots, \omega_k$ are irrational numbers. Throughout the paper, the symbol $\vec{\omega}$ refers to some parameter vector satisfying these two properties.

The aim of this paper is to present the asymptotic behavior of such perturbed renewal equations, illustrate the result by applications and carry out numerical studies of the applications. The case for $k = 2$ in (1) has been studied in the previous research (Ni, Silvestrov and Malyarenko 2008). Setting $k = 1$, the asymptotic scale (1) reduces to the standard polynomial asymptotic scale, and this case was first investigated in Silvestrov (1995). The present paper covers the general case where k can be any finite positive integer, that is, the case with "multivariate" non-polynomial perturbations. Other works on nonlinearly perturbed renewal equation with non-polynomial perturbations have been done by Englund and Silvestrov (1997) and Englund (2001), where the expansions of defect and moments have polynomial and mixed polynomial-exponential forms.

For a general theory of nonlinearly perturbed renewal equations with applications to non-linearly perturbed stochastic systems, we refer to the book by Gyllenberg and Silvestrov (2008) and references therein. Note that all expansions in this book are based on the standard polynomial asymptotical scale.

2. The Model

Let us consider the following perturbed renewal equation which holds for every $\varepsilon \geq 0$:

$$x_\varepsilon(t) = q_\varepsilon(t) + \int_0^t x_\varepsilon(t-s) F_\varepsilon(ds), \ t \geq 0, \tag{2}$$

where the force function $q_\varepsilon(t)$ refers to a measurable real-valued function on $[0, \infty)$ being bounded on every finite interval. The distribution function $F_\varepsilon(\cdot)$ generating this renewal equation has its support on $[0, \infty)$, is not concentrated at 0 and can be improper. It is known that there exists a unique solution which is both measurable and bounded on every finite interval solution, $x_\varepsilon(t)$, for equation (2).

The defect and moments for F_ε are defined as

$$f_\varepsilon = 1 - F_\varepsilon(\infty), \qquad m_{\varepsilon r} = \int_0^\infty s^r F_\varepsilon(ds), \qquad r \geq 1.$$

Assume that the following perturbation conditions hold for $F_\varepsilon(\cdot)$ and $q_\varepsilon(\cdot)$.

A. $F_\varepsilon(t) \Rightarrow F_0(t)$ as $\varepsilon \to 0$, where $F_0(t)$ is a proper and non-arithmetic distribution function.

B. (*Cramér type condition*) There exists $\delta > 0$ such that

$\overline{\lim}_{0\le\varepsilon\to 0}\int_0^\infty e^{\delta s}F_\varepsilon(ds)<\infty$.

C. (a) $\lim_{u\to 0}\overline{\lim}_{0\le\varepsilon\to 0}\sup_{|v|\le u}|q_\varepsilon(t+v)-q_0(t)|=0$ a.e. with respect to Lebesgue measure on $[0,\infty)$;

(b) $\overline{\lim}_{0\le\varepsilon\to 0}\sup_{0\le t\le T}|q_\varepsilon(t)|<\infty$ for every $T\ge 0$;

(c) $\lim_{T\to\infty}\overline{\lim}_{0\le\varepsilon\to 0}h\sum_{r\ge T/h}\sup_{rh\le t\le(r+1)h}e^{\gamma t}|q_\varepsilon(t)|=0$ for some $h>0$, $\gamma>0$.

Note that symbol $F_\varepsilon(\cdot)\Rightarrow F_0(\cdot)$ as $\varepsilon\to 0$ denotes weak convergence of the distribution functions. Notations $\overline{\lim}$ and $\underline{\lim}$ are equivalent to $\limsup$ and $\liminf$, respectively.

It follows from condition B that the exponential moment of F_ε, defined as

$$\phi_\varepsilon(\rho)=\int_0^\infty e^{\rho s}F_\varepsilon(ds),\ \rho\ge 0,$$

is finite for $\rho<\delta$ and ε small enough.

It is known that, under condition A and B there is a unique nonnegative root, ρ_ε, of the following characteristic equation:

$$\phi_\varepsilon(\rho)=\int_0^\infty e^{\rho s}F_\varepsilon(ds)=1, \tag{3}$$

for ε small enough and $\rho_\varepsilon\to 0$ as $\varepsilon\to 0$.

The following theorem (Silvestrov, 1976, 1978, 1979) serves as the starting point for the present study.

THEOREM 1

Let conditions A, B and C be satisfied. Then for any $0\le t_\varepsilon\to\infty$ as $\varepsilon\to 0$, the following asymptotical relation holds

$$\frac{x_\varepsilon(t_\varepsilon)}{\exp\{-\rho_\varepsilon t_\varepsilon\}}\to x_0(\infty)=\frac{\int_0^\infty q_0(s)ds}{m_{01}}\quad\text{as}\quad\varepsilon\to 0. \tag{4}$$

By condition A we have $f_\varepsilon\to f_0=0$ as $\varepsilon\to 0$ and by Condition B all moments of F_ε are finite, i.e. $m_{\varepsilon r}<\infty, r\ge 1$. Condition A and B also imply that for ε small enough, $m_{\varepsilon r}\to m_{0r}\in(0,\infty)$ as $\varepsilon\to 0, r\ge 1$. The basic idea of the present research is: by assuming some appropriate form of asymptotic expansions for f_ε and $m_{\varepsilon r}$, the corresponding asymptotic expansion of ρ_ε may be obtained, which can be used to improve the asymptotic relation (4) to a more explicit form.

For some real number $\alpha \geq 1$, notation $[\alpha]_{\vec{\omega}}$ is defined as: $[\alpha]_{\vec{\omega}} = \max(\vec{n} \cdot \vec{\omega} : \vec{n} \cdot \vec{\omega} \leq \alpha, \vec{n} \in \mathbf{N}_0^k)$, i.e. the last gauge function in (1) that has the order less than or equal to α is $\varepsilon^{[\alpha]_{\vec{\omega}}}$.

Given α and a specific parameter vector $\vec{\omega}$, by property (ii) of $\vec{\omega}$ we know that there exists a unique vector $\vec{n}$ such that $[\alpha]_{\vec{\omega}} = \vec{n} \cdot \vec{\omega}$, we denote this $\vec{n}$ by $\vec{f}(\alpha, \vec{\omega})$.

Notation $[\alpha]$ is used to denote the integer part of number α. Let us also define the following two sets:

$$\mathbf{R}_i(\vec{n}) = \{\vec{p} : p_1 \leq n_1, \ldots, p_k \leq n_k, \sum_{j=1}^{k} p_j \geq i\}, \quad \mathbf{R}_i'(\vec{n}) = \mathbf{R}_i(\vec{n}) \setminus \{\vec{n}\},$$

where $\vec{n}, \vec{p} \in \mathbf{N}_0^k$. For example, if $\vec{n} = (2,1), i = 2$ then $\mathbf{R}_i(\vec{n})$ refers to the set $\{(1,1),(2,0),(2,1)\}$ while $\mathbf{R}_i'(\vec{n})$ represents the set $\{(1,1),(2,0)\}$.

All vectors in this paper are k-dimensional (as for $\vec{\omega}$) row vectors unless stated otherwise, and they are represented with lowercase Roman/Greek letters with right-pointing arrows above. Symbol $\vec{0}$ is a vector with all components equal to zeros, and $\vec{e}_i$ refers to i-th unit vector, i.e. all components are zero except that the i-th component is equal to one.

We are now in a position to impose the following additional perturbation conditions which hold for a given real number $\alpha \geq 1$ and for some given parameter vector $\vec{\omega}$.

$\mathbf{P}_{\vec{\omega}}^{(\alpha)}$: (a) $1 - f_\varepsilon = 1 + \sum_{1 \leq \vec{n} \cdot \vec{\omega} \leq \alpha} b_{\vec{n},0} \varepsilon^{\vec{n} \cdot \vec{\omega}} + o(\varepsilon^{[\alpha]_{\vec{\omega}}})$, where all coefficients are finite.

(b) $m_{\varepsilon r} = m_{0r} + \sum_{1 \leq \vec{n} \cdot \vec{\omega} \leq \alpha - r} b_{\vec{n},r} \varepsilon^{\vec{n} \cdot \vec{\omega}} + o(\varepsilon^{[\alpha - r]_{\vec{\omega}}})$, for $r = 1, \ldots, [\alpha]$, where all coefficients are finite.

Remark 1. In condition $\mathbf{P}_{\vec{\omega}}^{(\alpha)}$, the defect and moments are expanded, up to order α, with respect to asymptotic scale (1). For convenience, notation: $b_{\vec{0},0} = 1, b_{\vec{0},r} = m_{0r}$ is also used.

3. The Main Result

The following theorem presents the exponential asymptotics for the solution to the perturbed renewal equation described in the previous section.

THEOREM 2

Let conditions A, B and $\mathbf{P}_{\vec{\omega}}^{(\alpha)}$ be satisfied. Then:

***(i)** There exists a unique non-negative solution to characteristic equation (3) for all ε that are small enough. Further, the following expansion for ρ_ε holds,*

$$\rho_\varepsilon = \sum_{1\le \vec{n}\cdot\vec{\omega}\le\alpha} a_{\vec{n}}\varepsilon^{\vec{n}\cdot\vec{\omega}} + o(\varepsilon^{[\alpha]_{\vec{\omega}}}), \tag{5}$$

where the coefficients can be calculated by the following recurrent formula: $a_{\vec{e}_1} = -b_{\vec{e}_1,0}/b_{\vec{0},1}$, *and in general for* $\vec{n}: 1 < \vec{n}\cdot\vec{\omega} \le \alpha$,

$$a_{\vec{n}} = -\frac{1}{b_{\vec{0},1}}(b_{\vec{n},0} + \sum_{\vec{p}\in\mathbf{R}_1'(\vec{n})} b_{\vec{n}-\vec{p},1}a_{\vec{p}} + \sum_{i=2}^{n_1+n_2+\ldots+n_k} \sum_{\vec{p}\in\mathbf{R}_i(\vec{n})} b_{\vec{n}-\vec{p},i}\Big(\sum_{\vec{j}(\vec{p})\in\mathbf{D}_i(\vec{p})} \prod_{\vec{r}\in\mathbf{R}_1'(\vec{p})} \frac{(a_{\vec{r}})^{j_{\vec{r}}}}{(j_{\vec{r}})!}\Big)), \tag{6}$$

where $\mathbf{D}_i(\vec{p})$ *is the set of all nonnegative and integer solutions,* $\vec{j}(\vec{p}) \equiv (j_{\vec{r}}, \vec{r}\in\mathbf{R}_1'(\vec{p}))$, *for the Diophantine system*

$$\begin{cases} \sum\limits_{\vec{r}\in\mathbf{R}_1'(\vec{p})} j_{\vec{r}} = i, \\ \sum\limits_{\vec{r}\in\mathbf{R}_1'(\vec{p})} \vec{r}\times j_{\vec{r}} = \vec{p} \end{cases} \tag{7}$$

(ii) *If the coefficients for the defect satisfy* $b_{\vec{n},0} = 0$ *for* $\vec{n}$ *such that* $\vec{n}\cdot\vec{\omega} \le \beta$ *for some* $1 \le \beta \le \alpha$, *then* $a_{\vec{n}} = 0$ *for* $\vec{n}$ *such that* $\vec{n}\cdot\vec{\omega} \le \beta$.

(iii) *If in addition condition C holds, then for any* $0 \le t_\varepsilon \to \infty$ *balanced with* $\varepsilon \to 0$ *in such a way that* $\varepsilon^{[\beta]_{\vec{\omega}}} t_\varepsilon \to \lambda_\beta \in [0,\infty)$ *where* $\beta \in [1,\alpha]$ *is a given real number, we have the following asymptotical relation:*

$$\exp\{(\sum_{1\le\vec{n}\cdot\vec{\omega}<[\beta]_{\vec{\omega}}} a_{\vec{n}}\,\varepsilon^{\vec{n}\cdot\vec{\omega}})t_\varepsilon\}\, x_\varepsilon(t_\varepsilon) \to e^{-\lambda_\beta a^{(1)}} x_0(\infty) \quad \text{as} \quad \varepsilon\to 0,$$

where $a^{(1)} = a_{\vec{p}}$ *with* $\vec{p} = \vec{f}(\beta,\vec{\omega})$.

Remark 2. The coefficient $a_{\vec{n}}$ can be calculated from the recurrent formula (6) if $\vec{n}$ satisfies $1 \le \vec{n}\cdot\vec{\omega} \le \alpha$. It can be directly seen from formula (6) that $a_{\vec{n}}$ depends on the set of coefficients $\{a_{\vec{p}} : \vec{p}\in\mathbf{R}_1'(\vec{n})\}$ which is obviously a subset to $\{a_{\vec{p}} : 1 \le \vec{p}\cdot\vec{\omega} < \vec{n}\}$. Also one can observe from (5) and (6) that the value of coefficient $a_{\vec{n}}$ does not depend on parameter vector $\vec{\omega}$ and parameter α.

Remark 3. For a given $\vec{\omega}$ and a given α, the expansion of ρ_ε (5) takes a unique form, and so is the sequence of coefficients $a_{\vec{n}}$, $1 \le \vec{n}\cdot\vec{\omega} < \alpha$. Let the terms in expansion (5) be ordered in terms of the powers of ε, a natural choice of recursive algorithm would be to first calculate the first-by-order coefficient in the expansion then the second-by-order coefficient and so on.

Remark 4. If the dimension of $\vec{\omega}$ is one, so that $\vec{\omega} = 1$, and let also α be some positive integer greater than one, we have the particular case where the defect and moments are expanded with respect to the standard polynomial asymptotic scale, up to and including the order α. This case has been studied in Silvestrov (1995) and Gyllenberg and Silvestrov (2008). Theorem 2 reduces, in this case, to the corresponding result obtained

there. Similarly when $\vec{\omega}$ has dimension 2, so that $\vec{\omega}=(1,\omega)$ for some irrational $\omega>1$, and let α be some real number, Theorem 2 reduces, in this case, to the corresponding result in Ni, Silvestrov and Malyarenko (2008). For convenience, we shall call the latter case, i.e. the case when $\vec{\omega}$ has dimension 2, as the "bivariate" case.

One can use recurrent formula (6) to calculate manually the coefficients $a_{\vec{n}}$ when parameter k and α are relatively small. For larger values of k and α, it is better to program formula (6). For instance, the corresponding MATLAB routine has been developed by the author. The algorithm takes inputs α and $\vec{\omega}$ (hence the dimension of $\vec{\omega}$, k), then determine the sequence of coefficients included in (5) by solving $\vec{n}\cdot\vec{\omega}\le\alpha$ for integer values of $\vec{n}$ and sorting the solutions, $\vec{n}$, in ascending order with respect to the value of $\vec{n}\cdot\vec{\omega}$. The next step is to determine recursively the coefficients using formula (6). Although tedious, most part of this formula are relatively easy to program. Let us only describe briefly the algorithm for solving the Diophantine system (7).

The second equation in system (7) leads to k equations since the dimension of vector $\vec{r}$ and $\vec{p}$ is k. The unknowns are $j_{\vec{r}},\vec{r}\in\mathbf{R}_1'(\vec{p})$. Denote $q=|\mathbf{R}_1'(\vec{p})|$, i.e. q is the number of vectors in the set $\mathbf{R}_1'(\vec{p})$. System (7) is indeed a Diophantine system of q unknowns in $k+1$ equations. Let us express this system in the matrix equation $\mathbf{A}\vec{x}=\vec{b}$, where $\mathbf{A}$ is $(k+1)\times q$ matrix, i.e. the matrix of coefficients for the system, $\vec{x}$ is the unknown column vector with q entries, and b is a column vector with $k+1$ entries. The problem is therefore: determine the set of non-negative integer solutions to $\mathbf{A}\vec{x}=\vec{b}$, and this can be efficiently solved by using a recursive algorithm.

4. Applications

The results may have many potential applications, for instance, to the analysis of nonlinearly perturbed risk processes and processes which are used to describe functioning of queueing systems. We present in this section two examples of perturbed classical risk processes, with bivariate and multivariate non-polynomial perturbations respectively. Theorem 2 is applied to obtain asymptotic behaviour for ruin probabilities and experimental numerical studies are carried out to gain insights into the asymptotical results. Since there's a duality of classical risk processes with the workload process of a $M/G/1$ queue, and with the dam/storage process, the results also have interpretation in these areas.

Let us consider the perturbed classical risk-process which describes the time evolution of the reserves of an insurance company

$$X_\varepsilon(t)=ct-\sum_{j=1}^{N(t)}Z_{j\varepsilon},\qquad t\ge 0, \tag{8}$$

where $c>0$ is the gross risk premium rate; $N(t),t\ge 0$ is the Poisson claim arrival process with rate λ; the claim sizes $Z_{j\varepsilon}$, $j=1,\ldots,N(t)$ are i.i.d. nonnegative random variables, independent of process $N(t)$, that follow a common distribution $G_\varepsilon(z)$ with a finite mean $\mu_\varepsilon=\int_0^\infty z\,G_\varepsilon(dz)<\infty$;

It is usually assumed that the moment characteristics of $G_\varepsilon(z)$ depend on the perturbation parameter ε but converge in some sense to the corresponding characteristics of limiting distribution $G_0(z)$ as $\varepsilon \to 0$. These continuity conditions allow us to consider the risk process $X_\varepsilon(t)$ for $\varepsilon > 0$ as a perturbed version of $X_0(t)$ for $\varepsilon = 0$.

The loading rate of claims are characterized by a constant α_ε or equivalently by the safety loading coefficient η_ε, defined respectively as

$$\alpha_\varepsilon = \frac{\lambda \mu_\varepsilon}{c}; \qquad \eta_\varepsilon = \frac{1-\alpha_\varepsilon}{\alpha_\varepsilon}. \tag{9}$$

Let $u \geq 0$ be the initial reserve of the insurance company, the object of our study is the ruin probability,

$$\Psi_\varepsilon(u) = \mathrm{P}\{u + \inf_{t \geq 0} X_\varepsilon(t) < 0\}.$$

The ruin probability is known to be equal to one if $\alpha_\varepsilon \geq 1$ or equivalently if the safety loading $\eta_\varepsilon \leq 0$. For $\alpha_\varepsilon \leq 1$, $\Psi_\varepsilon(u)$ as a function of initial reserve u, satisfies the following perturbed renewal equation (Feller, 1966),

$$\Psi_\varepsilon(u) = \alpha_\varepsilon(1-\widetilde{G}_\varepsilon(u)) + \alpha_\varepsilon \int_0^u \Psi_\varepsilon(u-s)\widetilde{G}_\varepsilon(ds), \qquad u \geq 0, \tag{10}$$

where $\widetilde{G}_\varepsilon(u)$ is the integrated tail distribution, i.e.

$$\widetilde{G}_\varepsilon(u) = \frac{1}{\mu_\varepsilon}\int_0^u (1-G_\varepsilon(s))ds. \tag{11}$$

Note that $\alpha_\varepsilon = 1$ is the trivial case since $\Psi_\varepsilon(u) \equiv 1$ is a solution to equation (10) if $\alpha_\varepsilon = 1$.

The distribution function that generates the perturbed renewal equation (10) is $F_\varepsilon(u) = \alpha_\varepsilon \widetilde{G}_\varepsilon(u)$.

Denote α_0 as α_ε for $\varepsilon = 0$. Let us assume the following condition holds.

D. $\alpha_0 = 1$.

Note that condition **D** implies that $\Psi_0(u) = 1$ for all $u \geq 0$.

Our aim is to obtain the asymptotic behavior of $\Psi_\varepsilon(u)$ as the perturbation parameter $\varepsilon \to 0$ simultaneously as the initial reserve $u \to \infty$ under some balancing condition. Let us use notation u_ε to emphasize that u is changing together with ε.

4.1. Perturbed Risk Process with Bivariate Non-polynomial Perturbations

We consider the perturbed risk process (8) and assume the following form for the limiting claim size distribution $G_0(z)$,

$$G_0(z) = \begin{cases} 1 - \dfrac{(T_0 - z)^\omega}{T_0^\omega}, & 0 \leq z < T_0, \\ 1, & z \geq T_0, \end{cases} \tag{12}$$

where T_0 is a constant parameter and parameter $\omega > 1$ is some irrational number.

The first moment μ_0 for $G_0(z)$ is,

$$\mu_0 = \int_0^\infty s\, G_0(ds) = \frac{T_0}{\omega+1}. \tag{13}$$

Let the perturbed claim size distribution $G_\varepsilon(z)$ for $\varepsilon \geq 0$ be given by

$$G_\varepsilon(z) = \begin{cases} 1 - \dfrac{(T-z)^\omega}{T^\omega}, & 0 \leq z < T, \\ 1, & z \geq T, \end{cases} \tag{14}$$

where T is a constant parameter and $T \leq T_0$. Let us use $\varepsilon \equiv T_0 - T \geq 0$ as the perturbation parameter.

In other words, $G_\varepsilon(z) = \mathrm{P}\{Z_{j0} \wedge T \leq z\}$ where Z_{j0} follows distribution $G_0(z)$, which can be caused for example by a excess-of-loss reinsurance with retention level T.

Taking account of (13), the first moment of $G_\varepsilon(z)$, μ_ε can be calculated as

$$\mu_\varepsilon = \int_0^\infty s G_\varepsilon(ds) = \mu_0 - \frac{\mu_0}{T_0^{\omega+1}} \varepsilon^{\omega+1}. \tag{15}$$

Note that it follows from (12) and (14) that $G_\varepsilon(z) \to G_0(z)$ as $\varepsilon \to 0$ for every $z \geq 0$. Also $\mu_\varepsilon \leq \mu_0$ and $\mu_\varepsilon \to \mu_0$ as $\varepsilon \to 0$ due to (15).

It follows from condition **D**, (9), (13) and (15) that $\alpha_\varepsilon \leq \alpha_0 = 1$ and $\alpha_\varepsilon \to \alpha_0 = 1$ as $\varepsilon \to 0$, which is the situation considered in a diffusion approximation for ruin probabilities.

Under condition **D** we have $\Psi_0(u) = 1$. Also we have $\alpha_\varepsilon \leq 1$, hence the ruin probability $\Psi_\varepsilon(u)$ satisfies the perturbed renewal equation (10).

Since for $\varepsilon > 0$ we have $\alpha_\varepsilon < 1$, the distribution function $F_\varepsilon(u) = \alpha_\varepsilon \widetilde{G}_\varepsilon(u)$ in this case is improper, i.e. defect $f_\varepsilon = 1 - F_\varepsilon(\infty) = 1 - \alpha_\varepsilon > 0$ for $\varepsilon > 0$. Obviously $f_\varepsilon \to f_0 = 0$ as $\varepsilon \to 0$.

It can be shown that the defect f_ε takes the following form,

$$f_\varepsilon = 1 - \alpha_\varepsilon = 1 - \frac{\lambda \mu_\varepsilon}{c} = \frac{1}{T_0^{1+\omega}} \varepsilon^{1+\omega}. \tag{16}$$

By repeatedly applying integration by parts, the r-th moment, $r \geq 1$, of $F_\varepsilon(u)$ can be calculated as:

$$m_{\varepsilon r} = \int_0^\infty s^r F_\varepsilon(ds) = \frac{(r!) \cdot T_0^r}{\prod_{i=2}^{r+1}(\omega+i)} + \sum_{k=0}^{r} (-1)^{k+1} \binom{r}{k} \frac{\omega+1}{\omega+k+1} T_0^{r-k-\omega-1} \varepsilon^{k+1+\omega}. \tag{17}$$

Let us define $[\xi]_\omega \equiv \max(n+m\omega : n+m\omega \leq \xi, n, m \in \mathbf{N}_0)$, where $\mathbf{N}_0$ is the set of non-negative integers. We now set $\xi = 4+3\omega$ so that $[\xi]_\omega = 4+3\omega$. Using (16), (17), the characteristics of $F_\varepsilon(u)$, namely the defect and moments, can be written down as the following perturbation condition.

$\mathbf{P}_{\omega}^{(\xi)}$: **(a)** $1-f_{\varepsilon}=1+\sum_{1\le n+m\omega\le 4+3\omega} b_{n,m,0}\varepsilon^{n+m\omega}+o(\varepsilon^{4+3\omega})$, where coefficients are given in (16);

(b) $m_{\varepsilon r}=m_{0r}+\sum_{1\le n+m\omega\le 4+3\omega-r} b_{n,m,r}\varepsilon^{n+m\omega}+o(\varepsilon^{4+3\omega-r})$, for $r=1,\ldots,[4+3\omega]$, where coefficients are given in (17).

We would like to note that (16), (17) can be rewritten in the form of $\mathbf{P}_{\omega}^{(\xi)}$ for any $\xi<\infty$. Note also that the perturbation condition $\mathbf{P}_{\omega}^{(\xi)}$ is a particular case of condition $\mathbf{P}_{\vec{\omega}}^{(\alpha)}$ for the case $\vec{\omega}=(1,\omega)$ and $\alpha=\xi=4+3\omega$. It can also be shown that condition A, B, C hold for the perturbed renewal equation (10) with $G_{\varepsilon}(z)$ given by (14). Applying Theorem 2 we obtain the following exponential asymptotic expansion for the ruin probability.

THEOREM 3

*Let the claim distributions $G_0(z)$ and $G_{\varepsilon}(z)$ be given by formulas (12) and (14). Let also condition **D** holds and $\varepsilon=T_0-T\ge 0$ be the perturbation parameter. Then there exists a unique non-negative solution, ρ_{ε}, to the characteristic equation (3) and the following asymptotical relation holds,*

$$\rho_{\varepsilon}=a_{11}\varepsilon^{1+\omega}+a_{2,2}\varepsilon^{2+2\omega}+a_{3,2}\varepsilon^{3+2\omega}++a_{3,3}\varepsilon^{3+3\omega}+a_{4,3}\varepsilon^{4+3\omega}+o(\varepsilon^{4+3\omega}), \quad (18)$$

where

$$a_{1,1}=\frac{\omega+2}{T_0^{\omega+2}},\quad a_{2,2}=\frac{(\omega+2)^3}{(\omega+3)T_0^{2\omega+3}},\quad a_{3,2}=-\frac{(\omega+1)(\omega+2)}{T_0^{2\omega+4}},$$
$$a_{3,3}=\frac{(\omega+2)^4}{T_0^{3\omega+4}}\left(\frac{\omega+1}{(\omega+3)^2}+\frac{\omega+5}{2(\omega+3)(\omega+4)}\right),$$
$$a_{4,3}=-\frac{3(\omega+1)(\omega+2)^3}{(\omega+3)T_0^{3\omega+5}}.$$

(i) *For any $0\le u_{\varepsilon}\to\infty$ in such a way that $\varepsilon^{[\beta]_{\omega}}u_{\varepsilon}\to\lambda_{\beta}\in[0,\infty)$ for some $1+\omega\le\beta<2+2\omega$, the following asymptotical relation holds,*

$$\Psi_{\varepsilon}(u_{\varepsilon})\to\exp\{-\lambda_{\beta}a_{1,1}\}\quad as\ \varepsilon\to 0.$$

(ii) *For any $0\le u_{\varepsilon}\to\infty$ in such a way that $\varepsilon^{[\beta]_{\omega}}u_{\varepsilon}\to\lambda_{\beta}\in[0,\infty)$ for some $2+2\omega\le\beta<3+2\omega$, the following asymptotical relation holds,*

$$\exp\left\{\left(a_{1,1}\varepsilon^{1+\omega}\right)u_{\varepsilon}\right\}\Psi_{\varepsilon}(u_{\varepsilon})$$
$$\to\exp\left\{-\lambda_{\beta}a_{2,2}\right\}\quad as\ \varepsilon\to 0.$$

(iii) *For any $0\le u_{\varepsilon}\to\infty$ in such a way that $\varepsilon^{[\beta]_{\omega}}u_{\varepsilon}\to\lambda_{\beta}\in[0,\infty)$ for $3+2\omega\le\beta<3+3\omega$,*

the following asymptotical relation holds,

$$\exp\left\{(a_{1,1}\varepsilon^{1+\omega}+a_{2,2}\varepsilon^{2+2\omega})\,u_\varepsilon\right\}\Psi_\varepsilon(u_\varepsilon)$$
$$\to \exp\left\{-\lambda_\beta\, a_{3,2}\right\} \quad as\ \varepsilon\to 0.$$

(iv) *For any* $0\le u_\varepsilon\to\infty$ *in such a way that* $\varepsilon^{[\beta]_\omega}u_\varepsilon\to\lambda_\beta\in[0,\infty)$ *for* $3+3\omega\le\beta<4+3\omega$,

the following asymptotical relation holds,

$$\exp\left\{(a_{1,1}\varepsilon^{1+\omega}+a_{2,2}\varepsilon^{2+2\omega}+a_{3,2}\,\varepsilon^{3+2\omega})\,u_\varepsilon\right\}\Psi_\varepsilon(u_\varepsilon)$$
$$\to \exp\left\{-\lambda_\beta\, a_{3,3}\right\} \quad as\ \varepsilon\to 0.$$

(v) *For any* $0\le u_\varepsilon\to\infty$ *in such a way that* $\varepsilon^{[\beta]_\omega}u_\varepsilon\to\lambda_\beta\in[0,\infty)$ *for* $\beta=4+3\omega$, *the*

following asymptotical relation holds,

$$\exp\left\{(a_{1,1}\varepsilon^{1+\omega}+a_{2,2}\varepsilon^{2+2\omega}+a_{3,2}\,\varepsilon^{3+2\omega}+a_{3,3}\varepsilon^{3+3\omega})\,u_\varepsilon\right\}\Psi_\varepsilon(u_\varepsilon)$$
$$\to \exp\left\{-\lambda_\beta\, a_{4,3}\right\} \quad as\ \varepsilon\to 0.$$

Remark 5. This example of perturbed risk process was first introduced in the author's earlier paper (Ni, Silvestrov and Malyarenko 2008). Theorem 3 above is an extended version of Theorem 3 in the aforementioned paper. The latter theorem gives the corresponding result for the perturbation condition $\mathbf{P}_\omega^{(\xi)}$ for $\xi=3+2\omega$ and hence presents the expansion of ρ_ε only up to and including the term of order $O(\varepsilon^{3+2\omega})$. In Theorem 3 above, we determine two more terms for the expansion of ρ_ε and consequently obtain two additional variants, i.e. statements (iv) and (v), of the exponential asymptotics for the ruin probability. The proof of these additional results follows the same line as the proof of Theorem 3 in Ni, Silvestrov and Malyarenko (2008).

4.2. Perturbed Risk Process with Multivariate Non-polynomial Perturbations

Let us suppose that the claim size distribution $G_\varepsilon(z)$ for the risk process (8) is a mixture of exponential distributions of the following form

$$G_\varepsilon(z) = 1-p_1e^{-z/\delta_1(\varepsilon)}-p_2e^{-z/\delta_2(\varepsilon)}-p_3e^{-z/\delta_3(\varepsilon)}, \tag{19}$$

where

$\delta_i(\varepsilon)=\delta_i-C_i\varepsilon^{\omega_i}>0,\ \delta_i>0, C_i>0,\ i=1,2,3,$ for $\varepsilon\ge 0,\ 0\le p_1,p_2,p_3\le 1,$

$p_1+p_2+p_3=1$, $\omega_1\equiv 1$, and $\omega_2,\omega_3>1$ take irrational values such that ω_2/ω_3 is an irrational number. Without loss of generality we assume $\omega_2<\omega_3$, so that we can introduce the vector parameter $\vec{\omega}=(1,\omega_2,\omega_3)$.

The perturbation above can be seen as an environmental factor that determines claim amounts and acts in a different form for different claim groups.

Note that if the perturbation parameter $\varepsilon=0$, $G_\varepsilon(z)$ reduces to

$$G_0(z)=1-p_1e^{-z/\delta_1}-p_2e^{-z/\delta_2}-p_3e^{-z/\delta_3}.$$

The first moment of the perturbed claim size distribution, μ_ε, takes the form

$$\mu_\varepsilon = p_1(\delta_1 - C_1\varepsilon) + p_2(\delta_2 - C_2\varepsilon^{\omega_2}) + p_3(\delta_3 - C_3\varepsilon^{\omega_3}). \quad (20)$$

By (19) and (20), we have $G_\varepsilon(z) \to G_0(z)$ as $\varepsilon \to 0$ for every $z \geq 0$, and $\mu_\varepsilon \leq \mu_0$ but $\mu_\varepsilon \to \mu_0$ as $\varepsilon \to 0$.

Obviously, $\alpha_\varepsilon \leq \alpha_0 = 1$ and $\alpha_\varepsilon \to \alpha_0 = 1$ as $\varepsilon \to 0$, so we have again the case of diffusion approximation for ruin probabilities.

Since $\alpha_\varepsilon < \alpha_0 = 1$ for $\varepsilon > 0$, the distribution $F_\varepsilon(u) = \alpha_\varepsilon \widetilde{G}_\varepsilon(u)$ is improper for $\varepsilon > 0$ but the limiting function $F_0(u)$ is proper, i.e. the defect $f_\varepsilon = 1 - \alpha_\varepsilon \to f_0 = 0$ as $\varepsilon \to 0$.

It can be shown that the defect f_ε and the r-th moment $m_{\varepsilon r}$ for the distribution function $F_\varepsilon(u) = \alpha_\varepsilon \widetilde{G}_\varepsilon(u)$ take the following form:

$$f_\varepsilon = \frac{p_1}{\mu_0} C_1\varepsilon + \frac{p_2}{\mu_0} C_2\varepsilon^{\omega_2} + \frac{p_3}{\mu_0} C_3\varepsilon^{\omega_3}, \quad (21)$$

$$m_{\varepsilon r} = m_{0r} + \frac{r!}{\mu_0} \sum_{i=1}^{3} [p_i \sum_{j=1}^{r+1} \binom{r+1}{j} (\delta_i)^{r+1-j} (-C_i\varepsilon^{\omega_i})^j], r \geq 1. \quad (22)$$

Relations (21) and (22) imply that, in this case, the perturbation condition $\mathbf{P}_{\vec{\omega}}^{(\alpha)}$ holds for any $\alpha \geq 1$. It can also be shown that condition A, B and C hold for the perturbed renewal equation (10) with $G_\varepsilon(z)$ given by (19).

Instead of reformulating Theorem 2 for this case, we illustrate the asymptotic result by a specific example where $\omega_2 = \sqrt{2}, \omega_3 = \sqrt{3}$, i.e. $\vec{\omega} = (1, \sqrt{2}, \sqrt{3})$ and $\alpha = 3$. In this case the following exponential asymptotic expansion for the ruin probability can be obtained by applying Theorem 2.

THEOREM 4

*Let the perturbed claim size distributions $G_\varepsilon(z)$ be given by formula (19) and ε be the perturbation parameter, let also condition **D** holds. Then :*

(i) *There exists a unique non-negative solution, ρ_ε, of the characteristic equation, (3) and the following asymptotical relation holds*

$$\rho_\varepsilon = a_{(1,0,0)}\varepsilon + a_{(0,1,0)}\varepsilon^{\sqrt{2}} + a_{(0,0,1)}\varepsilon^{\sqrt{3}} + a_{(2,0,0)}\varepsilon^2 + a_{(1,1,0)}\varepsilon^{1+\sqrt{2}} + a_{(1,0,1)}\varepsilon^{1+\sqrt{3}} + a_{(0,2,0)}\varepsilon^{2\sqrt{2}} + a_{(3,0,0)}\varepsilon^3 + o(\varepsilon^3) \quad (23)$$

where $a_{(1,0,0)} \ldots a_{(3,0,0)}$ can be calculated using recurrent formula (6) with the use of formulas (21) and (22), in particular,

$$a_{(1,0,0)} = \frac{p_1 C_1}{\mu_0 m_{01}}, \quad a_{(0,1,0)} = \frac{p_2 C_2}{\mu_0 m_{01}}, \quad a_{(0,0,1)} = \frac{p_3 C_3}{\mu_0 m_{01}},$$

$$a_{(2,0,0)} = \frac{p_1^2 C_1^2 (4\delta_1 m_{01} - m_{02})}{2\mu_0^2 m_{01}^3}, \ldots$$

(ii) *For any* $0 \le u_\varepsilon \to \infty$ *in such a way that* $\varepsilon^{[\beta]_\omega} u_\varepsilon \to \lambda_\beta \in [0, \infty)$ *for some* $1 \le \beta < 3$, *the following asymptotical relations holds,*

$$\exp\{(\sum_{1 \le \vec{n}\cdot\vec{\omega} < [\beta]_{\vec{\omega}}} a_{\vec{n}} \varepsilon^{\vec{n}\cdot\vec{\omega}}) u_\varepsilon\} \Psi_\varepsilon(u_\varepsilon) \to \exp\{-\lambda_\beta a^{(1)}\} \quad as\ \varepsilon \to 0, \tag{24}$$

where $a^{(1)} = a_{\vec{p}}$ *with* $\vec{p} = \vec{f}(\beta, \vec{\omega})$.

Remark 6. The expansion for ρ_ε, (23) is expanded only up to order $O(\varepsilon^3)$ in the example above. If needed, ρ_ε can be further expanded up to order of $O(\varepsilon^{[\alpha]_{\vec{\omega}}})$ for any real number $1 \le \alpha < \infty$.

Remark 7. Although the above asymptotic results are derived for specific parameter values, i.e. $\omega_2 = \sqrt{2}$ and $\omega_3 = \sqrt{3}$, similar results can be easily obtained for cases where the parameters ω_2 and ω_3 take other admissible values. Different choices of ω_2 and ω_3 only lead to different forms of the expansion for ρ_ε.

Remark 8. As in Theorem 3, statement **(ii)** of Theorem 4 leads to several variants of asymptotic relation (24) for different cases of the values for β, namely $1 \le \beta < \sqrt{2}$, $\sqrt{2} \le \beta < \sqrt{3}$, $\sqrt{3} \le \beta < 2$, ..., $2\sqrt{2} \le \beta < 3$ and finally $\beta = 3$. For instance, under the the balancing condition described in statement **(ii)**, we have, if $1 \le \beta < \sqrt{2}$, the asymptotic relation:

$$\Psi_\varepsilon(u_\varepsilon) \to \exp\{-\lambda_\beta a_{(1,0,0)}\} \quad as\ \varepsilon \to 0. \tag{25}$$

Similarly if $\sqrt{2} \le \beta < \sqrt{3}$ we obtain

$$\exp\{(a_{(1,0,0)}\varepsilon) u_\varepsilon\} \Psi_\varepsilon(u_\varepsilon) \to \exp\{-\lambda_\beta a_{(0,1,0)}\} \ as \ \ \varepsilon \to 0, \tag{26}$$

and following the same pattern, if $\sqrt{3} \le \beta < 2$ we obtain

$$\exp\{(a_{(1,0,0)}\varepsilon + a_{(0,1,0)}\varepsilon^{\omega_2}) u_\varepsilon\} \Psi_\varepsilon(u_\varepsilon) \to \exp\{-\lambda_\beta a_{(0,0,1)}\} \quad as \ \ \varepsilon \to 0,$$

and so on.

5. Experimental Study

In Section 5.1, we carry out experimental numerical studies for the example of perturbed risk process discussed in Section 4.1. The example introduced in Section 4.2 is investigated in Section 5.2.

5.1. Perturbed Risk Process with Bivariate Non-polynomial Perturbations

The asymptotic formulas given by statements **(i) - (v)** of Theorem 3 can serve as approximation methods for $\Psi_\varepsilon(u)$ for small value of ε and relatively large values of u. To gain insight into the accuracy and other properties of these asymptotic formulas, we

compare the corresponding approximations to the value of ruin probability estimated by computer simulation since the true value is difficult to compute.

Let us denote the simulated estimate of $\Psi_\varepsilon(u)$ by $\Psi_\varepsilon^s(u)$ with s standing for simulation. To obtain $\Psi_\varepsilon^s(u)$ we implement the conditional Monte Carlo simulation method, i.e. a variance reduced version of the Crude Monte Carlo, via the Pollaczeck-Khinchine formula for ruin probabilities. The description of this simulation method can be found in Asmussen (2000). The solution to our problem is $\Psi_\varepsilon^s(u) = \mathrm{E}(Z)$ where Z is the random variable generated in Algorithm 1 below. Note that u is the chosen value of the initial reserve, T is the constant parameter of $G_\varepsilon(z)$ given in (14) and hence the corresponding constant in $\widetilde{G}_\varepsilon(u)$ defined in (11).

Algorithm 1

1. Generate geometric random variable K, with $P(K=k) = (1-\alpha_\varepsilon)\alpha_\varepsilon^k$.
2. **if** $K = 0$ **then** $Z \leftarrow 0$.
3. **else if** $K = 1$ **then** $Z \leftarrow 1 - \widetilde{G}_\varepsilon(u)$.
4. **else**
5. Generate $L_1, \ldots, L_{K-1}$ from distribution $\widetilde{G}_\varepsilon(\cdot)$.
6. let $Y \leftarrow u - (L_1 + \cdots + L_{K-1})$.
7. **end if**
8. **if** $Y < 0$ **then** $Z \leftarrow 1$. **else if** $Y > T$ **then** $Z \leftarrow 0$.
9. **else** $Z \leftarrow 1 - \widetilde{G}_\varepsilon(Y)$.
10. **end if**

The main problem in Algorithm 1 is to simulate the random variable from the distribution $\widetilde{G}_\varepsilon(\cdot)$. We use the inverse method to generate outcomes of this random variable, i.e. to generate

$$x = T_0 - [T_0^{\omega+1}(1-v) + \varepsilon^{\omega+1} v]^{\frac{1}{\omega+1}} \qquad 0 \le v \le 1, \tag{27}$$

where v is a realization of a standard uniform random variable.

Set the parameters $T_0 = 1$, $\omega = (4+\sqrt{2})/5$, the simulation experiments have been carried out for different combinations of initial capital u and the perturbation parameter ε, with concentration on the cases where the ruin probability is of the magnitude 10^{-2}, 10^{-3} and 10^{-5}. For each simulation experiment, we execute the block in Algorithm 1 for 100 million times, i.e., to generate 100 million replicates of random variable Z. When the ruin probability is as small as of magnitude 10^{-5}, we increase the number of simulations to 500 million times.

Let us denote $\Psi_\varepsilon^j(u), j = 1, \ldots, 5$ as the approximated ruin probability via the j-th statement in Theorem 3. By inspecting statement **(i) - (v)** we note that $\Psi_\varepsilon^j(u)$ represents

the approximation where the j-th order expansion of ρ_ε is used in the corresponding asymptotic formula. Let us call $\Psi_\varepsilon^j(u)$ the j-th order approximation.

The relative errors $E_j(u,\varepsilon), j =,1\ldots,5$ are calculated in the traditional way as

$$E_j(u,\varepsilon) = \frac{\Psi_\varepsilon^j(u) - \Psi_\varepsilon^s(u)}{\Psi_\varepsilon^s(u)},$$

and they are presented in Table 1. The value of safety loadings, η_ε, are also given in the table.

Table 1. Relative errors of the approximation by Theorem 3 with $T_0 = 1, \omega = (4+\sqrt{2})/5$

ε (η_ε)			Relative errors $E_j(u,\varepsilon)$ (%)				
	u	$\Psi_\varepsilon^s(u)$	E_1	E_2	E_3	E_4	E_5
0.05 (0.2%)	500	0.0487	1.51	0.13	0.19	0.18	0.18
	800	0.0080	2.42	0.21	0.31	0.29	0.29
	1600	6.35×10^{-5}	4.37	-0.10	0.10	0.07	0.07
0.1 (0.8%)	100	0.0743	5.3	0.27	0.71	0.57	0.59
	200	0.0055	10.4	0.13	1.01	0.74	0.79
	400	3.121×10^{-5}	20.3	-1.11	0.64	0.08	0.18
0.15 (2.0%)	50	0.0450	14.6	0.41	2.21	1.32	1.57
	100	0.0021	29.7	-0.55	3.05	1.27	1.76
	170	2.69×10^{-5}	56.1	-0.58	5.62	2.54	3.37
0.2 (3.6%)	30	0.0301	30.6	0.28	5.13	1.87	3.05
	50	0.0030	52.9	-1.48	6.59	1.13	3.09
	90	2.86×10^{-5}	111.8	-4.03	10.58	0.59	4.13
0.3 (8.9%)	10	0.0492	65.1	2.54	16.52	2.06	9.75
	20	0.0026	155.7	-1.36	27.38	-2.28	13.01
	35	3.08×10^{-5}	394.0	-6.69	45.95	-8.22	18.37
0.35 (12.7%)	8	0.0290	115.7	4.60	31.22	-0.61	18.73
	15	0.0014	292.7	1.06	54.59	-8.18	28.15
	24	2.91×10^{-5}	747.5	-3.41	90.68	-17.14	41.25

The first impression of Table 1 is $E_1(u,\varepsilon)$ is far too large when $\varepsilon \geq 0.1$ for all chosen values of u. This suggests that the first order approximation $\Psi_\varepsilon^1(u)$ is not adequate unless ε is really small, thus the contribution of the second term in (18) is definitely not negligible. For $\varepsilon \leq 0.2$, the higher order approximations, namely $\Psi_\varepsilon^j(u), j = 2,\ldots,5$ are good, except that the approximation by $\Psi_\varepsilon^3(u)$ does not work so well for $\varepsilon = 0.2$, which may be caused by some special property of the expansion (18).

As seen from Table 1, if ε is relatively large, say $\varepsilon \geq 0.2$, even higher order approximations work poorly. Interestingly, in general approximation by $\Psi_\varepsilon^2(u)$ still seem to be applicable.

These experiments are done for T_0 normalized to 1, and ω set to equal to $(4+\sqrt{2})/5$. Similar experiments have been done for $T_0 = 1, \omega = \sqrt{2}$, and the general impression of the results is the same: first order approximation ought not to be used for moderate and large ε; all approximations get more accurate as ε gets smaller as expected. The quality of approximations seems to depend heavily on ε but not so much on the values of u in the chosen range.

It can be shown that in this model of a perturbed risk process, the safety loading η_ε is of the order $O(\varepsilon^{1+\omega})$. Therefore as shown in Table 1, the approximations are applicable only for η_ε being very small. This may not be the most interesting case in risk theory. However, we would like to note that there's a duality of the classical risk process with the virtual waiting time process in a M/G/1 queue, consequently the ruin probability can be interpreted as the steady-state limit of the virtual waiting time. The case when η_ε is very small corresponds to the interesting heavy traffic case in the queuing theory and thus the study of this case has its own value.

Finally, we compare the approximation by $\Psi_\varepsilon^2(u)$ to the classical diffusion approximation method (see for example Grandell 2000),

$$\Psi_\varepsilon^D(u) = \exp(-(2\beta_\varepsilon \eta_\varepsilon u)/\gamma_\varepsilon), \tag{28}$$

where $\beta_\varepsilon, \gamma_\varepsilon$ refer to the first and second moment of claim size distribution $G_\varepsilon(z)$, η_ε is the safety loading. The results are presented in Table 2, with $E_D(u,\varepsilon)$ refer to the relative error of the approximation by (28).

As seen from Table 2, in this numerical example, approximation by $\Psi_\varepsilon^2(u)$ works better. $E_D(u,\varepsilon)$ tends to get larger as ε get larger and also as u gets larger. This is the case for $\varepsilon > 0.2$ as well (not shown in the table).

5.2. Perturbed Risk Process with Multivariate Non-polynomial Perturbations

We consider a numerical example for the application in section (4.2). Suppose that $C_1 = C_2 = C_3 = 1$, $p_1 = 0.4, p_2 = 0.3, p_3 = 0.3$, $\delta_1 = 3, \delta_2 = 5, \delta_3 = 7$ in the perturbed claim size distribution (19).

Since the claim distribution $G_\varepsilon(z)$ is a mixture of three exponential distributions, exact formula of ruin probability for this case exists in terms of a matrix-exponential function (see for example Asmussen 2000). Let us denote the ruin probability calculated via the exact formula by $\Psi_\varepsilon^e(u)$. We then compare it to the approximated ruin probabilities via statement **(ii)** of Theorem 4. Let us denote $\Psi_\varepsilon^j(u), j = 1,\ldots 8$ as these approximated ruin probabilities with β chosen in such a way that the j-th order expansion of ρ_ε is used in (24). For example, $\Psi_\varepsilon^1(u)$, call it the first order approximation, is calculated using (25) in Remark 8 where the parameter β satisfies $1 \le \beta < \sqrt{2}$, and $\Psi_\varepsilon^2(u)$, i.e. the second order approximation, refers to the approximation by (26) and so on.

Table 2. Relative errors of $\Psi_\varepsilon^j(u)$ and $\Psi_\varepsilon^D(u)$

ε (η_ε)	u	$\Psi_\varepsilon^s(u)$	$E_2(u,\varepsilon)(\%)$	$E_D(u,\varepsilon)(\%)$
0.05 (0.2%)	500	0.0487	0.13	0.26
	800	0.0080	0.21	0.42
	1600	6.35×10^{-5}	-0.10	1.34
0.1 (0.8%)	100	0.0743	0.27	1.03
	200	0.0055	0.13	2.43
	400	3.121×10^{-5}	-1.11	6.12
0.15 (2.0%)	50	0.0450	0.41	2.96
	100	0.0021	-0.55	7.11
	170	2.69×10^{-5}	-0.58	11.47
0.2 (3.6%)	30	0.0301	0.28	6.44
	50	0.0030	-1.48	12.25
	90	2.86×10^{-5}	-4.03	22.08

The relative errors $E_j(u,\varepsilon), j=1,\ldots,8$, defined as

$$E_j(u,\varepsilon)=\frac{\Psi_\varepsilon^j(u)-\Psi_\varepsilon^e(u)}{\Psi_\varepsilon^e(u)},$$

are calculated for different combinations of ε and u and presented in Table 3. All calculations are done in MATLAB. Symbol η_ε in Table 3 refers to the safety loading coefficient.

As shown in Table 3, the higher order approximations, i.e. $\Psi_\varepsilon^j(u), j\geq 3$ are perfect for ε small, say $\varepsilon\leq 0.05$. For $0.05<\varepsilon\leq 0.15$, even higher order approximations $\Psi_\varepsilon^j(u), j\geq 6$ should be used. Also it is seen from the table that, for a fixed u and a fixed small ε, the relative errors appear to decrease, when we include more terms from the expansion ρ_ε in the approximation, with the exception that $E_8(u,\varepsilon)$ is oftentimes slightly larger than $E_7(u,\varepsilon)$.

Figure 1 illustrates how the approximation in general improves as we take approximations of higher orders. Approximation $\Psi_\varepsilon^1(u)$ is shown to be a very poor approximation and is therefore omitted in the figure.

Table 3. Relative errors of the approximation by Theorem 4 with $C_1=C_2=C_3=1$, $p_1=0.4, p_2=0.3, p_3=0.3,\ \delta_1=3,\delta_2=5,\delta_3=7$

ε (η_ε)	u	$\Psi_\varepsilon^e(u)$	Relative errors $E_j(u,\varepsilon)$ (%)							
			E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8
0.01 (0.0009)	17000	0.0499	44	7.1	0.14	0.13	0.12	0.11	0.10	0.11
	33000	0.0030	102	14	0.18	0.16	0.13	0.11	0.11	0.11
	57000	4.31×10^{-5}	237	26	0.23	0.21	0.14	0.11	0.11	0.11
0.05 (0.005)	3000	0.0481	103	22	0.99	0.96	0.79	0.67	0.63	0.64
	6000	0.0023	310	50	1.48	1.33	0.99	0.74	0.67	0.69
	10000	4.11×10^{-5}	947	95	1.91	1.83	1.26	0.84	0.72	0.76
0.15 (0.020)	1000	0.0273	258	62	4.65	4.57	3.75	2.88	2.61	2.71
	1500	0.0046	570		5.93	5.81	4.56	3.26	2.85	3.00
	3000	2.12×10^{-5}	4297	305	9.86	9.61	7.04	4.38	3.57	3.86
0.25 (0.037)	500	0.0343	320	85	9.34	9.23	7.75	5.95	5.36	5.59
	800	0.0046	870	162	12.7	12.5	10.04	7.11	6.16	6.53
	1500	4.37×10^{-5}	6735	487	20.8	20.4	15.6	9.86	8.05	8.75

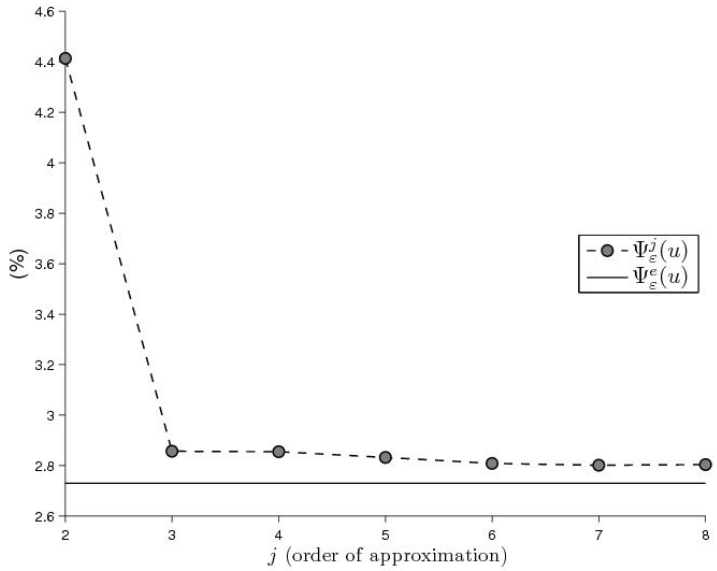


Figure 1. Approximation by $\Psi_\varepsilon^j(u)$ for $\varepsilon = 0.15; u = 1000.$

From Table 3 we note also that, for a fixed $\varepsilon \geq 0.05$ and a fixed approximation, if u takes a larger value, the corresponding relative error appears to be larger. This seems to be contradictory to the fact that formula (24) holds for $u \to \infty$ and $\varepsilon \to 0$ simultaneously. However, note that to use formula (24) we should have $u \to \infty$ and $\varepsilon \to 0$ balanced so that $\varepsilon^{[\beta]_\omega} u_\varepsilon \to \lambda_\beta \in [0, \infty)$ for some $1 \leq \beta < 3$. Hence the value of λ_β can have a subtle effect on the quality of approximation. The experiments suggest a relatively too large λ_β may not be desirable for a good approximation. For the purpose of illustration, let us consider the approximation by $\Psi_\varepsilon^3(u)$ with $\varepsilon = 0.15$ and varying u, of which the values of λ_β are given in Table 4:

Table 4. The values of λ_β for $\Psi_\varepsilon^3(u)$ with $\varepsilon = 0.15$ and varying u

ε	u	$\Psi_\varepsilon^e(u)$	$E_3(u,\varepsilon)$ (%)	λ_β
0.15	1000	0.0273	4.65	37.4
	1500	0.0046	5.93	56.1
	3000	2.12×10^{-5}	9.86	112.2

We note from Table 4 that when λ_β is as large as 112.2, the approximation is less accurate for the cases with smaller values of λ_β. To address questions like whether the values of λ_β always affect the quality of approximation, and if this is true which value of λ_β is optimal for the approximation, more comprehensive and extensive numerical experiments are required.

6. Conclusions and Future Research

We have studied the asymptotic behavior of nonlinearly perturbed equations with non-polynomial perturbations of the type $\mathbf{P}_{\tilde{\omega}}^{(\alpha)}$ which is a generalized type of the non-polynomial perturbations treated in the previous research (Ni, Silvestrov, Malyarenko 2008). The theoretical results have been applied to examples of nonlinearly perturbed risk processes and can have potential applications in various applied probability models. For the proofs of the results we refer to a forthcoming report by Ni (2010).

This article has dealt with asymptotically proper perturbed renewal equation, i.e. as described in condition A, $F_0(t)$ is assumed to be a proper distribution function. The case of asymptotically improper perturbed renewal equation where $F_0(t)$ can be improper leads to a further generalization of the theory and will be studied at the next stage of research. The study of asymptotic expansions for renewal limits follows naturally afterwards.

References

1. Asmussen, S. **Ruin Probabilities**, Singapore: World Scientific, 2000
2. Englund, E. **Nonlinearly perturbed renewal equations with applications**, Ph.D. Thesis, Umeå University, Sweden, 2001
3. Englund, E. and Silvestrov, D.S. **Mixed large deviation and ergodic theorems for regenerative processes with discrete time**, Theory Stoch. Process., 3(19), no. 1-2, 1997, pp. 164-176
4. Feller, W. **An Introduction to Probability Theory and Its Applications**, Vol. II, New-York: Wiley, 1966
5. Grandell, J. **Simple approximations of ruin probability**, Insurance: Math. Econom., 26(2-3), 2000, pp. 157-173
6. Gyllenberg, M. and Silvestrov, D.S. **Quasi-stationary phenomena in nonlinearly perturbed stochastic systems**, De Gruyter Expositions in Mathematics, 44, Berlin: Walter de Gruyter, 2008
7. Ni, Y., Silvestrov, D. and Malyarenko, A. **Exponential asymptotics for nonlinearly perturbed renewal equation with non-polynomial perturbations**, J. Numer. Appl. Math., 1(96), 2008, pp. 173-197
8. Ni, Y. **Nonlinearly perturbed renewal equations: The non-polynomial case**, Research Report, Mälardalen University, 2010 (to appear)
9. Silvestrov, D.S. **A generalization of the renewal theorem**, Reports of the Ukrainian Academy of Sciences, Ser. A, no. 11, 1976, pp. 978-982
10. Silvestrov, D.S. **The renewal theorem in a series scheme 1**, Theory Probab. Math. Statist., 18, 1978, pp. 155-172
11. Silvestrov, D.S. **The renewal theorem in a series scheme 2**, Theory Probab. Math. Statist., 20, 1979, pp. 113-130
12. Silvestrov, D.S. **Exponential asymptotic for perturbed renewal equations**, Theory Probab. Math. Statist., 52, 1995, pp. 153-162

JOURNAL OF APPLIED QUANTITATIVE METHODS

International Symposium on Stochastic Models in Reliability Engineering, Life Sciences and Operations Management (SMRLO'10)

A NON-PARAMETRIC TEST FOR A CHANGE-POINT IN LINEAR PROFILE DATA

Wolfgang BISCHOFF
Prof., Faculty of Mathematics and Geography,
Catholic University Eichstätt-Ingolstadt,
Eichstätt, Germany

E-mail: wolfgang.bischoff@ku-eichstaett.de

Andreas GEGG
PhD Candidate, Faculty of Mathematics and Geography,
Catholic University Eichstätt-Ingolstadt,
Eichstätt, Germany

E-mail: andreas.gegg@ku-eichstaett.de

Abstract: *We propose a change-point approach for testing the constancy of regression parameters in a linear profile data set (panel data in econometrics).*
Each sample collected over time in the historical data set consists of several multivariate observations for which a linear regression model is appropriate. The question now is whether all of the profiles follow a linear regression model with the same parameter vector or whether a change occurred in one or more model parameters after a special sample.
We use the partial sum operator in several dimensions to test the null hypothesis "H_0: no change-point occurred" and propose a non-parametric size α-test.
In Bischoff and Gegg (2010) we compared our proposed method with the likelihood-ratio-test by Mahmoud et al. (2007) in a simulation study. By these simulations we could show that our procedure can, in contrast to the likelihood-ratio-test, even be applied to the non-normal case. In this paper, however, we show how to compute our proposed test statistic step-by-step by considering an artificial data set.

Key words: *change-point problem; panel data; statistical process control; linear regression*

1. Introduction

We investigate a linear profile data set for change-points. In economics, linear profile data are also known as, "panel data". Note that linear profile data are assumed to be ordered in a natural way. In most of the applications the profiles will be sampled sequentially and so time is the ordering variable. We attack this problem by using the results given in Bischoff and Gegg (2010), where a linear regression model with p-variate response was considered. In order to find change-points in regression models we investigated there the partial sums of the least squares residuals. Without specifying the error a nonparametric test (as for instance a test of Kolmogorov-Smirnov type) can then be applied to the limit process of the partial sums in order to test whether a change-point does or does not occur. MacNeill (1978a,b) and Bischoff (1998, 2002) give basic theoretical results concerning the residual partial sums process for univariate response, whereas Bischoff (2010) demonstrates this approach in case of univariate regression by using an example from quality control. Note that the residual partial sums technique can also be used to check asymptotically for regression with multivariate correlated response (Bischoff and Gegg, 2010).

Our proposed method is a two-step-procedure: In a first step, we estimate the parameter vectors for every profile. In a second step we analyze these estimations which build a linear model with multiple correlated response under the null hypothesis that the profile data have no change-point.

Mahmoud et al. (2007) also attacked the described change-point problem and proposed a modification of a likelihood ratio test (LRT) for the case of simple linear regression with normally distributed error terms. By asymptotic considerations our method does not need assumptions about the distribution of the error terms and so it is more robust against departure from normal distribution, see Bischoff and Gegg (2010). A further advantage of our procedure is that the alternative hypothesis does not have to be specified.

2. Linear Profile Data

In practice, one often wants to test whether all of a fixed number m, say, of independent samples follow the same known linear model. To be more precise let

$$W^{(i)} = \boldsymbol{X}\,\beta^{(i)} + \varepsilon^{(i)}, \qquad \beta^{(i)} \in \mathbb{R}^p \text{ unknown}, \tag{1}$$

be a linear model for every profile $i \in \{1,\ldots,m\}$, where

(A1) $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the corresponding design matrix of explanatory variables with $\operatorname{rank}(\boldsymbol{X}) = p \le n$,

(A2) and $\varepsilon^{(i)}$ is the vector with iid components $\varepsilon_1^{(j)},\ldots,\varepsilon_n^{(j)}$ having mean 0 and variance σ^2.

Note that assumption **(A1)** in particular claims that the same design is used for each profile j. Furthermore, since different profiles are assumed to be independent, so are $\varepsilon^{(1)},\ldots,\varepsilon^{(m)}$. In the sequel we assume model (1), together with the assumptions **(A1) - (A2)**, to be true for each j. Model (1) is called the "j-th Linear Profile" and the aim is to test for a change-point in the parameter vector. Since the profiles have a natural ordering we can formulate the corresponding hypothesis by

$$H_0 : \beta = \beta^{(1)} = \ldots = \beta^{(m)} \quad \text{vs.} \quad H_1 : \exists m_0 \in \{1,\ldots,m-1\} : \beta^{(1)} = \beta^{(2)} = \ldots = \beta^{(m_0)} \neq \beta^{(m_0+1)} \tag{2}$$

and so the testing problem is indeed a change-point problem. In order to check (2) we estimate $\beta^{(i)}$ by the least-squares estimator $\hat{\beta}^{(j)}$. With our assumptions **(A1)-(A2)** we have $\hat{\beta}^{(j)} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T W^{(i)}$ with

$$\mathbf{E}(\hat{\beta}^{(j)}) = \beta^{(i)} \quad \text{and} \quad \mathbf{Cov}(\hat{\beta}^{(j)}) = \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1} =: \Sigma. \tag{3}$$

In case σ^2 is unknown, our proposed procedure can also be used by replacing σ^2 with a consistent estimator for σ^2 under H_0. So we can assume without loss of generality, Σ is a known positive definite matrix. Furthermore $\hat{\beta}^{(1)},\ldots,\hat{\beta}^{(m)}$ are independent since the different samples $W^{(1)},\ldots,W^{(m)}$ are assumed to be independent. Let $Y := \begin{pmatrix} \hat{\beta}^{(1)^T} \\ \vdots \\ \hat{\beta}^{(m)^T} \end{pmatrix}$ be the $m \times p$ matrix containing the least-squares estimations and let $Z := \begin{pmatrix} \hat{\beta}^{(1)^T} - \beta^T \\ \vdots \\ \hat{\beta}^{(m)^T} - \beta^T \end{pmatrix}$. If (1) and (2) hold true, then (3) leads to the following model:

$$Y = \mathbf{1}_m \beta^T + Z \quad \text{with} \quad \mathbf{EZ} = \mathbf{0},\ \mathbf{Cov}(\mathrm{vec}(Z^T)) = \mathbf{I}_m \otimes \Sigma \quad \text{and} \quad \beta \in \mathbb{R}^p \ \text{unknown parameter vector.} \tag{4}$$

Thereby $1_m \in \mathbb{R}^m$ is the vector whose components are all equal to 1, I_m is the $m \times m$ identity matrix, "$\otimes$" denotes the Kronecker-Product and "vec" is the well-known vec-operator (Harville, 1997).

Conversely, if (2) is false and (1) together with **(A1)–(A2)** still holds true, then a change-point occured and (4) does not hold. Therefore we can test hypothesis (2) by checking the linear model (4). Bischoff and Gegg (2010) formulated a procedure which can be used to check a more general model by using multiple partial sum processes. Below we apply this method to our problem.

3. Residual Partial Sums Process

In order to test the hypotheses (2), we investigate the partial sums of the p-dimensional residuals in model (4). For that we use the partial sum operator T_m, which embeds a vector $a = (a_1,\ldots,a_m)^T \in \mathbb{R}^m$ in the space $C([0,1])$ by

$$T_m\left(\begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}\right)(z) = \sum_{i=1}^{\lfloor mz \rfloor} a_i + (mz - \lfloor mz \rfloor) a_{\lfloor mz \rfloor + 1}, \quad z \in [0,1],$$

where $\lfloor z \rfloor := \max\{l \in Z \mid l \leq z\}$ and $\sum_{i=1}^{0} a_i = 0$. Figure 1 shows the resulting graph of the partial sum operator T_m applied to a vector $a \in R^m$. The partial sum operator $T_{m \times p}$ embeds

$R^{m\times p} = R^m \times \ldots \times R^m$ in the space $C([0,1], R^p) = C([0,1]) \times \ldots \times C([0,1])$.

We define $T_{m\times p}$ with the help of T_m . For this, let $A \in R^{m\times p}$ be an $m \times p$ matrix with columns $a^{(1)},\ldots,a^{(p)}$, then:

$$T_{m\times p} : \begin{cases} R^{m\times p} \to C([0,1]) \times \ldots \times C([0,1]) \\ A \mapsto T_{m\times p}(A)(z) = (T_m(a^{(1)})(z), \ldots, T_m(a^{(p)})(z)), \quad z \in [0,1] \end{cases} .$$

The partial sum operator has gained a lot of interest especially because of the well-known Donsker-Theorem for an iid sequence of centered random variables. Iglehart (1968) formulated a vector-valued version of this theorem:

Theorem 1

Let $(\xi_i)_{i\geq 1}$ *be an iid sequence of random variables with values in* R^p *and* $E\xi_1 = 0$, $Cov(vec(Z^T)) = I_m \otimes \Sigma$

with Σ *positive definite. Then:*

$$\frac{1}{\sqrt{m}} \Sigma^{-1/2} T_{m\times p} \left(\begin{pmatrix} \xi_1^T \\ \vdots \\ \xi_m^T \end{pmatrix}^T \right) \xrightarrow{D} B^p \quad \text{with } m \to \infty,$$

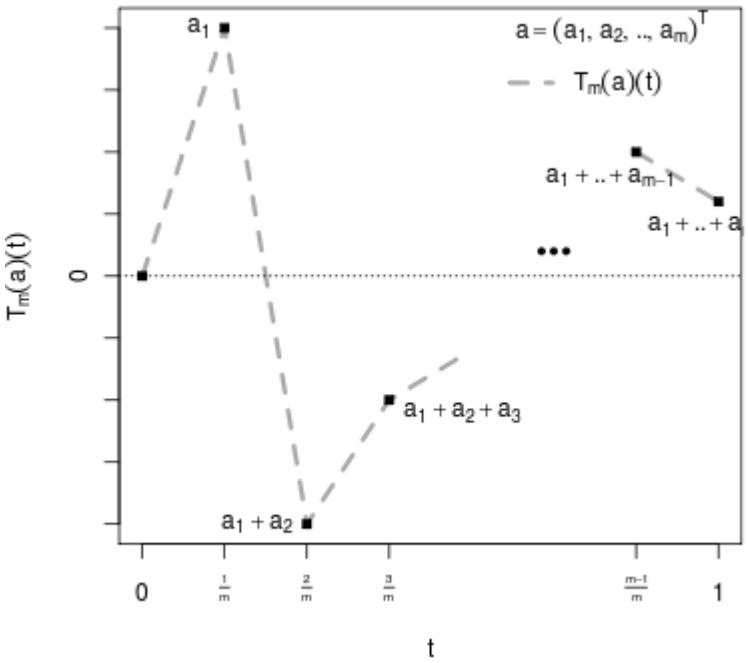


Figure1. Partial sum process $T_m(a)$

whereas B^p *is the p-dimensional Brownian motion with independent components and* „ $\xrightarrow{D}$ *''means weak convergence.*

The residuals of the linear model (4) are correlated and through this they do not fulfil the iid assumption of the preceding theorem. However, Bischoff and Gegg (2010) used the vector-valued version of the Donsker-Theorem to establish the *p*-dimensional residual

partial sums process in case of a multivariate linear model with multiple response. It is a projection of the Brownian motion B^p on a certain subspace. As a special case, we state the following result:

Theorem 2

Consider model (4), i.e. $Y = \mathbf{1}_m \beta^T + Z$ *with* $\mathbf{EZ} = \mathbf{0}$, $\mathbf{Cov}(vec(Z^T)) = \mathbf{I}_m \otimes \Sigma$ *and* $\beta \in R^p$ *unknown parameter vector.*

Then, under "$H_0 : \beta = \beta^{(1)} = \ldots = \beta^{(m)}$*", we have for the residuals* $Y - \hat{Y}$ *and* $m \to \infty$,

$$\frac{1}{\sqrt{m}} \Sigma^{-1/2} T_{m \times p} (Y - \hat{Y})^T \xrightarrow{D} B_0^p, \qquad (5)$$

where B_0^p *is the p-dimensional Brownian bridge.*

4. Test for Linear Profile Data

Under the null hypothesis the residual partial sums limit process (cf. Theorem 2) is given by B_0^p, the so called standard *p*-dimensional Brownian bridge on [0,1]. An intuitive one-dimensional test statistic is the maximum of the Euclidean norm of the *p*-dimensional process. To be more precise let

$$R_m(t) := \frac{1}{\sqrt{m}} \Sigma^{-1/2} T_{m \times p} (Y - \hat{Y})^T (t), \qquad t \in [0,1].$$

Then our proposed test statistic is $\max_{t \in [0,1]} \| R_m(t) \|$, where $\| \cdot \|$ is the Euclidean norm in R^p, i.e. $\| (x_1, \ldots, x_p)^T \|^2 = \sum_{i=1}^{p} x_i^2$. Because of the "Continuous Mapping Theorem" (Billingsley 1999), we have the following convergence under H_0:

$$\| R_m \| \xrightarrow{D} \| B^p \| \qquad \text{for } m \to \infty. \qquad (6)$$

Note that the limit process is the well-known Bessel bridge. In order to check (4) we apply a test of Kolmogorov-Smirnov type to the Bessel bridge and we get an asymptotic size α-test, $\alpha \in (0,1)$, by

Reject $H_0 \Leftrightarrow \sup_{t \in [0,1]} \| R_m(t) \| > k_\alpha$.

Thereby $\kappa_\alpha > 0$ is a constant such that $P\left(\sup_{t \in [0,1]} \| B_0^p(t) \| > k_\alpha\right) = \alpha$. Note that for given α, the corresponding value κ_α can be explicitly calculated. Kiefer (1959) gives a closed form for the cumulative distribution function of the Bessel bridge. We cite from his article concrete values for κ_α in case *p* = 2,...,5 and α = 0.1, 0.05, 0.01 in Table 1:

Table 1. Critical values for the *p*-dimensional Bessel bridge

κ_α	*p* = 2	*p* = 3	*p* = 4	*p* = 5
α=0.1	1.45399	1.61960	1.75593	1.87462
α=0.05	1.58379	1.74726	1.88226	2.00005
α=0.01	1.84273	2.00092	2.13257	2.24798

5. Numerical Example

5.1. Profile Data and Parameter Estimation

As a concrete example of application, we study the situation, that a change takes place after profile 4 of *m* = 6 profiles – both in the intercept and in the quadratic term in a quadratic model. For each profile, we simulated *n* = 10 observations according to

$$W_i^{(j)} = \alpha_1^{(j)} + x_i + \alpha_2^{(j)} x_i^2 + 2 \cdot \varepsilon_{ij}, i = 1,...,10, j = 1,...,6.$$

Thereby

- $\alpha_1^{(j)} = 0$ for *j* = 1,...,4 and $\alpha_1^{(j)} = 1$ for *j* = 5, 6 (shift in intercept)
- $\alpha_2^{(j)} = 0.1$ for j =1,...,4 and $\alpha_2^{(j)} = 0.12$ for *j* = 5, 6 (shift in quadratic term)
- $x_1 = 0, x_2 = \frac{10}{9}, x_3 = \frac{20}{9}, ..., x_{10} = 10,$
- ε_{ij} is a sequence of standardized iid random variables having lognormal distribution.

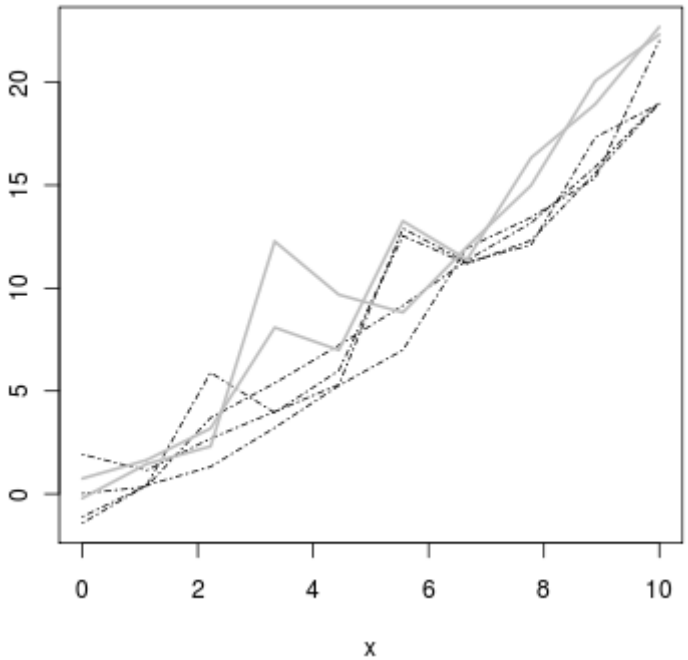


Figure 2. Simulated profiles $W^{(1)}, ..., W^{(4)}$ (color in plot: black) and $W^{(5)}, W^{(6)}$ (grey).

Figure 2 shows the simulated profile data under study. Let

$$\mathbf{X} := \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & & \\ 1 & x_{10} & x_{10}^2 \end{pmatrix} \in R^{m \times p}$$

be the design matrix for each profile. Consequently, we have $p = rank(\mathbf{X}) = 3$. We fit for each profile the model

$$W^{(j)} = \mathbf{X} \begin{pmatrix} \beta_0^{(j)} \\ \vdots \\ \beta_2^{(j)} \end{pmatrix} + \varepsilon^{(j)}, \qquad (7)$$

where $\varepsilon^{(j)}$ is a random vector with $\mathbf{E}\,\varepsilon^{(j)} = 0$ and $\mathbf{Cov}\,\varepsilon^{(j)} = \sigma^2 \mathbf{I}_p$. We estimate the coefficients by least squares method and get the values $\beta^{(1)}, \ldots, \beta^{(6)}$ shown in Figure 3. With these estimations, we can fit model (4) with $\Sigma = \sigma^2 (XX^T)^{-1}$.

Furthermore, by model (7), we get for each profile *j* an estimation for the variance, namely the usual variance estimation

$$\hat{\sigma}_j^2 := \frac{1}{10-3} (W^{(j)^T} (\mathbf{I}_3 - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) W^{(j)}), \; j = 1, \ldots, 6.$$

Consequently, with our assumptions **(A1)-(A2)**, we can estimate σ^2 by

$$\hat{\sigma}^2 := \frac{1}{m} \sum_{j=1}^{m} \hat{\sigma}_j^2 = \frac{1}{6} \sum_{j=1}^{6} \hat{\sigma}_j^2.$$

In case of the simulated data mentioned above, we have $\hat{\sigma}^2 = 3.004228.$

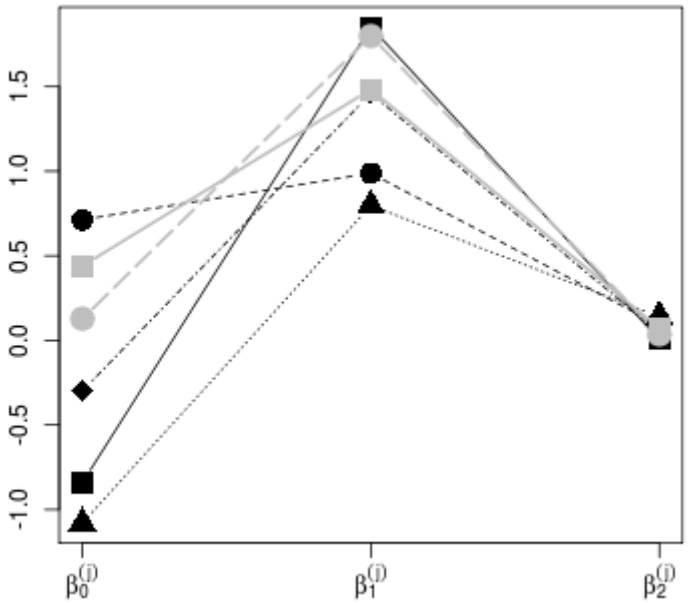


Figure 3. Estimations for parameter vector $\beta^{(1)}, \ldots, \beta^{(4)}$ (black) and $\beta^{(5)}, \beta^{(6)}$ (grey)

5.2. Test for a Change-Point

Now we are in the position, to calculate our proposed test statistic. Therefore, with $\mathrm{Y} := \begin{pmatrix} \hat{\beta}^{(1)^{\mathrm{T}}} \\ \vdots \\ \hat{\beta}^{(6)^{\mathrm{T}}} \end{pmatrix}$ and $\hat{Y} \in R^{6\times 3}$ being the matrix of estimated values in (4), we get:

$$\mathrm{R}_6(\mathrm{t}) := \frac{1}{\sqrt{6}\hat{\sigma}} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{1/2}\, \mathrm{T}_{6\times 3}(\mathrm{Y}-\hat{\mathrm{Y}})^{\mathrm{T}}(\mathrm{t}), \mathrm{t}\in[0,1].$$

Then we can determine the value of our test-statistic $\max_{t\in[0,1]} \| R_6(t) \|$ and compare with the critical values given in Table 1.

Figure 4 shows the process $\| R_6(t) \|$ for the data set under study. We get a value for the test statistic of 2.01976 and so we can reject the null hypothesis "H_0 no change-point" even for $\alpha = 0.01$ since the corresponding critical value is $\mathrm{k}_{0.01} = 2.00092.$

Note that, by using the same random numbers in case "no change-point" (i.e. $\alpha_1^{(\mathrm{j})} = 0, \alpha_2^{(\mathrm{j})} = 0.1$ for all *j* = 1,...,6), the test statistic is 0.5036005 and so we consistently cannot reject the null hypothesis to usual sizes.

Consequently, our proposed method leads to good results even in the case of small shifts (see Figure 2) and also in case of non-normal error terms (in the example, we have used log-normally distributed error terms).

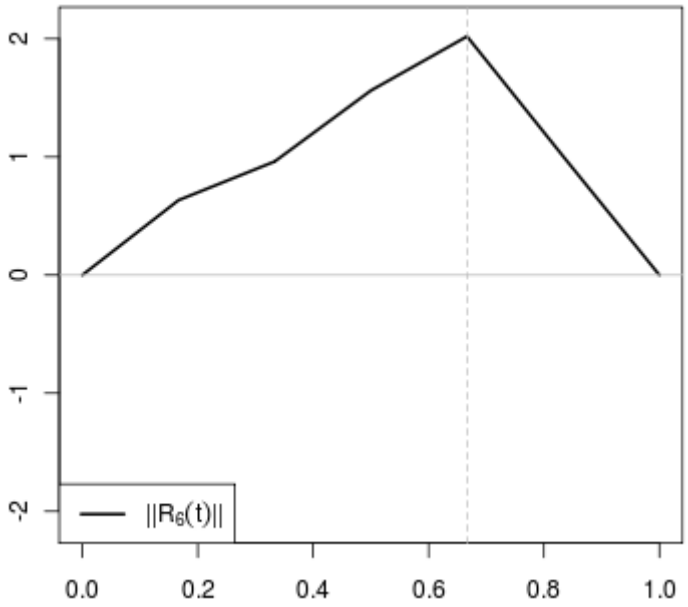


Figure 4. $\| R_6(t) \|$ with the true position of the change-point (dotted grey line)

References

1. Billingsley, P. **Convergence of Probability Measures** (2nd ed.), New York: Wiley, 1999
2. Bischoff, W. **A functional central limit theorem for regression models**, Ann. Stat., 26, 1998, pp. 1398-1410

3. Bischoff, W. **The structure of residual partial sums limit processes of linear regression models**, Theory Stoch. Process., 8, 24, N1-2, 2002, pp. 23-28
4. Bischoff, W. **Residual partial sums techniques to find change-points in linear regression**, Computer Modelling and New Technologies, 2010 (to appear)
5. Bischoff, W. and Gegg, A. **Partial sums process to check regression models with multiple correlated response with an application: Test for a change-point in profile data**, J. Multivariate Anal., 2010 (to appear)
6. Harville, D.A. **Matrix Algebra from a Statistician's Perspective**, New York: Springer, 1997
7. Iglehart, D. **Weak convergence of probability measures on product spaces with applications to sums of random vectors**, Tech. Rep., Dept. of Operations Res., Stanford Univ., 109, 1968
8. Kiefer, J. **K-sample analogues of the Kolmogorov-Smirnov and Cramer-v. Mises tests**, Ann. Math. Stat., 30, 2, 1959, pp. 420-447
9. MacNeill, I.B. **Properties of sequences of partial sums of polynomial regression residuals with applications to tests for change of regression at unknown times**, Ann. Statist., 6, 1978, pp. 422-433
10. MacNeill, I.B. **Limit processes for sequences of partial sums of regression residuals**, Ann. Prob., 6, 1978, pp. 695-698
11. Mahmoud, M., Parker,P., Woodall, W. and Hawkins, D. **A change point method for linear profile data**, Quality and Reliability Engineering International, 23, 2, 2007, pp. 247-268