Probability in Action

edited by

Bartosz Przysucha

Politechnika Lubelska

Lublin 2017

Probability in Action

Volume 3

Monografie – Politechnika Lubelska

Politechnika Lubelska
Wydział Zarządzania
ul. Nadbystrzycka 38
20-618 Lublin

Probability in Action

Volume 3

edited by
Bartosz Przysucha

Politechnika Lubelska
Lublin 2017

Reviewer:
prof. dr hab. Jurij Kozicki, Maria Curie Skłodowska University

Scientific editor: Bartosz Przysucha

Language editor: Małgorzata Maśkiewicz

Typesetting: Przemysław Kowalik

Publication approved by the Rector of Lublin University of Technology

ISBN: 978-83-7947-299-4

Publisher: Lublin University of Technology
ul. Nadbystrzycka 38D, 20-618 Lublin, Poland
Realization: Lublin University of Technology Library
ul. Nadbystrzycka 36A, 20-618 Lublin, Poland
tel. (81) 538-46-59, email: wydawca@pollub.pl
www.biblioteka.pollub.pl
Printed by : TOP Agnieszka Łuczak
www.agencjatop.pl

The digital version is available at the Digital Library of Lublin University of Technology: www.bc.pollub.pl
Impression: 100 copies

Content

Preface

We present to the readers the third volume of the book *Probability in Action*. The main purpose of the book discussed in the previous two parts has not changed. The book presents research carried out by the scientists of the Lublin University of Technology: Przemysław Kowalik, Tomasz Warowny, and Ruslan Motoryn (Department of Quantitative Methods in Management, Faculty of Management), Ernest Nieznaj, Małgorzata Murat and Iwona Malinowska (Department of Mathematics, Faculty of Electrical Engineering and Computer Science) and by Ewa Łazuka (Department of Applied Mathematics, Faculty of Fundamentals of Technology). This volume also includes the papers contributed by our collaborators from other universities – Tadeusz Banek (Faculty of Economical and Technical Sciences, Pope John Paul II State School of Higher Education in Biała Podlaska, the former head of the Department of Quantitative Methods in Management of the Lublin University of Technology and the creator of the Probability in Action series), Bartłomiej Stępień (Department of Mechanics and Vibroacoustics, AGH University of Science and Technology in Kraków), Andrii Kaminskyi and Konstantyn Pysanets (Faculty of Economics, Taras Shevchenko National University of Kyiv).

Research works discussed in the book involve probability theory, and statistics and its applications. These studies were conducted in the area of pure and applied mathematics in game theory, financial analysis, acoustics, and economics.

The book contains a presentation of a wide range of research on applied mathematics, organized in 10 thematically separate articles. The exception are two articles: *Disinformation – advanced weapon in political and military games*, which describe the same problem on two levels: a zero-sum matrix game and a non-matrix game – a stochastic view.

Tadeusz Banek[1]

Disinformation – advanced weapon in political and military games: basic ideas – zero sum matrix games

Keywords: disinformation, game theory, strategy, politics.

Abstract

A disinformation action is aimed at setting up a trap. It consists of creating a so-called "false game", which is perceived by the opponent as the original game, i.e. the one that describes the actual nature of the conflict. Even the best strategy created for the false game will not work in the original game, and so, the manipulated player, falsely assuming that he plays the original game, applies a strategy which is ineffective with the real threats, thus falling prey to the manipulator. The approach proposed here may be a powerful tool in the political conflict and the special operations. In the papers we give analytical tools for studying games with disinformation, and a methodology necessary for initiating, using and conducting them. We present a numerical example which illustrates the key idea of conducting disinformation in such games. Finally, we propose an algorithm which, step-by-step, describes our approach for integrating disinformation into the classical scheme of game theory. To make things easier it is done for matrix games and in particular for zero-sum matrix games.

1 Introduction

When writing about contemporary political games (with or without disinformation), one must start from a very general observation: it is hardly ever known what is at stake in a specific political game. Observers (and occasionally some players) are being informed about the actions being conducted, probably about implementations of the adopted strategies, but only those who are intended to be communicated, or those which relevant services managed to establish. Even though the general goal of the opponent is largely known, the stake in the current game usually is not. This is the basic difficulty in analysing such games and an irremovable barrier to implement the results of mathematical game theory ([1,2,4]). In the latter, it is assumed that the goals of the game are either known

[1] Faculty of Economical and Technical Sciences, Pope John Paul II State School of Higher Education in Biała Podlaska

to both players or there may be assigned known probabilities (stochastic games [3,4]). However, in stochastic games the entire analysis, and thus also the solutions obtained, depend on the probabilities assumed at the beginning; as one is not able to verify these probabilities, no serious use of these games is possible in the political and military domain.

This paper shows how to surmount that difficulty and take advantage of the potential of mathematical game theory. The paper is addressed to practitioners who want to apply disinformation methodically, particularly in politics in order to lead an opponent into a trap of the so-called false game (FG). The distorted image of the conflict is aimed at triggering a reaction of the opponent which will be detrimental for him in the real conflict.

1.1 Relations with traditional game theory

It should be emphasized that the proposed methodology goes beyond the framework of traditional game theory ([1,2]). That framework is shaped by the fundamental paradigm that the opponent is equally as cunning and intellectually able as we are. This can be explained as follows. Being 'equal' means being 'at most' and 'at least' simultaneously. Let us begin on 'at most' first. This paradigm is included implicitly in the fundamentals of game theory (see [1,2]), at two points: (1) players act 'rationally', (2) the game has some prescribed rules, (which are given and known to players, or not). As rationality can be defined differently, depending upon players' individual preferences and abilities, the point (1) means 'players act rationally according to his own rationality'. To create the conflict model, each player has not only to define his own rationality, but to estimate 'rationality' of the opponent as well. Together with the game rules taken from (2) all players' intellectual abilities are fixed in the model created by an individual player. Certainly, the model cannot include elements outside of creator's imaginations, but, on the other hand, it is created under the pressure of risk of opponent's underestimation. This proves the 'at most' part with a strong indication toward 'as close to equal as possible'. To see the 'at least' part look how the game is solved according to the game theory standards. The min-max strategy is the best response against opponent's best strategy, where the best strategy (or response) is defined via the maximization (or minimization) operation, i.e., it has an absolute meaning. Hence a min-max procedure forces a player to find the best response against an opponent whose skills are at least as high as ours. This proves the 'at least' part.

In order to apply game theory methodology in the practice of political and military conflicts we have to check all candidates' intellectual abilities first and select only those who passed successfully the 'equity' exam! What if we met an idiot or, much worse, had the misfortune to confront with Alcibiades, Talleyrand or Julius Cesar? How must we identify all strategies at the disposal of these opponents? This kind of player's limitations we shall call "intellectual constraints".

It is a huge paradox that a unique rigorous mathematical theory pretending to be useful in the practice of political and military conflicts is restricted by such unrealistic assumptions. Opposed to that, our methodology assumes that the opponent can be cheated. Such a successful disinformation action may give him a false image of the conflict. The history of political and military games offers enough examples to back such an assumption, whereas game theory itself offers no scheme for breaking free from this restrictive and unrealistic traditional paradigm. This paper is a step in that direction.

Inspired by two classic books, "The Art of War" ([6]), by Sun Tzu, and "New Lies for Old" ([5]) by Anatoliy Golitsyn, this paper can, in some restricted sense, be considered as a mathematical extract from these prototypes.

The basic part of the work is devoted to the use of disinformation in two-person zero-sum matrix games. We analyse the situation of a player who knows that he is being misled and that his image of the conflict is a false one, but he is also aware that he has no other. We pose the problem of getting the real picture of the conflict from the knowledge of the false one, and we show that the solution involves reversing the disinformation action.

2 Disinformation in games

We begin with some general remarks. Disinformation has always accompanied political and military conflicts. However, the scale and range of techniques used nowadays in such actions is a new phenomenon and it should be assumed that this increasing trend will continue and even accelerate, propelled by new developments in information technology.

Games may be played against:

1. an external opponent (a foreign country);
2. an internal opponent (political opposition);
3. a total opponent (a foreign country and its agents located in the home country).

Disinformation may concern:

a) intentions (e.g. the place of landing the enemy troops);
b) goals (e.g. a game aimed at breaking up an enemy alliance by creating and disclosing a conflict of interests between the allies).

This paper is limited to a detailed discussion of item 1b.Item 1a is briefly alluded in Digression 1.

3 Disinformation in zero-sum games

Disinformation is a part of a political or military action aimed at misleading an opponent and giving him a false image of the conflict. The real goals of the disinformant (D) are hidden from his opponent (O), and replaced by other ones, seemingly as important. In this way, the original game (OG) which D wants to win turns into two games: OG and the false game (FG). Unaware of that, O de-

velops the best (in his opinion) strategy (FGS) for the game FG, while that strategy is in fact implemented in the original game OG. As OG and FG differ one from another in the way which is known only to D, the strategy FGS will not be successful in OG. Moreover, that strategy is transparent to D, because he has created FG and is able to analyse it. To make matters worse, it is D who can develop a counterstrategy against FGS, the one that will work best in the original game too!

Examples

1. To illustrate the above ideas, let us consider the extremely simplified model of a zero sum game with the payoff matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 3 \\ 7 & 2 & 6 \\ 1 & 1 & 3 \end{bmatrix}.$$

D chooses one of the rows of $\boldsymbol{A}$ and O one of the columns. The payoff for D (and the loss for O) is the element in the chosen row and column. In this game, optimal strategies for the players are as follows: the second row for D, the second column for O. The result of the game is 2, which means that O pays D two units. Choosing the second row is the optimal strategy for D as it guarantees a win of 2 (if O plays the second column), or wins of 7 or 6 if O plays the first or third column, respectively. Choosing the second column is the best for O as it guarantees a loss no higher than 2 (if D plays the second row), or even smaller losses of 0 or 1 if D plays the first or third row, respectively. Guaranteed wins/losses are equal and result is the best possible outcome for both players.

Let us see what happens if we replace one sensitive element $a_{21} = 7$ in the second row and first column of $\boldsymbol{A}$, creating a false game with matrix

$$\boldsymbol{B} = \begin{bmatrix} 1 & 0 & 3 \\ 1 & 2 & 6 \\ 1 & 1 & 3 \end{bmatrix}.$$

In that game, the best strategy for O is to choose the first column as this guarantees a loss no higher than 1, while the other columns carry a risk of losing 2 or 6. However, applying that strategy (first column) is disastrous for O when the game is played on the matrix $\boldsymbol{A}$ of the original game. Now, playing the second row, D wins as many as 7 units, which is the best possible outcome in game $\boldsymbol{A}$, achieved as a result of game analysis and a successful disinformation action.

2. Suppose that in a game with matrix

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 3 \\ 7 & 2 & 6 \\ 0 & 1 & 3 \end{bmatrix}$$

D managed to “convince” O that D is bound not to play the second row, i.e. that the matrix

$$\begin{bmatrix} 0 & 1 & 3 \\ 0 & 1 & 3 \end{bmatrix}$$

is the original matrix of the game. Then the best strategy for O would be to choose the first column, giving him a loss of only zero. However, when executed on the matrix $\boldsymbol{A}$ that strategy gives D a win of 7 (and thus a loss of 7 and not 0 for O).

3. Now, let us consider a simple zero sum game with

$$\boldsymbol{G} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

As it is easy to compute, the strategies $(p, q) = \mathrm{col}(1/2, 1/2)$ are optimal and the value of the game $V(\boldsymbol{G})$ is 0. Now, let us assume that in the result of disinformation the player Q is convinced that the true playoff is

$$\boldsymbol{R} = \begin{bmatrix} 1 & -1 \\ x & 1 \end{bmatrix},$$

where x is a given number. Simple computations show that now the min-max strategies $(p(x), q(x))$ are

$$(p(x), q(x)) = \begin{bmatrix} 2/(3-x) \\ (1-x)/(3-x) \end{bmatrix}$$

for $x < 1$ and the value of the game in this case is $V(\boldsymbol{R}) = (1+x)/(3-x)$ and

$$p(x) = q(x) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

if $x > 1$ with the value of the game is $V(\boldsymbol{R}) = 1$.

4. Assume now, that P successfully applied disinformation described in the example above to the original game $\boldsymbol{G}$, i.e., P knows that Q plays on the matrix $\boldsymbol{R}$. Being sure that Q will play strategy q in the primary game $\boldsymbol{G}$, he may apply the best counterstrategy $p(\boldsymbol{GR})$against $q(x)$. This leads to the simple maximization problem

$$\max \left\{ \frac{1+x}{3-x}(2p-1); 0 \le p \le 1 \right\}$$

with the obvious solution for $x < 1$

$$p(\boldsymbol{GR}) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \text{ when } x \text{ belongs to } (-1,1),$$

$$p(\boldsymbol{GR}) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \text{ when } x \text{ belongs to } (-\infty, -1)$$

and the corresponding winning for P is

$$V(\boldsymbol{GR}) = \left|\frac{1+x}{3-x}\right|.$$

Moreover, it should be noted that not only the values $V(\boldsymbol{R})$and $V(\boldsymbol{GR})$ are different, but the strategy $p(\boldsymbol{GR})$ is deterministic as opposed to mixed strategies $p(\boldsymbol{G})$ or $p(\boldsymbol{R})$. This shows that the proper selection of x can be very profitable for P giving him the high wins with no risk of errors occurring when the initially intended disinformation playoff x is wrongly estimated by Q, as could be in the case when the game value has the form similar to $V(\boldsymbol{R})$.

Generally, the following claim is not surprising.

Claim. Disinformation is always profitable when applied to zero-sum games.

Proof. Indeed, let $(p(G), q(\boldsymbol{G}))$ be a pair of optimal min-max strategies in the original game G, i.e.,

$$p^T\boldsymbol{G}q(\boldsymbol{G}) \leq p^T(\boldsymbol{G})\ \boldsymbol{G}q(\boldsymbol{G}) \leq p^T(\boldsymbol{G})\boldsymbol{G}q$$

for any (p, q). Playing optimally $q(\boldsymbol{R})$ in the game $\boldsymbol{R}$, player Q may secure $p^T(\boldsymbol{R})\ \boldsymbol{R}q(\boldsymbol{R})$, where

$$p^T\boldsymbol{R}q(\boldsymbol{R}) \leq p^T(\boldsymbol{R})\boldsymbol{R}q(\boldsymbol{R}) \leq p^T(\boldsymbol{R})\boldsymbol{R}q.$$

However, since $q(\boldsymbol{R})$ generally differs from $q(\boldsymbol{G})$, we have

$$p^T(\boldsymbol{G})\ \boldsymbol{G}q(\boldsymbol{G}) \leq p^T(\boldsymbol{G})\boldsymbol{G}q(\boldsymbol{R}).$$

Moreover, if P knows $(\boldsymbol{R})$, then he can do better by applying $p(\boldsymbol{GR})$ defined in the condition

$$p^T(\boldsymbol{GR})\ \boldsymbol{G}q(\boldsymbol{R}) \leq \underbrace{\max}_{p}\ p^T\boldsymbol{G}q(\boldsymbol{R})$$

Obviously,

$$\underbrace{\max}_{p}\ p^T\boldsymbol{G}q(\boldsymbol{R}) \geq p^T(\boldsymbol{G})\boldsymbol{G}q(\boldsymbol{R})$$

hence

$$p^T(\boldsymbol{G})\boldsymbol{G}q(\boldsymbol{G}) \leq p^T(\boldsymbol{G})\ \boldsymbol{G}q(\boldsymbol{R}) \leq p^T(\boldsymbol{GR})\boldsymbol{G}q(\boldsymbol{R})$$

proving the claim.

Looking in the opposite direction one may say another claim.

Claim. If disinformation was used as a rule, then it was applied to the zero-sum games.

Let us now move from those particular cases to a general discussion.

For that purpose, let us consider a game in which payoffs (wins or losses) are presented in the form of a matrix with n rows and m columns:

$$\boldsymbol{A} \stackrel{\text{def}}{=} \begin{bmatrix} a_{11} \dots & \dots a_{1j} \dots & \dots a_{1m} \\ a_{i1} \dots & \dots a_{ij} \dots & \dots a_{im} \\ a_{n1} \dots & \dots a_{nj} \dots & \dots a_{nm} \end{bmatrix}$$

The game is as follows: one player (D) chooses one of the rows of the matrix and the other player (O) chooses one of the columns. The players do not inform each other of their choices which are being made simultaneously. If D chooses row i and O chooses column j, the payoff is the element a_{ij} in row i and column j. To fix ideas, we assume that a_{ij} is the payoff paid to D by O. If a_{ij} turned out to be negative, D would pay O a win of a_{ij}. In the literature, such games are called two person zero-sum matrix games. The mathematical theory of such games is widely known ([1]), as are the procedures for analysing and solving them. Let us assume that this very game with payoff matrix $\boldsymbol{A}$ is the original game for player D. If D can suggest that the game they are playing has payoffs specified in another matrix, e.g.

$$\boldsymbol{B} \stackrel{\text{def}}{=} \begin{bmatrix} b_{11} \dots & \dots b_{1j} \dots & \dots b_{1m} \\ b_{i1} \dots & \dots b_{ij} \dots & \dots b_{im} \\ b_{n1} \dots & \dots b_{nj} \dots & \dots b_{nm} \end{bmatrix}$$

with b_{ij} different from a_{ij}, then it is a classic example of a disinformation action mentioned at the beginning. In that case, O, believing that he is playing game $\boldsymbol{B}$, develops the best (in his opinion) strategy against D, which has been referred to before as the false game strategy (FGS). In fact, that strategy is being used in the original game A. Of course, player D is familiar with that strategy as he has created game $\boldsymbol{B}$ himself, knows the payoffs in it and is able to reconstruct O's strategy, i.e. FGS. With this knowledge, D can easily calculate the best response to FGS, while of course taking into account the fact that the real game has payoffs specified in the matrix $\boldsymbol{A}$. Basing on the above pattern, it is easy to formulate the Optimum Disinformation Problem.

It consists in distorting the matrix $\boldsymbol{A}$ into a matrix $\boldsymbol{B}$ in such a way that the strategy FGS calculated by O based on the matrix B is as advantageous as possible for D in the original game $\boldsymbol{A}$. Formally, the distorting Δ is an operation

$$M(n,m) \ni \boldsymbol{A} \rightarrow \Delta(\boldsymbol{A}) = \boldsymbol{B} \in M(n,m)$$

where $M(n,m)$ is a set of all n – row and m - column real matrices.

4 Mathematical model of the problem games

We assume that $\boldsymbol{A}$, $\boldsymbol{B}$ are matrices of the same dimensions. The standard n-dimensional simplex is defined by

$$S_n \stackrel{\text{def}}{=} \left\{ w \in \mathbb{R}^n;\ w_i \geq 0, \sum_{i=1}^{n} w_i = 1 \right\}.$$

The best strategy $q(\boldsymbol{B})$ of player O in the game on the matrix $\boldsymbol{B}$ is defined by the condition

$$\langle p, \boldsymbol{B}q(\boldsymbol{B})\rangle \leq \underbrace{\max}_{p} \langle p, \boldsymbol{B}q(\boldsymbol{B})\rangle \leq \underbrace{\max}_{p} \underbrace{\min}_{q} \langle p, \boldsymbol{B}q\rangle$$

for all $p \in S_n,\ q \in S_m$, where $\langle \cdot,\cdot \rangle$ denotes the scalar product.

As this strategy depends on the matrix $\boldsymbol{B}$, we have denoted it by $q(\boldsymbol{B})$. The best response of player D is the solution of the linear programming problem

$$\max\{\langle p, \boldsymbol{A}q(\boldsymbol{B})\rangle;\ p \in S_n\},$$

which depends on $q(\boldsymbol{B})$, and hence on $\boldsymbol{B}$. Let us therefore denote it by $p(q(\boldsymbol{B}))$. Let $\#\boldsymbol{A}$ signify the number of non-zero elements in the matrix $\boldsymbol{A}$. We introduce a matrix index by setting

$$W(\boldsymbol{A}) \stackrel{\text{def}}{=} (\#\boldsymbol{A})! \sqrt{\sum_{i,j=1}^{n,m} a_{ij}^2}$$

This index strongly distinguishes matrices depending on the number of non-zero elements. Disinformation actions which change payoffs in the game matrix from $\boldsymbol{A}$ to $\boldsymbol{B}$ may be costly, and the costs increase with the number of elements changed and the scale of the changes. The above index is aimed at measuring the scope and scale of the changes. It is of importance when constructing the false game matrix $\boldsymbol{B}$.

Remark: The assumption that the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same dimensions, i.e.

$$M(n,m) \ni \boldsymbol{A} \ \rightarrow \Delta(\boldsymbol{A}) = \boldsymbol{B} \ \in \ M(n,m)$$

is of technical nature, but it does not restrict the generality of our considerations. If $\boldsymbol{B}$ has different dimensions, e.g. k rows and r columns, it is enough to expand both $\boldsymbol{A}$ and $\boldsymbol{B}$ by adding zero entries to $\max(n,k)$ rows and $\max(m,r)$ columns. In the general variant, the matrix dimensions may be different. A larger matrix $\boldsymbol{B}$ corresponds to conflict escalation, a smaller to de-escalation.

a. The Optimum Disinformation Problem (ODP)

For a given matrix $\boldsymbol{A}$, and $\varepsilon > 0$, we search for

$$\max\{\langle p(q(\boldsymbol{B})), \boldsymbol{A}q(\boldsymbol{B})\rangle; W(\boldsymbol{B}-\boldsymbol{A}) \leq \varepsilon\}$$

over all matrices $\boldsymbol{B}$ of the same dimension as $\boldsymbol{A}$.

Explicitly, ODP may be presented as follows.

Find

$$\underbrace{\max}_{B}\left\{\underbrace{\max}_{p}\left\{\langle p, \boldsymbol{A}q(\boldsymbol{B})\rangle ; \langle w, \boldsymbol{B}q(\boldsymbol{B})\rangle \leq \underbrace{\max}_{w}\langle w, \boldsymbol{B}q\rangle\right\};\ W(\boldsymbol{B}-\boldsymbol{A}) \leq \varepsilon\right\}$$

Remark. Sometimes it would be desirable to consider an weaker version of this problem when the goal is to force the opponent to play some particular strategy q_s. The problem takes the form: Find

$$\underbrace{\min}_{B}\{\|\boldsymbol{A}[q_s - q(\boldsymbol{B})]\| \ ;\ W(\boldsymbol{B}-\boldsymbol{A}) \leq \varepsilon\}$$

over all matrices $\boldsymbol{B}$ of the same dimension as $\boldsymbol{A}$, if a difference $q_s - q(\boldsymbol{B})$ should be measured in the image of $\boldsymbol{A}$, or simply

$$\underbrace{\min}_{B}\{\|q_s - q(B)\| \ ;\ W(B-A) \leq \varepsilon\},$$

if not.

Disinformation directions.

After finding the solution $\boldsymbol{B}(\boldsymbol{A}, \varepsilon)$ of ODP, one can examine which elements of the matrix $\boldsymbol{A}$ should be modified, and by how much.

It is the position of the elements subject to change in the matrix $\boldsymbol{A}$ that indicates the directions of optimum disinformation, and the size of those elements corresponds to the intensity of disinformation action aimed at distorting the picture of the conflict.

5 The algorithm

a. Creating the matrix $\boldsymbol{A}$ of the original game

This is the task for political and/or military analysts in close cooperation with mathematicians. It consists in:

1. Identifying all major imaginable strategies of both players. This stage requires imagination and a capacity for creating operational combinations and forming them into strategies and counterstrategies.
2. Assessing the outcome of confronting each pair strategy vs. counterstrategy.
3. Assigning a numerical value (payoff) to each outcome of confrontation according to a single unit of measurement for all game outcomes.
4. Entering those numerical values into appropriate places in the matrix.

This stage is of utmost importance and should be verified many times based on the entire body of knowledge gathered in the all available information resources, as it is aimed at recreating the real background, scope and goals of the conflict. The game matrix is a model of the actual conflict of interests. The task is an extremely delicate and complex one, as in real political games particular strategies are usually implemented not through one move, but through an entire

sequence of actions coordinated in time. An additional factor to be taken into account is the state of information resources during the implementation of that sequence, and the fact that they may dynamically expand. Such a situation may be conveniently described by creating a so-called dendrite (directed graph) which makes it easy to write the final outcome (payoff) of confronting a pair of strategies in the form of an element of a matrix. In the literature, this is referred to as transition from the extensive form (of the game) to the normal form ([1,2]).

b. Defining optimal strategies in the original game $\boldsymbol{A}$

Having specified the matrix $\boldsymbol{A}$, we find a pair of optimal strategies for players, and the corresponding values of the game. This may be done with the use of a spreadsheet or – in the case of large matrices – by applying more specialized tools, e.g. MATLAB.

c. Creating the matrix B of the false game

This consists in transforming the matrix $\boldsymbol{A}$ into a matrix $\boldsymbol{B}$ by changing the values of some elements (payoffs). It can only be done after analysing in which subject areas, and to what extent, influencing the opponent's information centres is possible; this leads to selecting those elements of $\boldsymbol{A}$ that can be changed, and determines the scope of the changes. Only then can one mathematically analyse the impact of the selected elements, leading to the modification of elements of $\boldsymbol{A}$. Such analysis will be the subject of another paper. Its summary is presented below in the form of a procedure:

1. Arrange the elements of the matrix in non-increasing order.
2. Consider the largest element a_{ij}.
3. Analyse how many elements of column j must be modified and by how much to make it the dominating strategy for O in the modified game.
4. Check if the change is possible in the light of available impact means. If yes, complete the procedure. If not, go to 5.
5. Conduct the same reasoning for the second largest element, etc. If the outcome is satisfactory, finish; if not, go to 6.
6. Analyse the combined impact of two largest elements of $\boldsymbol{A}$, modifying item 3 in order to answer the question: how and by how much one should change the elements in columns in which those largest elements occur in order to make the strategy of choosing those columns the dominating strategy for player O. If the outcome is satisfactory, we finish the procedure; if not, go to 7.
7. Go back to 6, taking into consideration the next two, three elements etc.

Remark. Implementing steps 1 to 7 leads to solving the Optimum Disinformation Problem (ODP).

d. Solution of the false game $\boldsymbol{B}$.

Similarly to the procedure for the original game, we now calculate the optimal strategy "q" for player O in the false game $\boldsymbol{B}$.

e. Return to the original game $\boldsymbol{A}$

Having established the strategy "q" of the manipulated opponent (SOB, strategy of player O in the false game $\boldsymbol{B}$), we apply it to the original game $\boldsymbol{A}$. We then solve the linear programming problem

$$\{ \max\langle p, \boldsymbol{A}q \rangle \}$$

over all probability distributions on strategies p of D, the result of which is the best counteraction against the manipulated O.

f. Assessment of disinformation effectiveness

This consists in comparing the result from item (e) with the result from item (b). If the difference between the results, relative to the costs of the disinformation action, is satisfactory, we finish the calculation procedure and move on to the implementation stage.

If the difference is not satisfactory, we return loop-like to (c) and repeat all steps of the Algorithm until the result is satisfactory.

g. Implementation

Previous stages lead to finding the matrix $\boldsymbol{A}$ of the original game and the matrix $\boldsymbol{B}$ of the false game, which differs from $\boldsymbol{A}$ in one or more elements. These elements determine the payoffs when the players confront certain strategies. A disinformation action should now be conducted so as to make O think that the payoffs will be as calculated in the matrix $\boldsymbol{B}$. After completing the disinformation action, the actual game may be started, i.e. we apply the strategy calculated in (e).

Bibliography

[1] von Neumann J., Morgenstern O., *Theory of Games and Economic Behavior*, Princeton 1947.
[2] Luce R. D., Raiffa H., *Games and Decisions*, John Wiley & Sons, Inc., New York 1958.
[3] Harsanyi J., *Games With Incomplete Information Played by Bayesian Players*, *Part I*, "Management Science" 14 (3), 1967, 159–183; *Part II*: ibid., 14 (5), 320–334, *Part III*: ibid., 14 (7), 486–502.
[4] Fudenberg D., Tirole J., *Game Theory*, MIT Press 1991.
[5] Golitsyn A., *New Lies for Old*, Dodd, Mead & Company, New York 1984.
[6] Sun Tzu, *The Art of War*, Department of Oriental Printed Books and Manuscripts, British Museum, The Puppet Press 1910.
[7] Wets R. J.-B., *On the Relation between Stochastic and Deterministic Optimization,* in*: Control Theory, Numerical Methods and Computer Systems Modelling*, eds. Bensoussan A. and Lions J.L., "Lectures Notes in Economics and Mathematical Systems" 107, Springer-Verlag, Berlin 1975, 350–361.
[8] Davis M.H.A., Dempster M.A.H., Elliott R.J., *On the Value of Information in Controlled Diffusions Processes*, Liber Amicorum for M. Zakai, 125–138.

Ernest Nieznaj[1]

A few remarks about point processes on the line

Keywords: quasi Poisson point process, Poisson point process, Shepp example, Gumbel distribution, Burr distribution, generalized Poisson distribution.

Abstract

The purpose of this article is to provide the reader with several useful remarks about point processes. We deal, among other things, with a quasi-Poisson process, order statistics of point processes, the Gumbel, Burr and generalized Poisson distribution.

1 Introduction

The theory of point processes has many applications and models based on this theory can be applied in a broad range of disciplines. For example, in a book *Case Studies in Spatial Point Pattern Modelling* by A. Baddaley et al. (that is [1]), there is a bunch of articles concerning e.g.

- forestry and plant ecology (for modelling positions of trees and plants)
- epidemiology (home locations of infected patients)
- zoology (burrows or nests of animals)
- geography (positions of human settlements, towns or cities)
- seismology (epicentres of earthquakes)
- astronomy (locations of stars or galaxies).

There are also several chapters in [13] devoted to Poisson point process methods applied to tomographic imaging in medicine. Therefore the theory of point processes is not only for its own sake, but may be useful in applied sciences. This article is about several aspects of the theory. We finish this section with two definitions. A *point process* on a measurable space $(E, \mathcal{E})$ is a random variable, usually denoted by Π, whose realizations are subsets of E containing a finite or countable number of points. Thus for a fixed set $A \in \mathcal{E}$, the random variable $S(A) \coloneqq |\Pi \cap A|$ counts the points that „fall" in A. Assume that μ is a σ-finite measure on E A *Poisson process* on $(E, \mathcal{E})$ is a point process Π such that the following conditions hold

[1] Department of Mathematics, Faculty of Electrical Engineering and Computer Science, Lublin University of Technology, e-mail: e.nieznaj@pollub.pl

(i) for any $n > 2$ and any pairwise disjoint subsets $A_1, \dots, A_n \in \mathcal{E}$, the random variables $S(A_1), \dots, S(A_n)$ are independent;
(ii) if $A \in \mathcal{E}$ then $S(A)$ has the Poisson distribution with parameter $\mu(A)$.

A measure μ is often called the mean measure of Π, since $\mathbb{E}[S(A)] = \mu(A)$. For most applications E is $\mathbb{R}$, $\mathbb{R}^2$ or $\mathbb{R}^3$ with a standard Borel σ-algebra.

2 Shepp example of a quasi-Poisson process on ⟨0,1⟩

This example, given by a mathematician Lawrence Shepp (1936-2013), was first published in [6] by Goldman in 1967. It was later cited in several books, e.g. [12, 13]. We study in detail the construction of this process as opposed to the mentioned books. We use cdf as an acronym for a cumulative distribution function, i.e. $F(x) = \mathbb{P}(X \leq x)$, as well as pdf for probability density function, $f(x) = F'(x)$. If $X_1, \dots, X_n$ are independent and identically distributed, we write iid.

We construct a process S on the interval $\langle 0,1 \rangle$ in the following way. In the first step take $\lambda > 0$ and then choose a natural number n with probability $e^{-\lambda}\lambda^n/n!$, where $n = 0,1,2, \dots$. If $n = 0$ then nothing happens. If $n \geq 1$ and $n \neq 3$ let

$$F_n(x_1, \dots, x_n) = x_1 x_1 \dots x_n \tag{2.1}$$

be the cdf of the points $t_1, \dots, t_n$ of S where $x_1, \dots, x_n \in \langle 0,1 \rangle$. That means that we choose n points independently each with uniform cdf $F_n(x_i) = x_i$, $x_i \in \langle 0,1 \rangle$, $i = 1, \dots, n$. It is clear how cdf given by (2.1) looks on the outside of $\langle 0,1 \rangle^n$. If $n = 3$ we choose three points t_1, t_2, t_3 from $\langle 0,1 \rangle$ in such a way that their joint cdf is given by

$$\begin{aligned} F_3(x_1, x_2, x_3) = x_1 x_2 x_3 + \varepsilon(x_1 - x_2)^2(x_1 - x_3)^2(x_2 - x_3)^2 \\ \cdot x_1 x_2 x_3 (1 - x_1)(1 - x_2)(1 - x_3) \end{aligned} \tag{2.2}$$

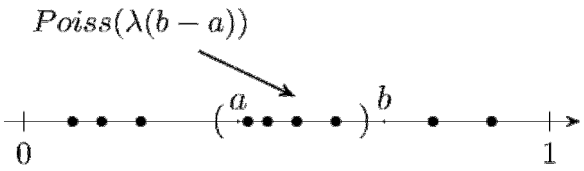


Figure 1. The counting random variable $S(A)$ has a Poisson distribution for any interval $A = (a, b)$.

Source: own elaboration.

We show in Remark 2.1 that it is in fact a cdf for certain $\varepsilon > 0$. Although one-dimensional marginals of F_3 are uniformly distributed on $\langle 0,1 \rangle$, it is not

a product of those marginals. Let $A = (a, b)$ be a subset of $\langle 0,1\rangle$. First we will prove that for $m = 0,1,2, \ldots$ we have

$$\mathbb{P}(S(A) = m) = e^{-\lambda(b-a)} \frac{(\lambda(b-a))^m}{m!}. \tag{2.3}$$

Fix m and denote (as in [5]) by $G_n(a, b, m)$ the probability that exactly m of the points $t_1, \ldots, t_n$ are in (a, b), $n \geq m$. Then

$$G_n(a,b,m) = \binom{n}{m} \mathbb{P}_n(\{t_1, \ldots t_m \in (a,b)\} \cap \{t_{m+1}, \ldots t_n \neq (a,b)\})$$

$$= \binom{n}{m} \mathbb{E}_n \left[\prod_{i=1}^{m} (X_b(t_i) - X_a(t_i)) \prod_{i=m+1}^{n} (X_a(t_i) + X_1(t_i) - X_b(t_i)) \right]$$

where

$$X_a(t) = \begin{cases} 1, & t \leq a \\ 0, & t > a. \end{cases}$$

The above equality follows from the fact that $\mathbb{P}(t \in (a,b)) = \mathbb{E}(X_b(t) - X_a(t))$ and $\mathbb{P}(t \notin (a,b)) = 1 - (b-a)$. The first product in the above expectation may be written in the form $\sum (-1)^l X_{a_1}(t_1) \ldots X_{a_m}(t_m)$ where a_i equals a or b for $i = 1, \ldots, m$ and l is the number of a's. The second product we may write as $\sum (-1)^l X_{a_{m+1}} X(t_{m+1}) \ldots X_{a_n}(t_n)$ where a_i equals a, b or 1 for $i = m+1, \ldots, n$ and l is the number of b's. The first sum is over all multi-indexes $(a_1, \ldots, a_m)$ and the second over all possible $(a_{m+1}, \ldots, a_n)$. So each term in the above expectation has the form

$$\sum \sum X_{a_1}(t_1) \ldots X_{a_m}(t_m) X_{a_{m+1}} X(t_{m+1}) \ldots X_{a_n}(t_n).$$

Next, we have

$$\mathbb{E}_n\left[X_{a_1}(t_1) \ldots X_{a_n}(t_n)\right] = F_n(a_1, \ldots, a_n) = a_1 \ldots a_n,$$

for any $n \geq 1$. If $n \neq 3$ then it follows directly from (2.1). It is also true for $n = 3$, because the second term in F_3 equals zero if at least two of its arguments are the same or at least one equals 1, i.e. $F_3(a,a,b) = aab$, $F_3(1,a,b) = ab$ and so on. Hence considered expectation takes the form

$$\sum \sum a_1 \ldots a_m a_{m+1} \ldots a_n = (\sum a_1 \ldots a_m)(\sum a_{m+1} \ldots a_n).$$

The first sum on the right hand side is $\sum_{i=0}^{m} \binom{m}{i} (-a)^i b^{m-i}$ which equals to $(b-a)^m$. The second one we write as

$$\sum_{i=0}^{n-m}\sum_{j=0}^{n-m-i}\binom{n-m}{i}\binom{n-m-i}{j}a^i(-b)^{n-m-i-j}$$

which equals to $(a+(1-b))^{n-m}$. Therefore

$$G_n(a,b,m)=\binom{n}{m}(b-a)^m(1-(b-a))^{n-m},\quad n\geq m.$$

From this one gets $\Delta=b-a$,

$$\begin{aligned}\mathbb{P}(S(A)=m)&=\sum_{n=m}^{+\infty}e^{-\lambda}\frac{\lambda^n}{n!}\binom{n}{m}\Delta^m(1-\Delta)^{n-m}\\&=e^{-\lambda}\frac{(\lambda\Delta)^m}{m!}\sum_{n=m}^{+\infty}\frac{(\lambda(1-\Delta))^{n-m}}{(n-m)!}=e^{-\lambda\Delta}\frac{(\lambda\Delta)^m}{m!},\end{aligned}$$

which means that this is the end of the proof of (2.3). Now we investigate lack of independence property of S. This is obviously caused by (2.2). Take for example $A_1=\left(\frac{1}{8},\frac{2}{8}\right)$, $A_2=\left(\frac{3}{8},\frac{4}{8}\right)$ and $A_3=\left(\frac{5}{8},\frac{6}{8}\right)$, see also Figure 2. We will show that $S(A_1)$, $S(A_2)$ and $S(A_3)$ are not independent by proving

$$\mathbb{P}\left(\bigcap_{i=1}^{3}\{S(A_i)=1\}\right)\neq\prod_{i=1}^{3}\mathbb{P}(S(A_i)=1)\tag{2.4}$$

First, observe that the left hand side of (2.4) equals

$$(S(A_1)=1,S(A_2)=1,S(A)=3),$$

where $A=A_1\cup A_2\cup A_3$. From this one gets

$$\begin{aligned}&\mathbb{P}\left(\bigcap_{i=1}^{3}\{S(A_i)=1\}\right)=\mathbb{P}(S(A_1)=1,S(A_2)=1,S(A)=3)=\\&=\mathbb{P}(S(A_1)=1,S(A_2)=1|S(A)=3)\cdot\mathbb{P}(S(A)=3)\end{aligned}$$

From (2.3) we have $\mathbb{P}(S(A_i)=1)=\frac{1}{8}\lambda e^{-\lambda/8}$ for $i=1,2,3$, and $\mathbb{P}(S(A)=3)=\left(\frac{3}{8}\lambda\right)^3e^{-\frac{3}{8}\lambda}/3!$, From (2.2) we have

$$\mathbb{P}(S(A_1)=1,S(A_2)=1|S(A)=3)=\int_{\frac{1}{8}}^{\frac{2}{8}}\int_{\frac{3}{8}}^{\frac{4}{8}}\int_{\frac{5}{8}}^{\frac{6}{8}}F_3(dx_1,dx_2,dx_3)$$

which equals $\left(\frac{1}{8}\right)^3+4428\varepsilon\left(\frac{1}{8}\right)^{11}$. Thus for any $\varepsilon>0$ the following holds

$$\mathbb{P}\left(\bigcap_{i=1}^{3}\{S(A_i)=1\}\right)=\left(\frac{1}{8}\right)^3+4428\varepsilon\left(\frac{1}{8}\right)^{11}\cdot\frac{1}{3!}\left(\frac{3}{8}\lambda\right)^3 e^{-\frac{3}{8\lambda}}$$
$$\neq\left(\frac{3}{8}\lambda\right)^3 e^{-\frac{3}{8\lambda}}=\prod_{i=1}^{3}\mathbb{P}(S(A_i)=1),$$

which proves (2.4).

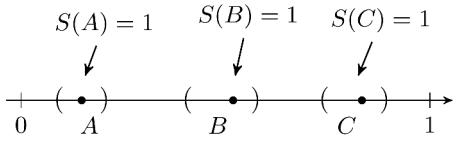


Figure 2. The situation in which $S(A)$, $S(B)$ and $S(C)$ are not independent. This is a consequence of (2.2).

Source: own elaboration.

Remark 2.1. We will show here that F_3 is a cumulative distribution function for certain $\varepsilon > 0$. Write $F_3(x_1,x_2,x_3) = x_1x_2x_3 + \varepsilon G_3(x_1,x_2,x_3)$, where

$$G_3(x_1,x_2,x_3) = x_1x_2x_3(x_1-x_2)^2(x_1-x_3)^2(x_2-x_3)^2\cdot (1-x_1)(1-x_2)(1-x_3).$$

Therefore the pdf of F_3 equals

$$f_3(x_1,x_2,x_3) = 1+\varepsilon g_3(x_1,x_2,x_3),\quad x_1,x_2,x_3\in\langle 0,1\rangle,$$

where $g_3(x_1,x_2,x_3) = \partial_{x_1x_2x_3}G_3(x_1,x_2,x_3)$. Since two dimensional marginals of F_3 are uniformly distributed on $\langle 0,1\rangle^2$, i.e. $F_3(x_1,x_2) = x_1x_2$, $F_3(x_1,x_3) = x_1x_3$ and $F_3(x_2,x_3) = x_2x_3$ (which implies $f_3(x_1,x_2) = 1$ and so on) we have

$$\int_0^1 g_3(x_1,x_2,x_3)dx_i = 0,\quad i=1,2,3. \tag{2.5}$$

Denote by m the minimum value of g_3 on $\langle 0,1\rangle^3$ and observe that $m<0$ by (2.5). Then $M+g_3\geq 0$ for any $M\geq|m|$. Therefore the function $(M+g_3)/A$ is a pdf defined on $\langle 0,1\rangle^3$, where $A=\int_{\langle 0,1\rangle^3}(M+g_3)dx_1x_2x_3$. But from (2.5) it follows that $A=M$, so $f_3 = 1+\frac{1}{M}g_3$ is a pdf. This means that F_3 is a cdf for any $\varepsilon\in(0,\frac{1}{M}\rangle$. ■

Example 2.2. Consider (less complicated) two dimensional cdf version of (2.2). Take a random vector (X,Y) with joint cdf given by

$$F(x,y) = xy + \underbrace{xy(x-y)^2(1-x)(1-y)}_{=G(x,y)}, \quad x,y \in \langle 0,1\rangle.$$

We write x, y instead of x_1, x_2, see Figure 3. Hence X and Y are uniformly distributed on $\langle 0,1\rangle$. The pdf of (X,Y) equals

$$f(x,y) = 1 - 4x^3 + 8x^3y - 4y^3 + 8y^3x + 3(x^2+y^2) - 8xy + \\ + 6x^2y - 18x^2y^2 + 6xy^2,$$

where $g(x,y) = f(x,y) - 1$. We have $\min\limits_{x,y\in D} g(x,y) = -1$, where $D = \langle 0,1\rangle^2$. Hence for any $M \geq 1$ the function $1 + \frac{1}{M}g(x,y)$ is a pdf and $xy + \varepsilon G(x,y)$ is a cdf for any $\varepsilon \in (0,1\rangle$. From the joint cdf of X, Y it follows are they are dependent with $cov\,(X,Y) = \frac{1}{360}$ and the correlation coefficient $\rho = \frac{1}{30}$. ■

Suppose that the joint pdf of (X,Y) can be expressed in the form

$$f(x,y) = f(x)g(y) + \varphi(x)\psi(y) - \varphi(y)\psi(x), \tag{2.6}$$

where f, g are one dimensional density functions and φ, ψ are odd and integrable functions. Assume also that $\varphi \neq \psi$. Then the pdf of X is f and pdf of Y is g, however X, Y are not independent. If $\mathbb{E}(XY)$ exists then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ due to $\int\int xy\varphi(x)\psi(y)dxdy = \int\int xy\varphi(y)\psi(x)dxdy$. In consequence $\operatorname{cov}(X,Y) = 0$ and X, Y are uncorrelated, see [5] and [11].

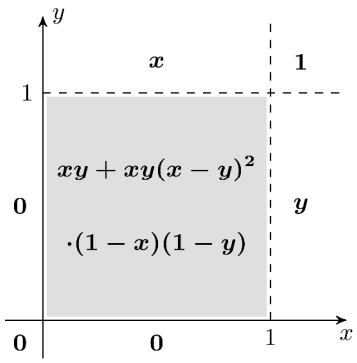


Figure 3. The cdf of a random vector (X,Y), where X and Y are dependent and uniformly distributed on $\langle 0,1\rangle$ with $cov\,(X,Y) = \frac{1}{360}$. The shaded region is the support of (X,Y).

Source: own elaboration.

Example 2.3. Consider a random vector (X,Y) with joint pdf given by

$$f(x,y) = 1 + \left(\frac{1}{2} - x\right)\sin(2\pi y) - \left(\frac{1}{2} - y\right)\sin(2\pi x), \quad x,y \in \langle 0,1\rangle.$$

It has the form of (2.6) with $\varphi(x) = \frac{1}{2} - x$ and $\psi(y) = \sin(2\pi y)$. It is easy to see that $\int_0^1 \varphi(x)dx = 0$ and $\int_0^1 \psi(y)dy = 0$. Since $|\varphi(x)\psi(y)| \leq \frac{1}{2}$, it is in fact a pdf. Hence X, Y are uniformly distributed on $\langle 0,1\rangle$, dependent and uncorrelated, see also Figure 4. One more example in this fashion. Consider a vector (X_1, Y_1) with pdf

$$f_1(x,y) = \frac{1}{2} + \frac{1}{4}(|2(x-1)-1|)\sin(2\pi y) - \frac{1}{4}(|2(y-1)-1|)\sin(2\pi x),$$

where $(x,y) \in \langle 0,2\rangle \times \langle 0,1\rangle$ with $\varphi_1(x) = \frac{1}{4}|2(x-1)-1|$ and $\psi_1(y) = \psi(y)$ from the previous example. The joint pdf has also the form of (2.6) because of $\int_0^1 \varphi_1(x)dx = \int_0^2 \varphi_1(x)dx = 0$ and $|\varphi_1(x)\psi(y)|\frac{1}{4}$. Hence $\operatorname{cov}(X_1, Y_1) = 0$, X_1 is uniformly distributed on $\langle 0,2\rangle$ and Y_1 on $\langle 0,1\rangle$. ■

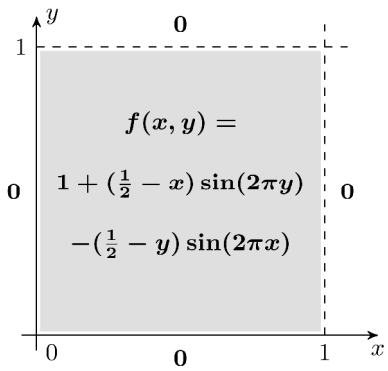


Figure 4: The joint pdf of dependent, uncorrelated random variables X, Y with uniform distribution on $\langle 0, 1\rangle$. The construction is based on (1.6).

Source: own elaboration.

3 Point processes on the line

We need here a few facts about order statistics. So we begin with a short introduction to this topic. Suppose that $X_1, \ldots, X_n$ are iid random variables each with cdf $F(x)$. If we arrange them in order of magniture $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ we call $X_{(k)}$ the k-th order statistic, $k = 1, \ldots, n$. In particular $X_{(1)} = \min\{X_1, \ldots, X_n\}$ and $X_{(n)} = \max\{X_1, \ldots, X_n\}$. Define the range W as $X_{(n)} - X_{(1)}$. Denote by $F_{(r)}(x)$ the cdf of $X_{(r)}$, $r = 1, \ldots, n$. It is well known that $F_{(n)}(x) = F^n(x)$ and $F_{(1)}(x) = 1 - (1 - F(x))^n$. But there is the general formula (see e.g. [4], Chapter 2)

$$F_{(r)}(x) = \sum_{k=1}^{n} \binom{n}{k} F^k(x)(1-F(x))^{n-k}, \quad r = 1, \dots, n$$

or in terms of pdf

$$f_{(r)}(x) = \frac{n!}{(r-1)!\,(n-r)!} f(x) F^{r-1}(x)(1-F(x))^{n-r}, \tag{3.1}$$

where $f(x) = F'(x)$. If $1 \le r < s \le n$ then the joint pdf of $X_{(r)}$ and $X_{(s)}$ equals

$$f_{(r)(s)}(x,y) = \frac{n!}{(r-1)!\,(s-r-1)!\,(n-s)!} f(x) F^{r-1}(x) \cdot (F(y)-F(x))^{s-r-1} f(y)(1-F(y))^{n-s}, \quad x \le y.$$

Example 3.1. We consider here a Poisson process on $\langle 0,1\rangle$ with intensity λ. So, take $\lambda > 0$ and choose n with probability $e^{-\lambda}\lambda^n/n!$, $n = 0,1, \dots$ *If* $n \ge 1$ we choose n points $t_1, \dots, t_n$ independently with uniform distribution on $\langle 0,1\rangle$. So this time (3.1) is applied for each n. Let $F_{S_{(M)}}(x) = \mathbb{P}(S_{(M)} \le x | A)$ be a conditional cdf, where A denotes an event that the number of points occuring is at least one. Because $\mathbb{P}(A) = 1 - e^{-\lambda}$ we have

$$F_{S_{(M)}}(x) = \frac{1}{1-e^{-\lambda}} \sum_{n=1}^{+\infty} e^{-\lambda} \frac{\lambda^n}{n!} \mathbb{P}(X_{(n)} \le x) = \frac{e^{-\lambda}}{1-e^{-\lambda}} \sum_{n=1}^{+\infty} \frac{\lambda^n}{n!} x^n = \frac{e^{\lambda x} - 1}{e^{\lambda} - 1}, \quad x \in \langle 0,1\rangle,$$

and the pdf of $S_{(M)}$ equals $f_{S_{(M)}}(x) = \lambda(e^{\lambda}-1)^{-1} e^{\lambda x}$. As for $S_{(1)}$ we have

$$F_{S_{(1)}}(x) = \frac{e^{-\lambda}}{1-e^{-\lambda}} \sum_{n=1}^{+\infty} \frac{\lambda^n}{n!} (1-(1-x)^n) = \frac{1-e^{-\lambda x}}{1-e^{-\lambda}}$$

and the pdf $f_{S_{(1)}}(x) = \lambda e^{\lambda}(e^{\lambda}-1)^{-1} e^{-\lambda x}$ where $x \in \langle 0,1\rangle$. By the symmetry it follows that the distribution of $S_{(M)}$ is the same as $1 - S_{(1)}$. Indeed, we have $F_{S_{(M)}}(x) = 1 - F_{S_{(1)}}(1-x)$. Computation gives

$$\mathbb{E}(S_{(M)}) = \frac{e^{\lambda}}{e^{\lambda}-1} - \frac{1}{\lambda}, \quad \lambda \in (0, +\infty).$$

Note that $\lim_{\lambda \to 0+} \mathbb{E}(S_{(M)}) = \frac{1}{2}$ and $\lim_{\lambda \to +\infty} \mathbb{E}(S_{(M)}) = 1$. Due to $S_{(M)} \stackrel{d}{=} 1 - S_{(1)}$ we have $\mathbb{E}(S_{(M)}) = 1 - \mathbb{E}(S_{(1)})$, so

$$\mathbb{E}(S_{(1)}) = \frac{1}{\lambda} - \frac{1}{e^{\lambda}-1}, \quad \lambda \in (0, +\infty),$$

see also Figure 6. In a similar way we find the pdf of $S_{(2)}$ (we assume here that there are at least two points in $\langle 0,1\rangle$). Because $\mathbb{P}(S(\langle 0,1\rangle) \geq 2) = 1 - e^{-\lambda}(1+\lambda)$ we have

$$f_{S_{(2)}}(x) = \frac{e^{-\lambda}}{1-e^{-\lambda}(1+\lambda)} \sum_{n=2}^{+\infty} \frac{\lambda^n}{n!} n(n-1)x(1-x)^{n-2} =$$
$$= \frac{\lambda^2 e^{\lambda}}{e^{\lambda}-1-\lambda} x e^{-\lambda x}, \quad x \in \langle 0,1\rangle.$$

We used here the fact that $f_{(2)}(x) = n(n-1)x(1-x)^{n-2}$ by (3.1). From this we obtain

$$\mathbb{E}(S_{(2)}) = \frac{2}{\lambda} - \frac{\lambda}{e^{\lambda}-1-\lambda}, \quad \lambda \in (0,+\infty).$$

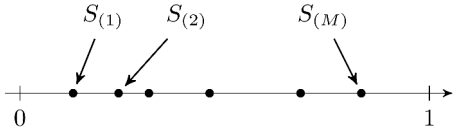


Figure 5. S is a Poisson process on $\langle 0,1\rangle$ with intensity $\lambda > 0$ and $S_{(1)}, \ldots, S_{(M)}$ are its order statistics.

Source: own elaboration.

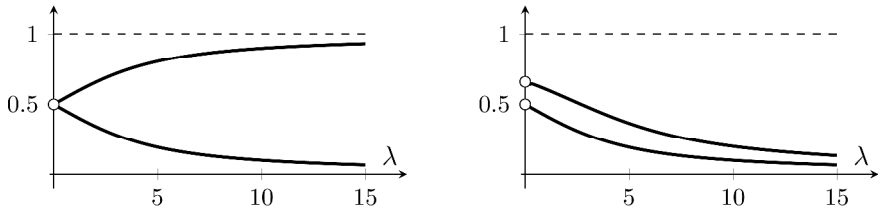


Figure 6. Comparison of $\mathbb{E}(S_{(M)})$ to $\mathbb{E}(S_{(1)})$ (left) and $\mathbb{E}(S_{(2)})$ to $\mathbb{E}(S_{(1)})$ (right).

Source: own elaboration.

Note that $\lim_{\lambda\to 0+} \mathbb{E}(S_{(2)}) = \frac{2}{3}$, $\lim_{\lambda\to+\infty} \mathbb{E}(S_{(2)}) = 0$ and obviously $\mathbb{E}(S_{(2)}) \geq \mathbb{E}(S_{(1)})$, see Figure 6. The joint pdf of $X_{(1)}$ and $X_{(n)}$ equals $n(n-1)(y-x)^{n-2}$ for $x \leq y$ and $n \geq 2$. Hence the joint pdf of $S_{(1)}$ and $S_{(M)}$ equals

$$h(x,y) = \frac{e^{-\lambda}\lambda^2}{1-e^{-\lambda}(1+\lambda)} \sum_{n=2}^{+\infty} \frac{\lambda^{n-2}}{(n-2)!}(y-x)^{n-2} =$$
$$= \frac{\lambda^2}{e^{\lambda}-1-\lambda} e^{\lambda(y-x)}, \quad 0 \le x \le y \le 1.$$

Denote by $f_W(z)$ the pdf of $W = S_{(M)} - S_{(1)}$ (the range of S). Then $\mathbb{P}(W \le z) = \int\int_D h(x,y)dxdy$, where $D = \{(x,y): y \le x+z\}$. Hence

$$F_W(z) = \int_0^{1-z}\int_x^{z+x} h(x,y)dxdy + \int_{1-z}^{1}\int_x^{1} h(x,y)dxdy$$
$$= \frac{\lambda}{e^{\lambda}-1-\lambda}\left[e^{\lambda z}\left(1-z+\frac{1}{\lambda}\right)-1-\frac{1}{\lambda}\right],$$

where $z \in \langle 0,1\rangle$. Obviously $F_W(z) = 1$ for $z \ge 1$ and $F_W(z) = 0$ if $z < 0$. Thus the pdf of W equals

$$f_W(z) = \frac{\lambda^2}{e^{\lambda}-1-\lambda} e^{\lambda z}(1-z), \quad z \in \langle 0,1\rangle. \tag{3.2}$$

If $\lambda \to +\infty$ then the graph of $f_W(z)$ is moved to 1, see Figure 7. ■

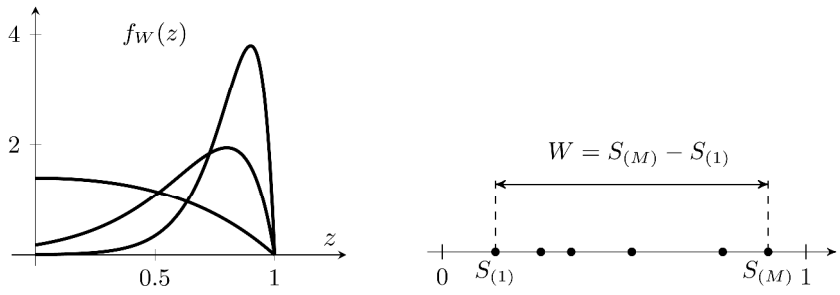


Figure 7. The pdf's of W (the range of S) for λ=1, 5 and 10.

Source: own elaboration.

Example 3.2. Define a point process in a half line in the following way. Choose $\lambda > 0$ and n with probability $e^{-\lambda}\lambda^n/n!$. If $n \ge 1$ we choose n points $t_1, \dots, t_n$ independently each with exponential distribution $\beta e^{-\beta x}$, where $x \ge 0$ and $\beta > 0$. This time $S_{(M)} \in (0, +\infty)$ and one gets

$$F_{S_{(M)}}(x) = \frac{e^{-\lambda}}{1-e^{-\lambda}} \sum_{n=1}^{+\infty} \frac{(\lambda(1-e^{-\beta x}))^n}{n!} = \frac{e^{\lambda}}{e^{\lambda}-1}\left(e^{\lambda(1-e^{-\beta x})}-1\right),$$

where $x \geq 0$. The pdf of $S_{(M)}$ is then

$$f_{S_{(M)}}(x) = \frac{\lambda\beta e^{\lambda}}{e^{\lambda}-1} e^{-\beta x-\lambda e^{-\beta x}}, \quad x \geq 0. \tag{3.3}$$

Remark 3.3. The probability distribution with the density

$$f_G(x) = \lambda\beta e^{-\beta x-\lambda e^{-\beta x}}, \quad x \in \mathbb{R}, \tag{3.4}$$

where $\beta, \lambda > 0$ is called the Gumbel distribution. Therefore (3.3) is the pdf of a truncated Gumbel distribution if we restrict its support to $(0, +\infty)$. Indeed, from the fact that $\int_0^{+\infty} f_G(x)dx = (e^{\lambda}-1)/e^{\lambda}$ we get (3.3), see Figure 8.

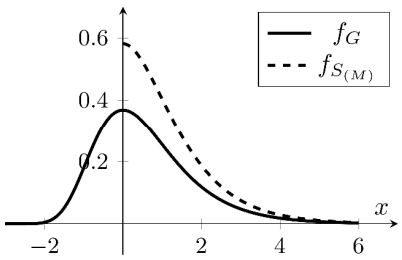


Figure 8. The pdf of Gumbel and truncated Gumbel distribution for $\lambda = 1$, $\beta = 1$, see (3.3) and (3.4).

Source: own elaboration.

As for $S_{(1)}$, the situation is different. Since $F(x) = 1 - e^{-\beta x}$ then $f_{(1)}(x) = 1 - e^{-\beta n x}$ and

$$f_{S_{(1)}}(x) = \frac{\lambda\beta}{e^{\lambda}-1} e^{-\beta x+\lambda e^{-\beta x}}, \quad x \geq 0.$$

Observe that $\lim_{x\to-\infty} e^{-\beta x+\lambda e^{-\beta x}} = +\infty$ therefore the above pdf cannot be extended to the whole line in a natural way, see also Figure 9. The joint pdf of $S_{(1)}$ and $S_{(M)}$ equals (again we assume that there are at least two points in $(0, +\infty)$)

$$f(x,y) = \frac{\beta^2\lambda^2}{e^{\lambda}-1-\lambda} e^{-\beta x-\beta y} e^{\lambda(e^{-\beta x}-e^{-\beta y})}, \quad 0 \leq x \leq y. \quad \blacksquare$$

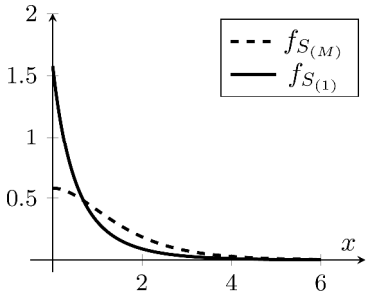


Figure 9. The pdf's of $f_{S_{(M)}}$ and $f_{S_{(1)}}$ for $\lambda = 1$, $\beta = 1$. The support of both densities is $\langle 0, +\infty)$.

Source: own elaboration.

In the next two examples we use the geometric distribution. So take a random variable X with $\mathbb{P}(X = n) = p(1-p)^{n-1}$, where $n = 1,2,3 \ldots$ and $p \in (0,1)$. Then $\mathbb{E}(X) = 1/p$ and

$$\lim_{p\to 0+} \mathbb{P}(X > k_0) = \lim_{p\to 0+} (1-p)^{k_0} = 1,$$

for any fixed $k_0 \in \mathbb{N}$. And of course $\lim_{p\to 0+} \mathbb{E}(X) = +\infty$.

Example 3.4. Sometimes the following Burr distribution (in fact its pdf) is very useful in the theory of point processes applied to physical phenomena, see e.g. [14],

$$f_B(x) = \frac{\alpha k \left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}}{\beta\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{k+1}}, \quad x \in (\gamma, +\infty), \tag{3.5}$$

where $k, \alpha > 0$ are shape parameters, $\beta > 0$ is a scale parameter and $\gamma \in \mathbb{R}$ is a location parameter, see Figure 10. The cdf of (3.5) is then

$$F_B(x) = 1 - \left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{-k}, \quad x \in (\gamma, +\infty). \tag{3.6}$$

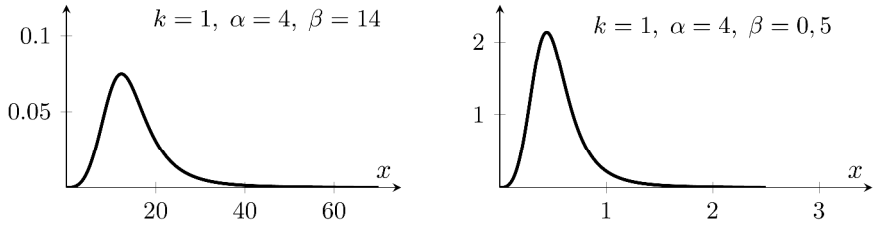


Figure 10. The pdf's of the Burr distribution for different parameters. In both cases the location parameter $\gamma = 0$.

Source: own elaboration.

We define a point process S on say $(\gamma, +\infty)$ in the following way. Take $p \in (0,1)$ and choose n with probability $p(1-p)^{n-1}$, $n \geq 1$. Next choose n points independently each with a given cdf $F(x)$. Assume that the support of $F(x)$ is also $(\gamma, +\infty)$. Then we have

$$F_{S_{(M)}}(x) = p\sum_{n=1}^{+\infty} q^{n-1}F^n(x) = \frac{p}{q}\sum_{n=1}^{+\infty}(qF(x))^n = \frac{pF(x)}{1-qF(x)},$$

where $q = 1-p$. We want to find $F(x)$ for which $F_{S_{(M)}}(x) = F_B(x)$. The solution to $pF(x)/(1-qF(x)) = F_B(x)$ is $F(x) = F_B(x)/(p+qF_B(x))$ so the answer is

$$F(x) = 1 - \frac{p}{(1+((x-\gamma)/\beta)^{\alpha})^k - q}, \quad x \in (\gamma, +\infty). \quad \blacksquare$$

Example 3.5. Construct a point process in a half line in the following way. As in the previous example take $p \in (0,1)$ and n with probability $p(1-p)^{n-1}$, $n \geq 1$. Next we choose n points $t_1, \ldots, t_n$ from $(0, +\infty)$ independently, each with exponential distribution $\beta e^{-\beta x}$. Hence if $x > 0$ then

$$F_{S_{(M)}}(x) = \frac{p}{q}\sum_{n=1}^{+\infty}\left(q(1-e^{-\beta x})\right)^n = \frac{p(e^{\beta x}-1)}{pe^{\beta x}+q},$$

and from this

$$f_{S_{(M)}}(x) = \frac{p\beta e^{\beta x}}{(pe^{\beta x}+q)^2}, \quad x \geq 0.$$

Computation gives

$$(S_{(M)}) = \frac{-\ln(p)}{\beta(1-p)}, \quad p \in (0,1), \qquad \beta > 0. \tag{3.7}$$

Observe that since p and β are independent parameters we have $\lim_{p\to 0+} \mathbb{E}(S_{(M)}) = +\infty$ for fixed β and $\lim_{\beta\to+\infty} \mathbb{E}(S_{(M)}) = 0$ if p is fixed. However if $\beta = \frac{-\ln(p)}{1-p}$ then $\mathbb{E}(S_{(M)}) = 1$, so one can control (3.7). ■

4 Generalized Poisson distribution

A random variable X has a generalized Poisson distribution if

$$P(X = n) = \frac{\lambda(\lambda + n\beta)^{n-1}}{n!} e^{-\lambda - n\beta}, \quad n = 0,1,2, ..., \tag{4.1}$$

where $\lambda > 0, \beta \in \langle 0,1\rangle$ are constants. One can also define this distribution for $\beta < 0$, see [2]. However if $\beta > 1$ then the right hand side of (4.1) is not a probability distribution and it was shown in [10] that

$$\sum_{n=0}^{+\infty} \frac{\lambda(\lambda + n\beta)^{n-1}}{n!} e^{-\lambda - n\beta} = e^{\lambda(\mu/\beta - 1)} < 1,$$

where μ is the solution to the equation $\mu e^{-\mu} = \beta e^{-\beta}$, see Figure 11. Note that $0 < \mu < \beta$ therefore the above sum is smaller than 1. For $\lambda = 1$ and $\beta = 1.1$ this sum is about 0.83. There are generally two ways of proving that (4.1) defines a probability distribution. The first type of proof uses the Lagrange expansion

$$\Phi(z) = \Phi(0) + \sum_{n=1}^{+\infty} \frac{\zeta^n}{n!} \cdot \frac{d^{n-1}}{dz^{n-1}} [f^n(z)\Phi'(z)]_{z=0},$$

where $\Phi(z)$ is an analytic function and $\zeta = z/f(z)$, see Jensen [8] (equation 6, p. 309; in fact this is the above expansion with $\Phi(z) = e^{\lambda z}$ and $f(z) = e^{\beta z}$). To the second type of proof (that is based on direct summation) one needs to apply the formula

$$\sum_{k=0}^{n} \binom{n}{k} (-1)^k (z+k)^l = \begin{cases} 0, & 0 \le l \le n-1 \\ (-1)^n n!, & l = n, \end{cases}$$

true for any complex number z, see [10], Lemma 2, or [15] for details. The reader should also look into an interesting article by Gould [7].

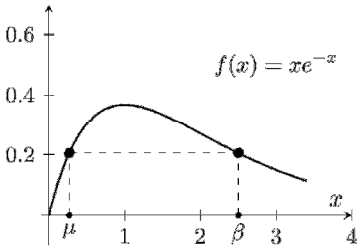


Figure 11. If $\beta > 1$ then the formula (4.1) does not define a probability distribution. Its sum equals $e^{\lambda(\mu/\beta-1)}$ where $\mu e^{-\mu} = \beta e^{-\beta}$.

Source: own elaboration.

Example 4.1. The moment generating function of X with a pdf given by (4.1) equals

$$M(t) = \mathbb{E}(e^{tX}) = e^{-\frac{\lambda}{\beta}W(-\beta e^{-\beta+t})-\lambda},$$

where $\beta \in (0,1\rangle$ and W is the Lambert's function defined as $W(x)e^{W(x)} = x$. The reader can find this for example in [2]. Hence $'(t) = M(t)\lambda e^{-\beta+t}W'(-\beta e^{-\beta+t})$. Note that $W(-\beta e^{-\beta}) = -\beta$ (it follows from the definition of W and this function is usually denoted by W_0 in this interval) and therefore $M(0) = e^{\lambda-\lambda} = 1$. In order to find $W'(-\beta e^{-\beta})$ we use the fact that $W'(x) = W(x)/(x(1 + W(x)))$. We have

$$W'(-\beta e^{-\beta}) = \frac{W(-\beta e^{-\beta})}{(-\beta e^{-\beta})(1 + W(-\beta e^{-\beta}))} = \frac{1}{e^{-\beta}(1-\beta)}.$$

Therefore $\mathbb{E}(X) = M'(0) = \lambda/(1-\beta)$. ■

Table 1. Basic facts about Poisson and generalized Poisson distribution. The Lambert W function is defined as $W(x)e^{W(x)} = x$.

distribution of X	Poisson	generalized Poisson
density	$e^{-\lambda}\lambda^n/n!$	$(\lambda(\lambda+n\beta)^{n-1})/n! \cdot e^{-\lambda-n\beta}$
support	$n = 0,1,2,\ldots$	$n = 0,1,2,\ldots$
parameters	$\lambda > 0$	$\lambda > 0$, $\beta \in \langle 0,1\rangle$
$\mathbb{E}(X)$	λ	$\lambda/(1-\beta)$
$D^2(X)$	λ	$\lambda/(1-\beta)^3$
$\mathbb{E}(e^{itX})$	$e^{\lambda(e^{it}-1)}$	$e^{-\frac{\lambda}{\beta}W(-\beta e^{-\beta+it})-\lambda}$

Source: own elaboration.

5 Multinomial distribution

In the theory of point processes, especially in the construction of a Poisson point process, a role is played by a multinomial distribution, see Kingman [9]. We will investigate one aspect of it in Example 5.1. Recall that a random vector $\vec{X} = (X_1, \dots, X_k)$ has a multinomial distribution with parameters n and $\vec{p} = (p_1, \dots, p_k)$ if

$$\mathbb{P}(X_1 = n_1, \dots, X_k = n_k) = \frac{n!}{n_1! \dots n_k!} p_1^{n_1} \dots p_k^{n_k}, \tag{5.1}$$

where $n_1 + \dots + n_k = n$ and $n_1, \dots, n_k \geq 0$. We assume also that $\sum_{i=1}^{k} p_i = 1$ and $p_i \in (0,1)$, $i = 1, \dots, k$. Note that X_i has the binomial distribution with parameters n and p_i, so $\mathbb{E}(X_i) = np_i$ and $Var\,(X_i) = np_i(1 - p_i)$. Denote $\mathbf{t} = (t_1, \dots, t_k)$. The characteristic function of $\vec{X}$ is $\varphi(\mathbf{t}) = f^n(\mathbf{t})$ where $f(\mathbf{t}) = p_1 e^{it_1} + \dots + p_k e^{it_k}$. From this we have

$$\varphi^{(l)}_{t_{i_1} t_{i_2} \dots t_{i_l}}(\mathbf{t}) = (\sqrt{-1})^l \frac{n!}{(n-l)!} p_{i_1} p_{i_2} \dots p_{i_l} e^{i(t_{i_1} + \dots + t_{i_l})} f^{n-l}(\mathbf{t}),$$

where $l \leq k$ and all $i_1, \dots, i_l$ are different. Hence we have

$$\mathbb{E}(X_{i_1} X_{i_2} \dots X_{i_l}) = \frac{n!}{(n-l)!} p_{i_1} p_{i_2} \dots p_{i_l},$$

and in particular $\operatorname{cov}(X_i, X_j) = -np_i p_j$ for $i \neq j$. The distribution (5.1) is in fact degenerate since if $n_1, \dots, n_{k-1}$ are fixed then n_k is completely determined.

Example 5.1. We will find the joint distribution of X_1, X_2 if the vector $\vec{X} = (X_1, X_2, X_3, X_4)$ has the following distribution

$$\mathbb{P}(\vec{X} = (n_1, n_2, n_3, n_4)) = \frac{4!}{n_1! n_2! n_3! n_4!} \left(\frac{1}{4}\right)^{n_1} \left(\frac{3}{8}\right)^{n_2} \left(\frac{1}{4}\right)^{n_3} \left(\frac{1}{8}\right)^{n_4}, \tag{5.2}$$

where $n_1 + n_2 + n_3 + n_4 = 4$. As we have already mentioned, the distribution of (X_1, X_2, X_3) is the same as $\vec{X}$, namely

$$\mathbb{P}(X_1 = n_1, X_2 = n_2, X_3 = n_3) = \\ = \frac{4!}{n_1! n_2! n_3! (4 - (n_1 + n_2 + n_3))!} \left(\frac{1}{4}\right)^{n_1} \left(\frac{3}{8}\right)^{n_2} \left(\frac{1}{4}\right)^{n_3} \left(\frac{1}{8}\right)^{4-(n_1+n_2+n_3)},$$

where $n_1 + n_2 + n_3 \leq 4$. From this we get

$$\mathbb{P}(X_1 = n_1, X_2 = n_2) = f(n_1, n_2) \left(\frac{1}{4}\right)^{n_1} \left(\frac{3}{8}\right)^{n_2}, \quad n_1 + n_2 \leq 4, \tag{5.3}$$

where coefficients $f(n_1, n_2)$ are given in Table 2. ■

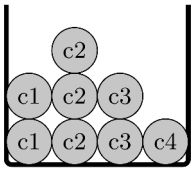


Figure 12. The urn contains 2 balls of color 1, 3 balls of colour 2, 2 balls of colour 3 and 1 ball of colour 4. We draw 4 balls with replacement. This urn model leads to multinomial distribution (5.2).

Source: own elaboration.

Table 2. The joint distribution of X_1, X_2 given by (5.3).

$X_1 \backslash X_2$	0	1	2	3	4
1	$\frac{81}{8^4}p_1^0p_2^0$	$\frac{108}{8^3}p_1^0p_2^1$	$\frac{54}{8^2}p_1^0p_2^2$	$\frac{12}{8}p_1^0p_2^3$	$p_1^0p_2^4$
2	$\frac{108}{8^3}p_1^1p_2^0$	$\frac{108}{8^2}p_1^1p_2^1$	$\frac{36}{8}p_1^1p_2^2$	$4p_1^1p_2^3$	0
3	$\frac{54}{8^2}p_1^2p_2^0$	$\frac{36}{8}p_1^2p_2^1$	$6p_1^2p_2^2$	0	0
4	$\frac{12}{8}p_1^3p_2^0$	$4p_1^3p_2^1$	0	0	0
5	$p_1^4p_2^0$	0	0	0	0

Source: own elaboration.

Example 5.2. One can apply (2.6) also for discrete random variables. For example, let X_1, X_2 be rv's with the following binomial distributions

$$\mathbb{P}(X_i = k) = \binom{2}{k} p_i^k (1-p_i)^{2-k}, \quad k = 0,1,2, \tag{5.4}$$

where $i = 1,2$ and $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{3}$. We use the algorithm described in [11] to construct the distribution of (X_1, X_2), see Figure 13. We have $\mathbb{E}(X_1X_2) = \frac{2}{3}$ with $\mathbb{E}(X_1) = 1$, $\mathbb{E}(X_2) = \frac{2}{3}$ hence $\operatorname{cov}(X_1, X_2) = 0$.

$X_2 \backslash X_1$	**0**	**1**	**2**
0	$\frac{16}{144}$	$\frac{35}{144}$	$\frac{13}{144}$
1	$\frac{13}{144}$	$\frac{32}{144}$	$\frac{19}{144}$
2	$\frac{7}{144}$	$\frac{5}{144}$	$\frac{4}{144}$

Figure 13. The joint pdf of dependent, uncorrelated random variables X_1, X_2 with marginal distributions given by (5.4).

Source: own elaboration.

Bibliography

[1] Baddeley A., Gregori P., Mateu J., Stoica R., Stoyan D. (editors), *Case Studies in Spatial Point Pattern Modelling*, "Lecture Notes in Statistics", No. 185, Springer, 2006.

[2] Ambagaspitiya, R. S., Balakrishnan, N. *On the compound generalized Poisson distributions*, "ASTIN Bull.", 24(2), 255-263, 1994.

[3] Conusl P. C., *Generalized Poisson distributions*, New York, Dekker, 1989.

[4] David H.A., Nagaraja H. N., *Order statistics*, Wiley, 2003.

[5] Gnedenko B.V., *The theory of probability*, 1967.

[6] Goldman J. R. *Stochastic point processes: limit theorems*, "Ann. of Math. Stat.", 38(3), 1967, 771–779.

[7] Gould H. W., *Euler's formula for n^{th} differences of powers*, "Am. Math. Monthly", Vol. 85, No. 6, 450–467, 1978.

[8] Jensen J. L. W. V., *Sur une identite d'Abel et sur d'autres formules analogues*, "Acta Math.", Vol. 26, 307–318, 1902.

[9] Kingman J. F. C., *Poisson processes*, Oxford, 1993.

[10] Lerner B., Lone A., Rao M., *On generalized Poisson distributions*, "Prob. Math. Stat.", vol. 17, 377-385, 1997.

[11] Serfling R. J., *Construction of dependent uncorrelated random variables with prescribed margin distributions*, Report M268, Florida State University, 1973

[12] Stoyanov J. M., *Counterexamples in Probability*, Wiley, 1997.

[13] Streit R.L., *Poisson point processes*, Springer, 2010.

[14] Tadikamalla P. R., *A look at the Burr and related distributions.*

[15] Tuenter H. J., *On the generalized Poisson distribution*, "Stat. Neerlandica", 54(3), 374–376, 2000.

Bartłomiej Stępień[1]

New algorithms for evaluation of standard uncertainty of long-term noise indicators

Keywords: uncertainty, non-classical statistics, interval estimation, bootstrap method

Abstract

This paper discusses various issues connected with the assessment of environmental noise pollution indicators and their uncertainty. Particular attention is paid to the process of interval estimation, which can be a promising tool for determining the standard uncertainty. Moreover, the two classical models of interval estimation and five non-classical algorithms based on the method of bootstrap resampling are mentioned. A theoretical basis and the methodology of determining the confidence intervals by using the proposed models is presented in detail. The most important properties of presented algorithms are discussed in terms of their applicability and effectiveness in the task of probabilistic environmental noise analysis. In addition, the possibility of using them to determine confidence intervals of acoustic indicators describing the environmental risk of noise and their standard uncertainty is analysed. Based on the analysis, alternative models of currently used estimation methods have been identified and can be used successfully in the statistical analysis of environmental noise pollution.

1 Introduction

Directive 2002/49/EC of the European Parliament [15] obligates the European Union countries to implement the common long-term policy of the environment protection against noise. Its realisation is based on the estimation of long-term noise indicators in the areas under protection. The two basic indicators are: the average A-weighted long-term day-evening-night level L_{DEN}, and the average A-weighted long-term night-time level L_{N}.

The basis for creating noise maps for sites under protection are the values of the above-mentioned long-term noise indicators. Any plans to prevent and reduce the harmful effects of noise in the environment are then associated with their values. These indicators characterise the acoustic climate over a long period. Most often it is assumed that this is one full calendar year, so values of the indicators depend on many factors (i.e. traffic intensity, structure of the vehicle

[1] Department of Mechanics and Vibroacoustics, Faculty of Mechanical Engineering and Robotics, AGH University of Science and Technology, e-mail: Bartlomiej.Stepien@agh.edu.pl

stream, average vehicle velocity, type and technical condition of the road surface, distance of the nearest buildings from the road edge, technical condition of the vehicles). Estimation of long-term noise hazard indicators requires access to results of an all-year-long sound level monitoring program. In practice, it is almost impossible to meet such a requirement. Therefore estimations of indicators are usually done on the basis of highly limited random sample. They are obtained as results of environmental sampling inspections. Sample size n is very small and ranges from few to a dozen or so elements [31,17,28].

The necessity of validation of the obtained results, which requires the analysis of uncertainty budget of estimation, is connected with the process of calculating the average long-term noise indicators determined by values L_{DEN} and L_{N}. An essential component of such budget is the type A standard uncertainty defined as the standard deviation of the mean from the inspections results. The rules given in the ISO/IEC Guide 98-3:2008 [1] are based on the point estimation methods and commonly used in the calculations. They are based on the classic variance estimators under restrictive assumptions (i.e. normality of measurements results, adequate sample size, lack of correlation between elements of the sample and observation equivalence). Results of acoustical measurements usually do not meet these assumptions [11,18,7].

A point estimation of noise indicators using classical [21,20,9] and non-classical [16,8,25] approach has been already performed. It should be noted that the probability of point estimation of a parameter being equal to the actual value of the estimated parameter is close to zero. There is no information about the distance between expected value of the estimated parameter and the true value of the population parameter in the point estimation. Overrating or underestimating values of noise indicators can have notable social and financial consequences.

For this reason, it seems to be necessary to examine the issue of confidence intervals of the expected value of long-term noise indicators. Because the point estimate is unlikely to be exactly correct, a range of values is usually specified in which the population parameter is likely to be. The confidence interval will include the true value of the population parameter with some probability. The interval estimation takes into account the estimation error for a given confidence level, as opposed to the point estimation.

For this reason, the interval data analysis is used in acoustics. This approach has been successfully applied, among other things, to real-time analysis of acoustic signal [19] and in uncertainty determination of the directional sound diffusion coefficient [26] as well as to planning measurement strategies [24,23]. The interval arithmetic finds application in modelling the railway noise [3] and in determination of other acoustic parameters such as reverberation time of rooms [4] and partitions sound insulation [5] and its uncertainty [6]. However, interval estimation algorithm based on kernel density estimator [29] is used in the analysis of long-term noise indicators.

The interval estimation can be applied to uncertainty evaluation what is presented in ISO/IEC Guide 98-3/Suppl.1:2008 [2]. This document describes application of the Monte Carlo method in case of typical problems of the uncertainty evaluation to which include:

- the contributory uncertainties are not of approximately the same magnitude,
- it is difficult or inconvenient to provide the partial derivatives of the model,
- the probability density function for the output quantity is not a Gaussian distribution or a scaled and shifted *t*-distribution,
- an estimate of the output quantity and the associated standard uncertainty are approximately of the same magnitude,
- the models are arbitrarily complicated,
- the probability density functions for the input quantities are asymmetric.

However, Monte Carlo applications within the framework of the long-term indicators assessment present two problems. First, Monte Carlo analysis requires knowledge of the probability distributions of the parameters under consideration and such distributions are rarely well defined. Second, acoustical measurements are expensive and Monte Carlo requires an often unfeasibly large number of measured realizations to obtain statistically meaningful results.

For this reason, it seems to be necessary to implement solutions of non-classical statistic for solving these problems. These techniques are based on non-parametric statistical methods, allowing determining the distribution of a random variable without any information on belonging or not to any specific class of distributions and with a limited sample size. For this reason, in the following it is proposed to use different models for constructing confidence intervals by means of the bootstrap method. The bootstrap method has been successfully applied to point estimation of expected value and uncertainty of noise indicators [16,25].

2 Selected models of interval estimation

While analysing the measurement data we need to remember that estimation of the mean value and standard deviation of normal distribution or estimation of an exponential distribution parameter is basically equivalent to the estimation of the probability distribution of population from which the random sample is taken. Actually, the estimation of the aforementioned parameters is equivalent to estimation of the density function of population. The fact that to estimate the probability density function it suffices to calculate a finite number of numerical estimators is a result of an assumption of a relatively accurate knowledge of a probabilistic model which governs the examined phenomenon – we have assumed that we know this model with the accuracy to a finite number of numerical parameters. In cases presented above, the estimation of parameters which define the unknown distribution of population can therefore be called a parametric estimation of probability distribution. The parametric estimation requires an

adequate random sample size. In practice $n \geq 30$ is often considered an adequate random sample size [22].

At the same time, a histogram is another estimation of the unknown population density. Being discrete, the histogram can be replaced with a continuous estimator, e.g. an adequate kernel density estimator or an estimator based on splines. These estimators of unknown probability density function do not need any assumptions about the sought form of a function and therefore are called non-parametric estimators. This family includes also distribution estimators based on the jackknife and bootstrap methods and on the Bayes' theorem widened to include the probability distributions. The basic advantage of non-parametric estimators is possibility to draw inference from a small random sample of a few to a dozen elements which does not have asymptotic properties. It is for this advantage that non-parametric estimators are increasingly used in the probabilistic analysis of environmental noise.

Below, it will be presented in detail two generally used parametric models and five different techniques for constructing confidence intervals using the bootstrap method. The discussed parametric models are based on the assumption that the analysed sample comes from the normal distribution population with a known ($N\sigma_{\mathrm{k}}$ model) or unknown ($N\sigma_{\mathrm{u}}$ model) standard deviation. However, the first two non-parametric methods are based on bootstrap "tables", and bootstrap percentiles are used in the next three algorithms.

2.1 Parametric models ($N\sigma_{\mathrm{k}}$ and $N\sigma_{\mathrm{u}}$ models)

Consider a random sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ from a normal distribution $N(\mu, \sigma)$ with a known standard deviation σ ($N\sigma_{\mathrm{k}}$ model). It is known that the mean from the sample $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ has a normal distribution $N\left(\mu, \sigma/\sqrt{n}\right)$. Therefore, the random variable

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx N(0,1) \tag{1}$$

has a standard normal distribution $N(0,1)$ [22]. The interval to which values of random variable Z belong with probability $1 - \alpha$, where α is a known number from the interval (0, 1) is given

$$p\left(z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}\right) = 1 - \alpha, \tag{2}$$

where $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are the $(100 \cdot \alpha/2)$th and $100(1 - \alpha/2)$th percentile points of a standard normal distribution, respectively. These values are given in the standard normal table (e.g. $z_{0.025} = -1.960$). After substituting the right-hand side of the expression (1) in place of Z and rearranging, the confidence interval for the $N\sigma_{\mathrm{k}}$ model is given by [22]

$$p\left(-z_{1-\alpha/2} \le Z \le z_{1-\alpha/2}\right) = p\left(\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha. \quad (3)$$

Most frequently, the standard deviation of population density is unknown ($N\sigma_{\mathrm{u}}$ model). Therefore, the random variable Z given by the equation (1) can be replaced with the random variable [22]

$$T = \frac{\bar{x}-\mu}{S/\sqrt{n}} \approx t_{n-1}. \quad (4)$$

This idea is not only natural but appropriate as well because the distribution of the random variable T does not depend on the unknown parameter σ and is known. Namely, it is a so-called *t*-distribution (also called Student's distribution or Student's *t*-distribution) with $n-1$ degrees of freedom.

Knowing the random variable T and its distribution t_{n-1}, the confidence interval for μ can be written analogously to the previous case. The $N\sigma_{\mathrm{u}}$ confidence interval of intended coverage $1-\alpha$ is defined by [22] as

$$p\left(\bar{x} - t_{1-\alpha/2,n-1}\frac{\hat{s}}{\sqrt{n}} \le \mu \le \bar{x} + t_{1-\alpha/2,n-1}\frac{\hat{s}}{\sqrt{n}}\right) = 1-\alpha, \quad (5)$$

where $t_{1-\alpha/2,n-1}$ indicates the $100(1-\alpha/2)$th percentile point of a t_{n-1} distribution and $p\left(T \le t_{1-\alpha/2,n-1}\right) = 1-\alpha/2$, whereas $\hat{s}$ is an unbiased estimator of standard distribution whose value is defined as

$$\hat{s} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}. \quad (6)$$

3 Assumptions and ideas of the bootstrap method

Consider an observed random sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ from an unknown probability distribution F with intent to estimate a parameter of interest $\theta = t(F)$ on the basis of **x**. For this purpose, let an estimate $\hat{\theta} = s(x)$ from **x** be calculated.

The bootstrap method was introduced in 1979 by Bradley Efron [12] as a computer-based method for estimating the standard error of $\hat{\theta}$. The bootstrap estimate of standard error requires no theoretical calculations and is available no matter how mathematically complicated the estimator $\hat{\theta} = s(\mathbf{x})$ may be.

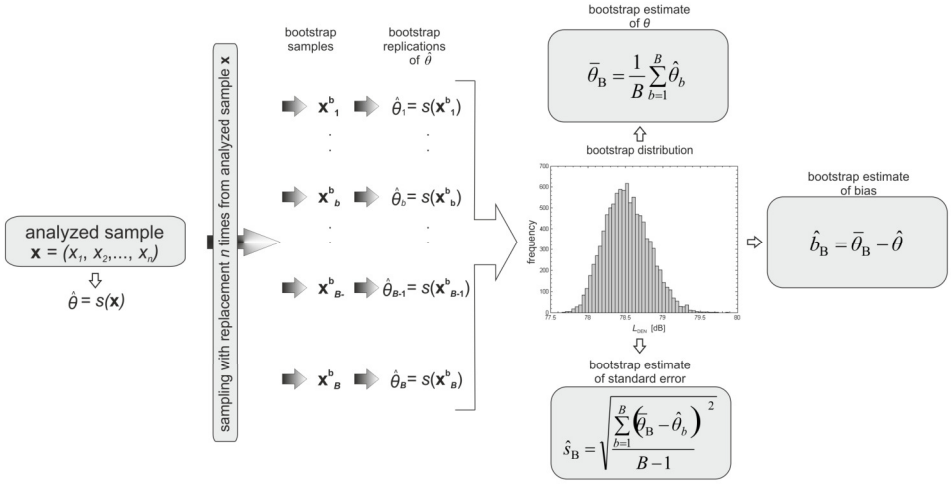


Figure 1: Schematic diagram of the bootstrap method.

Source: own elaboration.

Bootstrap methods depend on the concept of a bootstrap sample. Let $\hat{F}$ be the empirical distribution, assigning probability $1/n$ to each of the observed values $x_i, i = 1,2, \dots, n$. A bootstrap sample is defined as a random sample of size n drawn from $\hat{F}$, say $\mathbf{x}^b = (x_1^{\mathrm{b}}, x_2^{\mathrm{b}}, \dots, x_n^{\mathrm{b}})$ [14],

$$\hat{F} \rightarrow (x_1^{\mathrm{b}}, x_2^{\mathrm{b}}, \dots, x_n^{\mathrm{b}}). \tag{7}$$

The symbol "**b**" indicates that $\mathbf{x}^b$ is not the actual data set **x**, but rather a re-sampled version of **x**.

Symbolic expression (1) can be also verbalised as follows: the bootstrap data points $x_1^{\mathrm{b}}, x_2^{\mathrm{b}}, \dots, x_n^{\mathrm{b}}$ are a random sample of size n drawn with replacement from the population of n objects $(x_1, x_2, \dots, x_n)$. The bootstrap data set $(x_1^{\mathrm{b}}, x_2^{\mathrm{b}}, \dots, x_n^{\mathrm{b}})$ consists of elements of the original data set $(x_1, x_2, \dots, x_n)$.

Corresponding to a bootstrap data set $\mathbf{x}^b$ is a bootstrap replication of $\hat{\theta}$

$$\hat{\theta}_b = s(\mathbf{x}^b). \tag{8}$$

The quantity $s(\mathbf{x}^b)$ is the result of applying to $\mathbf{x}^b$ the same function $s(\bullet)$ as this applied to **x**.

3.1 Point estimation of distribution parameters by bootstrap method

Point estimation of an unknown distribution parameter θ of the examined variable is based on assuming that the estimator value of this parameter at the given sample is its estimation. Figure 1 presents a schematic diagram of the algorithm of point estimation of bootstrap distribution parameters. By applying the

Monte Carlo method to the bootstrap, a bootstrap sample B is generated. The bootstrap samples are generated from the original data set (analysed sample). Each bootstrap sample has n elements generated by sampling with replacement n times from the analysed sample. Bootstrap replications $\hat{\theta}_1, \ldots, \hat{\theta}_b, \ldots, \hat{\theta}_B$ are obtained by calculating the value of the statistics $s(\mathbf{x})$ on each bootstrap sample. The mean of these values can be assumed to be an assessment of parameter θ. Thus, the assessment of parameter θ can be expressed as [14]

$$\bar{\theta}_{\mathrm{B}} = \frac{1}{B}\sum_{i=1}^{B} \hat{\theta}_b. \tag{9}$$

The bootstrap estimate of the standard error is the standard deviation of the bootstrap replications [14]:

$$\hat{s}_{\mathrm{B}} = \sqrt{\frac{\sum_{b=1}^{B} (\bar{\theta}_{\mathrm{B}} - \hat{\theta}_b)^2}{B-1}}. \tag{10}$$

Further, the bootstrap estimate of bias $\hat{b}_B$ based on the B replications is defined by

$$\hat{b}_{\mathrm{B}} = \bar{\theta}_{\mathrm{B}} - \hat{\theta}, \tag{11}$$

where $\bar{\theta}_{\mathrm{B}}$ is bootstrap estimate of parameter θ and $\hat{\theta}$ is estimate of parameter θ from the original sample $\mathbf{x}$. Note that the algorithm of Figure 1 applies exactly to calculation of (5), except that at the last step, $\bar{\theta}_{\mathrm{B}} - \hat{\theta}$ is calculated rather than $\hat{s}_{\mathrm{B}}$. Of course, both $\hat{s}_{\mathrm{B}}$ and $\hat{b}_{\mathrm{B}}$ can be calculated from the same set of bootstrap replications.

4 Confidence intervals based on bootstrap method for the expected value

This section describes different techniques for constructing confidence intervals using the bootstrap method presented in Section 3. The first two methods are based on bootstrap ”tables”, and bootstrap percentiles are used in the next three algorithms.

4.1 Normal approximated interval with bootstrapped bias and standard error method (*norm* model)

Consider a one-sample situation where the data are obtained by random sampling from an unknown distribution. Let $\hat{\theta}$ be the estimate of a parameter from the original sample of interest θ, and let $\bar{\theta}_{\mathrm{B}}$ be the bootstrap estimate of parameter θ. Let $\hat{s}_{\mathrm{B}}$ be the bootstrap estimate of standard error of $\bar{\theta}_{\mathrm{B}}$, and $\hat{b}_{\mathrm{B}}$ be the bootstrap estimate of bias. Within this model, the confidence interval [10]

$$P\{\hat{\theta} - \hat{b}_{\mathrm{B}} + \hat{s}_{\mathrm{B}} \cdot z_{\alpha/2} < E(\theta) < \hat{\theta} - \hat{b}_{\mathrm{B}} - \hat{s}_{\mathrm{B}} \cdot z_{\alpha/2}\} = 1 - \alpha, \tag{12}$$

where $z_{\alpha/2}$ indicates the $(100 \cdot \alpha/2)$th percentile point of a $N(0,1)$ distribution, as given in the standard normal table (e.g. $z_{0.025} = -1.960$).

4.2 Studentized confidence interval method (*bootstrap-t* or *stud* model)

Through the use of the bootstrap method, it is possible to obtain accurate intervals without necessity to make normal theory assumptions. This section describes one way to get such intervals, namely the *bootstrap-t* or *stud* approach. This procedure estimates the distribution of $Z = \frac{\hat{\theta}^* - \theta}{\hat{s}}$ directly from the data. In principle, it builds a bootstrap table. The bootstrap table is built by generating B bootstrap samples, and then computing the bootstrap version of Z for each sample. The bootstrap table consists of percentiles of these B values.

More specifically, in the *bootstrap-t* model B bootstrap samples are generated and for each sample the quantity

$$\hat{Z}_b = \frac{\hat{\theta}_b - \hat{\theta}}{\hat{s}_b} \approx \hat{t} \tag{13}$$

is computed, where $\hat{\theta}_b$ is the value of $\hat{\theta}$ for the bootstrap sample and $\hat{s}_b$ is the estimated standard error of $\hat{\theta}_b$ for the bootstrap sample. The $(\alpha/2)$th percentile of $\hat{Z}_b$ is estimated by the value $\hat{t}_{\alpha/2}$ such that

$$\frac{\#\{\hat{Z}_b \leq \hat{t}_{\alpha/2}\}}{B} = \frac{\alpha}{2}. \tag{14}$$

Finally, the *bootstrap-t* confidence interval [14]

$$P\{\hat{\theta} - \hat{t}_{1-\alpha/2} \cdot \hat{s}_{\mathrm{B}} < E(\theta) < \hat{\theta} - \hat{t}_{\alpha/2} \cdot \hat{s}_{\mathrm{B}}\} = 1 - \alpha, \tag{15}$$

where $\hat{\theta}$ is the estimate of a parameter from the original sample, and $\hat{s}_{\mathrm{B}}$ is the bootstrap estimate of standard error of $\bar{\theta}_{\mathrm{B}}$.

4.3 Basic percentile method (*per* model)

This and the next two subsections present another approach to bootstrap confidence intervals based on percentiles of the bootstrap distribution of a statistic. The methods represent a different view of the standard normal-theory interval which leads to the percentile interval as a generalisation based on the bootstrap technique.

Let $\hat{G}$ be the cumulative distribution function of $\hat{\theta}$. The $1 - \alpha$ percentile interval is defined by $\alpha/2$ and $1 - \alpha/2$ percentiles of $\hat{G}$ [14]:

$$[\hat{\theta}_{\%,lo}, \hat{\theta}_{\%,up}] = [\hat{G}^{-1}(\alpha/2), \hat{G}^{-1}(1-\alpha/2)]. \tag{16}$$

Since, by definition, $\hat{G}^{-1}(\alpha/2) = \hat{\theta}_{\alpha/2}$, the $(100 \cdot \alpha/2)$th percentile of the bootstrap distribution, the percentile interval can be also written as

$$P\{\hat{\theta}_{(\alpha/2)} < E(\theta) < \hat{\theta}_{(1-\alpha/2)}\} = 1-\alpha. \tag{17}$$

Expressions (16) and (17) refer to the ideal bootstrap situation in which the number of bootstrap replications is infinite. In practice, it is necessary to use some finite number B of replications. The approximate $1-\alpha/2$ percentile interval [14]

$$P\{\hat{\theta}_{\mathrm{B}}^{(\alpha/2)} < E(\theta) < \hat{\theta}_{\mathrm{B}}^{(1-\alpha/2)}\} = 1-\alpha, \tag{18}$$

where $\hat{\theta}_{\mathrm{B}}^{(\alpha/2)}$ and $\hat{\theta}_{\mathrm{B}}^{(1-\alpha/2)}$ are the $(100 \cdot \alpha/2)$th and $100 \cdot (1-\alpha/2)$th empirical percentiles of the $\hat{\theta}_b$ values, respectively.

4.4 Bias-corrected percentile method (*cper* model)

Construction of percentile intervals according to *cper* method is more complicated than this of percentile intervals (*per*), but their use is almost as easy.

The *cper* interval endpoints are also given by percentiles of the bootstrap distribution, but not necessarily the same ones as in (18). The percentiles used depend on one number $\hat{z}_0$ called the bias-correction [14]. Later it will be described how $\hat{z}_0$ is obtained, but first the definition of *cper* interval endpoints must be given.

The *cper* interval of intended coverage $1-\alpha$ is defined by [13] as

$$P\{\hat{\theta}_{\alpha 1} < E(\theta) < \hat{\theta}_{\alpha 2}\} = 1-\alpha \tag{19}$$

where

$$\begin{aligned} \alpha 1 &= \Phi(2\hat{z}_0 - z_{\alpha/2}), \\ \alpha 2 &= \Phi(2\hat{z}_0 + z_{1-\alpha/2}) \end{aligned} \tag{20}$$

Here $\Phi(\bullet)$ is the standard normal cumulative distribution function, and $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are the $(100 \cdot \alpha/2)$th and $100 \cdot (1-\alpha/2)$th percentile points of a standard normal distribution, respectively.

The value of the bias-correction $\hat{z}_0$ is obtained directly from the ratio of bootstrap replications less than the original estimate $\hat{\theta}$ [13],

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\hat{\theta}_b < \hat{\theta}\}}{B}\right) \tag{21}$$

where $\Phi^{-1}(\bullet)$ denotes the inverse function of a standard normal cumulative distribution function.

4.5 Bias-corrected and accelerated percentile method (*BCa* model)

Another algorithm is based on knowledge of percentiles of the bootstrap distribution. The *BCa* interval endpoints used depend on two numbers, $\hat{a}$ and $\hat{z}_0$ (21), called the acceleration and the bias-correction [10], respectively.

The *BCa* interval of intended coverage $1-\alpha$ is given by [14]

$$P\{\hat{\theta}_{\alpha 1} < E(\theta) < \hat{\theta}_{\alpha 2}\} = 1-\alpha \tag{22}$$

where

$$\begin{aligned} \alpha 1 &= \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1-\hat{a}(\hat{z}_0 + z_{\alpha/2})}\right), \\ \alpha 2 &= \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1-\hat{a}(\hat{z}_0 + z_{1-\alpha/2})}\right) \end{aligned} \tag{23}$$

Here $\Phi(\bullet)$ is the standard normal cumulative distribution function, and $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are the $(100\cdot\alpha/2)$th and $100\cdot(1-\alpha/2)$th percentile points of a standard normal distribution, respectively. The value of the bias-correction $\hat{z}_0$ is obtained from (21).

There are various ways to compute the acceleration $\hat{a}$. The easiest can be expressed in terms of jackknife values $\hat{\theta}$. A simple formula for the acceleration has the form [14]

$$\hat{a} = \frac{\sum_{i=1}^{n}(\hat{\theta}_{\text{jack}} - \hat{\theta}_{i\text{jack}})^3}{6\{\sum_{i=1}^{n}(\hat{\theta}_{\text{jack}} - \hat{\theta}_{i\text{jack}})^2\}^{3/2}}, \tag{24}$$

where $\hat{\theta}_{\text{jack}}$ is the jackknife estimate of parameter $\hat{\theta}$, and $\hat{\theta}_{i\text{jack}}$ are jackknife replications of $\hat{\theta}$.

It must be explained at this point how $\hat{\theta}_{\text{jack}}$ and $\hat{\theta}_{i\text{jack}}$ from (24) are to be obtained by means of the jackknife method. Suppose a sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and an estimator $\hat{\theta} = s(\mathbf{x})$ are given. The jackknife is focused on samples that leave out one observation at a time [14]

$$x_{i\text{jack}} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \tag{25}$$

for $i = 1,2,\ldots,n$, called the jackknife samples. The *i*th jackknife sample consists of the data set with the *i*th observation removed. Let

$$\hat{\theta}_{i\text{jack}} = s(x_{i\text{jack}}) \tag{26}$$

be the *i*th jackknife replication of $\hat{\theta}$.

The jackknife estimate $\hat{\theta}_{\text{jack}}$ of parameter $\hat{\theta}$ is defined as [14]

$$\hat{\theta}_{\text{jack}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{i\text{jack}}. \tag{27}$$

5 Conclusion

This paper discusses various issues connected with the assessment of environmental noise pollution indicators and their uncertainty. Particular attention is paid to the process of interval estimation, which can be a promising tool for determining the type A uncertainty. Currently, it is determined by using the classical variance estimators, which raises many doubts in relation to the results of acoustic research. In practice of measurement, the analysis of environmental noise pollution is based on a small sample size (results of the control tests), which does not have asymptotic properties. For this reason, attention has been paid to the necessity of implementation of non-classical statistical solutions. The most important accompanying assumption is the lack of limitations concerning the form and properties of analysed statistics as well as the size of analysed samples.

A theoretical basis and the methodology of determining the confidence intervals for the two classical models and five non-classical algorithms based on the bootstrap resampling method was presented in this paper. The possibility of using them to determine the confidence intervals of acoustic indicators describing the environmental pollution by noise and their type A standard uncertainty was analysed.

In a classical approach, otherwise called frequency framework ($N\sigma_{\text{k}}$, $N\sigma_{\text{u}}$ models), confidence interval for the parameter θ is called a random interval $(\theta_{lo}, \theta_{up})$, where θ_{lo} and θ_{up} are the functions of a random variable that $P(\theta_{lo} < \theta < \theta_{up}) = 1 - \alpha$. The number $1 - \alpha$ is called the confidence level. In the frequency approach, the estimated parameter θ is a fixed constant, therefore an unknown value of this parameter can be covered by this interval or not. However, during a long series of observed trials, the frequencies of the compartments containing the actual value of the unknown parameter θ is approximately equal to $1 - \alpha$, that is $100(1 - \alpha)\%$ of compartments will contain the estimated parameter. Thus, in the frequency approach, the resulting confidence interval $(\theta_{lo}, \theta_{up})$ can be interpreted as follows: confidence interval $(\theta_{lo}, \theta_{up})$ is one of these intervals, which with confidence probability $1 - \alpha$ contain the estimated parameter θ. However, it is unacceptable to interpret that the parameter values change from θ_{lo} to θ_{up} because parameter θ is not a random variable, but only an unknown constant.

In conclusion, in classical approach, the interpretation of the confidence interval should indicate the variability of the limits of the confidence interval rather than the parameter itself. However, it should be emphasized that, in prac-

tice, we have at our disposal only one random sample, on the basis of which we determine the endpoints of a confidence interval, and then we do not know whether the estimated parameter belongs to this interval or not.

Non-classical statistics offers different approach to the confidence intervals. In the bootstrap approach, the estimated parameter θ is a random variable, therefore, it has the potential to assume different values with certain probabilities. In this approach, confidence intervals can be calculated precisely at a predetermined level of probability. On this basis, it can be argued that with probability $1-\alpha$, the estimated parameter will belong to the estimated confidence interval. In view of the above, conclusions which are probabilistic statements can be formulated.

Without any prior theoretical assumptions it should be stated that non-classical methods of interval estimation based on the distribution of parameters, are an important and attractive alternative to the classical estimation method, more reliable than the other two estimation methods presented in this paper.

Additionally, it must be emphasised that the bootstrap method can be successfully applied to a small random sample not having any asymptotic properties.

Analysing further properties of the obtained confidence intervals it should be also mentioned that percentiles of distributions $N(0,1)$ and t_{n-1} used to determine the confidence interval limits in the $N\sigma_{\mathrm{k}}$, $N\sigma_{\mathrm{u}}$ and *norm* models are symmetrical with respect to 0 (zero). As a result, the obtained confidence intervals are symmetrical with respect to point estimate of noise level indicators. Percentiles used in the other models (*per*, *cper*, *BCa*, and *stud*) are asymmetrical with respect to 0 (zero). This is the reason for which intervals shifted more or less to the left or right with respect to the point estimate are obtained. This asymmetry reflects probabilistic properties of the examined noise indicators.

The *stud* model can be characterised poor resistance to outlying values occurring in the original sample. Outlying observations have a very substantial effect on value of $\hat{Z}_b$ determined from (13), and as a consequence, on values of percentiles $\hat{t}_{1-\alpha/2}$ and $\hat{t}_{\alpha/2}$ necessary to establish the lower and the higher limit of the confidence interval (15).

Accordingly, three models, i.e. *per*, *cper*, *BCa*, can be recommended for the determination of confidence intervals and type A uncertainty, not only of noise indicators, but also other acoustic parameters.

Acknowledgements

The project described in this paper has been executed within the project No. 11.11.130.955.

Bibliography

[1] *ISO/IEC Guide 98-3:2008: Uncertainty of measurement. Part 3: Guide to the expression of uncertainty in measurement (GUM:1995)*, ISO/IEC Guide 98-3, 2008.

[2] *ISO/IEC Guide 98-3/Suppl.1:2008: Propagation of distributions using a Monte Carlo method*, ISO/IEC Guide 98-3/Suppl.1:2008.

[3] Batko W., Pawlik P., *New approach to the uncertainty assessment of acoustic effects in the environment*, "Arch. Acoust.", 37(1), 57–61, 2012.

[4] Batko W., Pawlik P., *Uncertainty evaluation in modelling of acoustic phenomena with uncertain parameters using interval arithmetic*, "Acta Phys. Pol. A", 121(1-A), A–152–A–155, 2012.

[5] Batko W., Pawlik P., *New method of uncertainty evaluation of the sound insulation of partitions*, "Acta Phys. Pol. A", 123(6), 1012–1015, 2013.

[6] Batko W., Pawlik P., *Uncertainty of sound insulation estimation*, "Measurement Automation and Monitoring", 59(1), 26–27, 2013.

[7] Batko W., Przysucha B.*, Determination of the probability distribution of the mean sound level*, "Arch. Acoust.", 35(4), 543–550, 2010.

[8] Batko W., Stępień B., *Non-parametric methods of estimation of type A uncertainty of the environmental noise hazard indices*, "Arch. Acoust.", 34(3), 295–303, 2009.

[9] Borkowski B., Czajka I., Pluta M., Suder-Dębska K., *The conceptual design of dynamic acoustic maps to assess noise exposure*, "Pol. J. Environ. Stud.", 25(4), 1415–1420, 2016.

[10] Davison A. C., Hinkley D. V., *Bootstrap methods and their application*, Cambridge University Press, New York, 1997.

[11] Don C. G., Rees I. G., *Road traffic sound level distributions*, "J. Sound Vibr.", 100(1), 41–53, 1985.

[12] Efron B., *Bootstrap methods: another look at the jackknife*, "Ann. Stat.", 7(1), 1–26, 1979.

[13] Efron B., *The jackknife, the bootstrap and other resampling plans*, Society for Industrial and Applied Mathematics, Philadelphia, 1982.

[14] Efron B., Tibshirani R. J., *An introduction to the bootstrap*, Chapman & Hall/CRC, New York, 1993.

[15] European Parliament, *Directive 2002/49/EC of the European Parliament and of the Council of 25 June 2002 relating to the assessment and management of environmental noise*, 2002.

[16] Farrelly F. A.. Brambilla G., *Determination of uncertainty in environmental noise mesurements by bootstrap method*, "J. Sound Vibr.", 268(1), 167–175, 2003.

[17] Gaja E., Giménez A., Sancho S., Reig A., *Sampling techniques for the estimation of the annual equivalent noise level under urban traffic conditions*, "Appl. Acoust.", 64(1), 43–53, 2003.

[18] Giménez A., González M., *A stochastic model for the noise levels*, "J. Acoust. Soc. Am.", 125(5), 3030–3037, 2009.
[19] Heiss A., Krapf K. G., *Quantification of uncertainty by real time confidence limits in separation of sound immission levels*, "Noise Control Eng. J.", 55(1), 149–158, 2007.
[20] Jiménez S., Genescà M., Romeu J., Sanchez A., *Estimation of night traffic noise levels*, "Acta Acust united Ac", 94(4), 563–567, 2008.
[21] Kephalopoulos S., Paviotti M., Knauss D., Bérengier M., *Uncertainties in long-term road noise monitoring including meteorological influences*, "Noise Control Eng. J"., 55(1), 133–141, 2007.
[22] Koronacki J., Mielniczuk J., *Statistics for students of technical and natural fields* (in Polish), Scientific and Technical Publishers, Warsaw 2004.
[23] Liguori C., Ruggiero A., Russo D., Sommella P., *Estimation of the minimum measurement time interval in acoustic noise*, "Appl. Acoust.", 127, 126–132, 2017.
[24] Liguori C., Ruggiero A., Russo D., Sommella P., *Innovative bootstrap approach for the estimation of minimum measurement time interval in road traffic noise evaluation*, "Measurement", 98, 237–242, 2017.
[25] Mateus M., Dias Carrilho J. A., Gameiro da Silva M. C., *Assessing the influence of the sampling strategy on the uncertainty of environmental noise measurements through the bootstrap method,* "Appl. Acoust.", 89, 159–165, 2015.
[26] Pilch A., *Sources of measurement uncertainty in determination of the directional diffusion coefficient value*, "Appl. Acoust.", 129, 268–276, 2018.
[27] Romeu J., Jiménez S., Genescà M., Pàmies T., Capdevila R., *Spatial sampling for night levels estimation in urban environments*, "J. Acoust. Soc. Am.", 120(2), 791–800, 2006.
[28] Schomer P. D., DeVor R. E., *Temporal sampling requirements for estimation of long-term average sound levels in the vicinity of airports*, "J. Acoust. Soc. Am.", 69(3), 713–719, 1981.
[29] Stępień B., *Confidence intervals for the long-term noise indicators using the kernel density estimator*, "Arch. Acoust.", 41(3), 517–525, 2016.

Tadeusz Banek[1]

Disinformation – advanced weapon in political and military games: basic ideas – non-matrix games

Keywords: disinformation, game theory, probability theory, strategy, politics.

Abstract

We analyse the problem how to defend against the opponent who uses disinformation intentionally. It appears to be a reverse to the primal problem considered in [1]. Beginning from Section 8 we make a next step toward reality and consider an approach which takes into account our limited access to real data. This forces us to abandon deterministic word and use probability theory instead. The new apparatus helps us to model information flows under several obstacles introduced intentionally or not. Limited access to information has several consequences in games of strategy and we analyse some of the most significant. Information transfer via public communication channels called mass media and other tubes of propaganda can be used for creation false conflict pictures, disinformation of public opinion, and in the consequence, forcing politicians to make decisions far from being optimal. In this connection several new mathematical problems are stated, some of them are very challenging and requires new concepts and ideas. To sketch possible approaches we analyse several related examples ranging from simple to advanced ones. In the last part we consider the games against Alcibiades. By Alcibiades we mean an exceptional intelligent and capable player whose skills are unlimited. Playing against Alcibiades is a great challenge requiring new methods and concepts. We adopt a notion of the Lagrange multipliers used in optimization theory for problems with constraints to model the intellectual constrains which an average player experiences having the bad luck by playing against Alcibiades.

1 Disinformation in non-matrix games

In [1] was considered the matrix games. In this paper we shall extend the game model to general case. Let P, Q are the sets of admissible strategies for D and O respectively, and a mapping $P \times Q \ni (p,q) \to A(p,q) \in \mathbb{R}$, is a payoff of the player D in the zero-sum game between D and O. We want to find the payoff $P \times Q \ni (p,q) \to B(p,q) \in \mathbb{R}$, in some admissible class $\mathfrak{B}$, such that

[1] Faculty of Economical and Technical Sciences, Pope John Paul II State School of Higher Education in Biała Podlaska.

$$\underbrace{\arg\max}_{B\in\mathfrak{B}} \boldsymbol{A}(p_B, q_B) = \boldsymbol{B}(p, q)$$

where p_B, q_B are defined by the inequalities

$$\underbrace{\max}_{p} \boldsymbol{B}(p, q_B) \leq \underbrace{\max}_{p} \boldsymbol{B}(p, q),$$
$$\underbrace{\max}_{p} \boldsymbol{A}(p, q_B) = \boldsymbol{A}(p_B, q_B).$$

Putting $\boldsymbol{A}(p, q) = \langle p, \boldsymbol{A}q\rangle$ and $B(p, q) = \langle p, \boldsymbol{B}q\rangle$, we turn the general model into the previous matrix form.

2 Disinformation in non-zero sum games

The conflict not always has the most acute form and the players' goals are not always exactly contradictory. There are conflicts in which players define the goals and scopes of the conflict differently and differently assess the win/loss values. It is then natural that the conflict is defined not by one, but by two matrices. A matrix $\boldsymbol{A}$ represents the wins of player D, and a matrix $\boldsymbol{B}$ the wins of player O, and $\boldsymbol{A} + \boldsymbol{B}$ is not a zero matrix. How can disinformation be built into such games? We will now answer this question, reasoning from the perspective of the disinformer D.

Nash games ([2]). Let us assume that the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the original game are known to D. Such games are called Nash games. Standard analysis leads to specifying a pair of Nash strategies (d, p) which are in equilibrium and each of them is the best against the other. There may be more than one such pair (non-unique solution). The further procedure of adding disinformation is analogous to the case of zero-sum games, the only difference being that we now modify both the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

3 Disinformation in n -person games

Let us denote by $L_j(s_1, \dots, s_n)$, $j = 1, \dots, n$ the payoff of the jth player, corresponding to strategies $s_1, \dots, s_n$ used by players $G_1, \dots, G_n$. Let us assume that the ith player, denoted by D, is the disinformer, and his strategy is denoted by d. We assume that he may misinform the other players using manipulation techniques which distort the picture of the conflict. As a result, for the other players, false payoffs $F_j(s_1, \dots, s_n)$ take place of the real ones $L_j(s_1, \dots, s_n)$.

To sum up, we assume that the players who are being misinformed know only the payoffs $F_j(s_1, \dots, s_n)$, $j = 1, \dots, n$, and the sets of their acceptable strategies S_j, $s_j \in \mathsf{S}_j$. Player D knows $L_j(s_1, \dots, s_n)$, for all $j = 1, \dots, n$. Then D faces the following problem: how to choose $F_j(s_1, \dots, s_n)$, depending on $L_j(s_1, \dots, s_n)$, $j = 1, \dots, n,\ j \neq i$, in order to achieve the Nash optimum.

$$\max\{L_i(S_1,\dots,d,\dots,S_n);\ d \in 6_i\} = L_i(S_1,\dots,S_i,\dots,S_n)$$

under the conditions

$$F_j(S_1,\dots,S_j,\dots,S_n) \geq F_j(S_1,\dots,s_j,\dots.,S_n),$$

where $S_j, s_j \in 6_j$ and $j = 1,\dots,n,\ j \neq i$.

The task still requires specifying in more details to what extent it is possible to "modify" the payoffs L_j into F_j. Disinformation procedures may be costly if conducted on many fronts and at a huge scale. It is obvious that the less F_j differs from L_j, the lower the cost of the disinformation action. The modification may thus be measured using the following index:

$$W_i(F_1,\dots,F_n) \stackrel{\text{def}}{=} \sum_{1\leq j\leq n}^{,j\neq i} (\#F_j)!\,\left|F_j(s_1,\dots,s_n) - L_j(s_1,\dots.,s_n)\right|$$

where $\#F_j$ is the number of non-zero modifications of L_j corresponding to each sequence of strategies $s_1,\dots,s_n$. Let $K(W)$ be the cost of the disinformation operation. We then search for the Nash optimum as above with the additional condition $K(W) \leq k$, where k is a specified financial limit.

The algorithm specified for two-person games should now be adjusted, starting from modifying the payoff of the player whose strategies have the greatest impact on the payoffs of player D.

4 Goals, strategies and the matrix of the game. Remarks on modelling

In the classical theory, the structure of each decision task (during a conflict or not) consists of the following, independent elements: the goal function (strategy assessment criterion) and the set of acceptable strategies. A decision task consists in choosing the strategy in such a manner that the goal function reaches, or approaches, the optimum. Thus, the goal function determines the decision task, and the set of acceptable strategies defines the player's possibilities. Both elements (the goal function and the set of acceptable strategies) are present in the game matrix. The set of strategies is explicitly represented by rows and columns, and the goal function is represented only implicitly, by assigning numerical values to the results of confronting particular strategy pairs. We formulate the goals of the game by entering sufficiently high payoff values into those matrix elements which correspond to our preferred results of confronting particular strategy pairs. We also discredit undesirable confrontation results by entering sufficiently low (negative) values into the corresponding matrix elements. Both kinds of elements are entered simultaneously, thus creating the game matrix.

The above algorithm clearly shows that the problem is so complex that it can be solved only by appropriate computer simulations conducted on a case-by-case basis. This has several reasons.

Firstly, such games feature so called mixed strategies. This means that, e.g., player O choosing columns has to "mix" them, i.e. pick various columns with different probabilities. Thus, his general strategy is a probability distribution on columns instead of choosing a single column. It turns out that it is only in such an extended set of strategies that an optimal strategy can be found. Consequently, this form of strategy should be expected for FGS as well.

Secondly, while the FGS can be calculated in a general manner using the known procedure, the analysis of the impact of disinformation on that strategy, resulting in the false matrix $\boldsymbol{B}$, is much more difficult.

Thirdly, examining the impact of that action on the outcome of the original game with matrix $\boldsymbol{A}$, is even more complex.

Due to lack of general theoretical results, it should thus be assumed that computer simulations will be the basic research method – unless such results will soon appear, which is unpredictable.

Of course, the feasibility of such a complex and ambitious analytical, mathematical and simulation task depends on our means and powers, but also on the complexity, scope and scale of the original game itself.

Applying the above methodology also enables one to obtain conclusions of a qualitative nature, sometimes even without a complicated simulation analysis. Below we present an example of such a reasoning inspired by the actions of the player mentioned at the beginning of the present study**.**

5 Disinformation and Russia

At least since the Bolshevik times, Russia has played political games tainted by disinformation against all countries. For that purpose, it uses an extremely extensive spy network located in political circles, intellectual elites, propaganda centres and mass media of the attacked countries. However, disinformation, as explained above, leads to creating a false image of goals – and this is always profitable only in zero-sum games, in which the win of one player means the same loss of the other. A zero-sum game describes the most acute conflict of all – players treat their opponents as full-blown enemies. As disinformation is guaranteed to be profitable only in that case, and Russia uses it always, regardless of its opponent – there can only be one conclusion: this player treats all the other players as its full-blown enemies!

6 Defence in games with disinformation

So far, we have been analysing the position of player D, who – having appropriate means and powers – successfully disinforms others. In this section, we will analyse the situation from the perspective of player O, who assumes that he

is being misinformed. That player is aware that the matrix $\boldsymbol{B}$, he is presented with differs from the matrix $\boldsymbol{A}$ of the original game, and that its elements have been purposely changed in order to lead him into the trap of the false game. Let us therefore follow the reasoning which has been applied so far by player D, in order to find an effective antidote.

Let us start from the obvious. As $\boldsymbol{B}$ is the matrix of the false game, O's optimal strategy for $\boldsymbol{B}$, and all nearly-optimal strategies, are immediately suspected of being a trap, and their corresponding payoffs are the consequence of the false image of the game. Also those elements of the matrix $\boldsymbol{B}$ which give very disadvantageous payoffs for O, and may therefore be aimed at discouraging O from using the corresponding strategies, are suspicious. Thus, it seems that the most reliable payoffs are those which are neither very encouraging nor very discouraging. Those payoffs should constitute a basis for redefining the game matrix using techniques applied by relevant services. If this is not possible, the optimum disinformation problem (ODP) from item 4a should be reversed as follows.

Let $\boldsymbol{B}(\boldsymbol{A}, \varepsilon)$ be the solution of the ODP, i.e. for a given matrix $\boldsymbol{A}$ and $\varepsilon > 0$, we search for

$$\max\{\langle p(q(\boldsymbol{B})), \boldsymbol{A}q(\boldsymbol{B})\rangle; W(\boldsymbol{B}-\boldsymbol{A}) \le \varepsilon\} \stackrel{\text{def}}{=} M < \infty$$

over all matrices $\boldsymbol{B}$ of the same dimension as $\boldsymbol{A}$, i.e.,

$$M = \langle p\left(q\big(\boldsymbol{B}(\boldsymbol{A},\varepsilon)\big)\right), \boldsymbol{A}q\big(\boldsymbol{B}(\boldsymbol{A},\varepsilon)\big)\rangle\,.$$

The problem of optimal identification of the matrix $\boldsymbol{A}$ (POI-A) is as follows: Given $\boldsymbol{B}$, we search for the solution $(\boldsymbol{A}, \varepsilon)$ of the equation

$$\boldsymbol{B}(\boldsymbol{A}, \varepsilon) = \boldsymbol{B}$$

or simply

$$\arg\left\{\underbrace{\max}_{B}\{\{\langle p(q(\boldsymbol{B})), \boldsymbol{A}q(\boldsymbol{B})\rangle; W(\boldsymbol{B}-\boldsymbol{A}) \le \varepsilon\}\}\right\} = B$$

Calculation aspects.

Even if ODP had a unique solution $\boldsymbol{B}(\boldsymbol{A}, \varepsilon)$, without the knowledge of ε, and without an explicit form of the index $W(\boldsymbol{A})$, one should not expect uniqueness of solution of POI-A. Additionally, the problem is very unstable, in the sense that small changes of parameters may result in significant changes of the solution. In that case, the only way to proceed is to use computer simulations, leading to approximate solutions differing from the optimal one by a tolerable error.

Example. In order to illustrate the problem and the difficulties in solving it, let us consider the example from Section 3 from [1], the only modification being that O is now aware that the matrix

$$\boldsymbol{B} = \begin{bmatrix} 1 & 0 & 3 \\ 1 & 2 & 6 \\ 1 & 1 & 3 \end{bmatrix}$$

is the matrix of the false game. In that game, the optimal strategy for the disinformation-unaware O would be to choose the first column. However, if O knows that $\boldsymbol{B}$ is the matrix of the false game, it is exactly the elements corresponding to that "optimal" strategy that are the most suspicious. Thus, the first column containing $b_{11} = b_{21} = b_{31}$=1 is in fact different, so some of its elements are different from 1. It is therefore reasonable to delete that column. After its removal, the game matrix is

$$\boldsymbol{C} = \begin{bmatrix} 0 & 3 \\ 2 & 6 \\ 1 & 3 \end{bmatrix}.$$

Clearly, choosing the second row is now the optimal strategy for D, guaranteeing the payoff $V = 2$ (and the same loss for O). It follows that the payoffs in the first column of $\boldsymbol{B}$ are falsified in such a way that the value of at least one of b_{11}, b_{21}, b_{31} is greater than 2, as only in that case does it make sense to set a trap resulting in a greater payoff. This conclusion is consistent with the real picture of the conflict expressed by the matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 3 \\ 7 & 2 & 6 \\ 1 & 1 & 3 \end{bmatrix}$$

in which $a_{21} = 7 > 2 = V > 1 = b_{21}$.

7 The identification problem for non-matrix games

On the level of generality adopted in section 6, the reconstruction problem for the matrix $\boldsymbol{A}$ having $\boldsymbol{B}$ as given turns to the following.

Consider a nonempty class $\mathfrak{B}$ of admissible mappings of the form $P \times Q \ni (p, q) \to A(p, q) \in \mathbb{R}$, and some $\mathbf{B} \in \mathfrak{B}$. For $\boldsymbol{A} \in \mathfrak{B}$, we define strategies $(p_B, q_B) \in P \times Q$, by the conditions

$$\underbrace{\max}_{p} \boldsymbol{B}(p, q_B) \leq \underbrace{\max}_{p} \boldsymbol{B}(p, q),$$
$$\underbrace{\max}_{p} \boldsymbol{A}(p, q_B) = \boldsymbol{A}(p_B, q_B),$$

and define

$$\boldsymbol{B}(\boldsymbol{A}) \stackrel{\text{def}}{=} \underbrace{\arg\max}_{B\in\mathfrak{B}} \boldsymbol{A}(p_B, q_B).$$

The identification problem: Find $\boldsymbol{A}$ satisfying the equation

$$\boldsymbol{B}(\boldsymbol{A}) = \boldsymbol{B}.$$

8 A stochastic view: alchemy of disinformation

In the model of disinformation which we have proposed, the matrix $\boldsymbol{B} = (b_{ij})$ of the false game is deterministic. The player O knows numerical values of the entries b_{ij} exactly. Certainly, this assumption is far from reality and was adopted for convenience in preliminary stages of the theory. In practice, an expert think tank working for the player O has to estimate the numerical values of b_{ij} by taking into account several events observed in the past and present. In the mathematical terms, these events form the so called sigma-sub-field on which conditional expectation operations should be performed. This suggests the following mathematical framework on which all mathematical models should be build. The elements b_{ij} are random variables defined on some probability space $(\boldsymbol{\Omega},\mathbb{P})$. There is some sigma-sub-field $\mathfrak{F}$ of $\mathbf{F}$,, formed by observations made by O. Then an estimator $\widehat{b_{ij}}$ of b_{ij} is computed and placed in the false game matrix. If the estimation has to be least squares optimal, then in the class of random variables having finite second moments, we have

$$\widehat{b_{ij}} = E\{\, b_{ij} \mid \mathfrak{F} \,\}$$

where $E\{\,\cdot \mid \mathfrak{F}\,\}$ is the conditional expectation operator. In this setting $\mathfrak{F}$ is the sigma-field of events which were suggested to O by D as describing the true conflict picture. Hence, the estimator $\widehat{b_{ij}}$ of b_{ij} computed from $\mathfrak{F}$ are the true elements of the false game matrix $\hat{B} = \left(\widehat{b_{ij}}\right)$.

Remark 1. It should be made clear that in the stochastic extension of the optimal disinformation we arrive to a new type of problems. First, when the sigma-sub-field $\mathfrak{F}$ is selected from **F**, in order to implement disinformation in this case, it is enough to use the appropriate events which took place in the past and publicize these outcomes via mass media and other tubes of propaganda. Doing that, we make some hypotheses about reality more probable then others. Bayes theorem applies and the false picture of the conflict is created. However, it may appear there are no such appropriate events in the past. In this case, the new and appropriate events should be created and the initial sigma field **F** should be extended to, say, $\mathfrak{B}$. In the both cases we have the complex interplay between highly abstract mathematical objects such as probability spaces, sigma-sub-fields of events, probability measures and random variables from the one side and the

disinformation actions taken in reality on the other. These actions can be viewed as generation of information flows produced in the real world and resulting from the appropriate sigma-sub-fields $\mathfrak{B}_t, t\epsilon[0,T], T>0$, of events taken from the model. $\mathfrak{B}_t$ is called a filtration when $\mathfrak{B}_s \sqsubset \mathfrak{B}_t$, for $s<t$. Set $\mathfrak{P}=\mathfrak{B}_T$, and note that there are many different filtrations $\mathfrak{B}_t$ called a 'bridge' such that $\mathfrak{B}_T=\mathfrak{P}$.

9 The stochastic disinformation problem (SDP)

To state this problem precisely, let the random variable b_{ij} taking values in a finite interval $[a,b]$ be defined on ($\mathbf{\Omega}$,,$\mathbb{P}$).

For any $\varepsilon>0$, and a real number $\beta_{ij}\,\epsilon[a,b]$ find:

1) a filtration $\mathfrak{B}_t, t\epsilon[0,T], T>0$, where $\mathfrak{B}_0$ is trivial;

2) a family of probability measures $\mathbb{P}_t$ on $\mathfrak{B}_t$, $t\epsilon[0,T]$ such that

$$\mathbb{P}_T\left(\left|\, E_T\left\{\, b_{ij}\,\middle|\,\mathfrak{B}_T\right\}-\,\beta_{ij}\,\right|\leq\,\varepsilon\right)\geq 1-\,\varepsilon$$

where the expectation E_T is taken with respect to the measure $\mathbb{P}_T$.

Having the information flow $\mathfrak{B}_t, t\epsilon[0,T], T>0$ and treating the probability measures $\mathbb{P}_t$ on $\mathfrak{B}_t$, $t\epsilon[0,T]$ as the objective probabilities, the opponent is forced to accept β_{ij} as the approximation of b_{ij} close to the best.

Moreover, each filtration corresponds to a sequence of events in the real world and only some of them can be arranged in reality. Since the arrangements are costly we arrive in this way to a new type of control optimization problems. In the classical control problems a state of a system is a point in some finite (or not finite) dimensional state space and the goal is to minimize a cost of transferring the system form an initial to final state. Here we have the 'bridge' problem of transferring an initial sigma-sub-field $\mathfrak{B}_0$ to the final $\mathfrak{P}$ along the trajectory $\mathfrak{B}_t, t\epsilon[0,T]$, which has to be realizable in practice and cost-minimal.

In order to find $\mathfrak{B}_T$ and $\mathbb{P}_T$ the following observations may be helpful. For fixed $\varepsilon>0$, denote

$$\mathfrak{F}_{\,b_{ij},\varepsilon}=\sigma\left\{\left|\;b_{ij}-\;\beta_{ij}\,\right|\leq\;\varepsilon\right\}$$

a sigma-sub-field of **F**. Next step is to assign high probabilities to the events in $\mathfrak{F}_{\,b_{ij},\varepsilon}$ and small outside $\mathfrak{F}_{\,b_{ij},\varepsilon}$.

Example. In order to see what one can expect let us consider an simplified version of the problem. Take one random variable b instead of the collection b_{ij} and a parameter t equal 0 or 1 instead $t\epsilon[0,T]$. The last simplification changes information flows into two-steps procedure. Assume

$$\{b(\omega);\;\omega\epsilon\Omega\}=(p,q)\text{ and }\beta\,\epsilon\,(p,q).$$

Next, define a set

$$A_{\,\beta,\varepsilon}=\{\omega\,\epsilon\,\Omega;|\;b(\omega)-\;\beta\,|\leq\;\varepsilon\}$$

and put

$$Q(A) = \frac{1-\varepsilon}{\mathbb{P}(A_{\beta,\varepsilon})}\,\mathbb{P}\big(A \cap A_{\beta,\varepsilon}\big) + \frac{\varepsilon}{1-\mathbb{P}(A_{\beta,\varepsilon})}\,\mathbb{P}\big(A \cap \overline{A_{\beta,\varepsilon}}\big)$$

for $A \in \mathbf{F}$. It is easy to check that Q is a probability measure on $\mathbf{F}$. Moreover,

$$E_Q\{b(\omega) \,|\, \mathbf{F}|A_{\beta,\varepsilon}\} = \beta + E_Q\{b(\omega) - \beta \,|\, \mathbf{F}|A_{\beta,\varepsilon}\} = \beta + b_{\varepsilon}$$

where $\mathbf{F}|A_{\beta,\varepsilon}$ is a restriction of $\mathbf{F}$ on $A_{\beta,\varepsilon}$, $b_{\varepsilon} = E_Q\{b(\omega) - \beta \,|\, \mathbf{F}|A_{\beta,\varepsilon}\}$ and

$$|b_{\varepsilon}| \le E_Q\{|b(\omega) - \beta| \,|\, \mathbf{F}|A_{\beta,\varepsilon}\} \le \varepsilon(1-\varepsilon),$$

since $Q(A_{\beta,\varepsilon}) = 1-\varepsilon$.

The assumptions above guarantee that $A_{\beta,\varepsilon}$ is nonempty and $\mathbf{F}|A_{\beta,\varepsilon}$ is a nontrivial sigma-field for any $\beta \in (p,q)$.

In conclusion we have shown that by reaping the events from $\mathbf{F}|A_{\beta,\varepsilon}$ via the tubes of propaganda according to the new measure Q one may convince the opponent that β is a good estimator of $b(\omega)$ with error not bigger than ε.

Example. In the more general case we have a random variable $\{b(\omega);\ \omega\in\Omega\} \subset (p,q)$ on $(\mathbf{\Omega},, \mathbb{P})$, a number $\beta \in (a,b)$ and want to define a new measure Q on $\mathbf{F}$ such that

$$E_Q\{b_{\varepsilon}(\omega) \,|\, \mathbf{F}|A_{\beta,\varepsilon}\}$$

is close to β, for some b_{ε} close to b. The point is that (p,q), (a,b) are not related to each other. Denote F_{b} distribution of b under $\mathbb{P}$, and f_{b} its density with respect to the Lebesgue measure (assumed to exists). Let

$$f_{\varepsilon}(x) = (1-\varepsilon)f_{\mathrm{b}}(x) + \frac{\varepsilon}{|R|}\,\mathbb{1}_R(x)$$

where $S = (a,b) \cup supp\, f_{\mathrm{b}}(x), R = S \cap (a,b), \mathbb{1}_R$ is an indicator function and $supp\, f_{\mathrm{b}}(x) = \{x\in (p,q);\ f_{\mathrm{b}}(x) > 0\}$. Let b_{ε} be a random variable having density f_{ε}. Define a measure $\boldsymbol{u}$ on $\mathbb{R}$ by the formula

$$u(\mathrm{P}) = \int_P f_{\varepsilon}(x)dx$$

and a new measure Q on ($\mathbf{\Omega}$, $\mathbf{F}$) by the formula

$$Q(A) = u(x;\ b_{\varepsilon}^{-1}(x)\in A)$$

Then

$$E_Q\{b_{\varepsilon}(\omega) \,|\, \mathbf{F}|A_{\beta,\varepsilon}\} = \beta + E_Q\{b_{\varepsilon}(\omega) - \beta \,|\, \mathbf{F}|A_{\beta,\varepsilon}\}$$

where

$$A_{\beta,\varepsilon} = \{\omega \in \Omega; |\ b_\varepsilon(\omega) - \beta| \leq \varepsilon\}.$$

Since

$$Q(A_{\beta,\varepsilon}) = u(x;\ b_\varepsilon^{-1}(x) \epsilon A_{\beta,\varepsilon}) = \int_{\beta-\varepsilon}^{\beta+\varepsilon} \left[(1-\varepsilon) f_b(x) + \frac{\varepsilon}{|R|} \mathbb{1}_R(x)\right] dx$$
$$\leq 2(1-\varepsilon)\,\varepsilon \max f_b(x) + \frac{2\varepsilon^2}{|R|},$$

then

$$E_Q\,\{|b_\varepsilon(\omega) - \beta|; A_{\beta,\varepsilon}\} \leq \varepsilon Q(A_{\beta,\varepsilon}) \leq \varepsilon\, o(\varepsilon).$$

Remark. As the above examples clearly indicate, the 'false conflict picture', a main goal of disinformation actions, consists in creation of a new probability measure Q which differs locally from the objective measure $\mathbb{P}$. Q assigns higher probability than $\mathbb{P}$ does to some events and lowest to others making some hypothesis more probable, or convincing then other. More about this point in the next examples is given.

10 The 'right proportion' optimization problem (RPOP)

In order for disinformation to be credible and convicting it cannot be bothersome or naïve. It cannot be false ostentatiously too, i.e., it should be a mixture of a truth and lies in a right proportion. This proportion can be a subject of the RPOP, or a family of problems when many criteria are taken into account.

The question how to mix 'a truth and lies' in a right proportion appears in many hazard games the most famous being poker. In this game the mix problem takes the form of bluffing. It is instructive to recall the famous version of asymmetric poker analysed in Section 19.14 of [1]. The optimal strategy for Player 1 is to 'bluff' by making a higher bid, say α, despite having a weak hand. The right proportion between 'bluffing' and 'non-bluffing', i.e., making a lower bid, say β is given in the formula

$$u = \frac{\gamma(1-\gamma)}{1+3\gamma}, \text{ where } \gamma = \beta/\alpha \text{ and } 0 < \beta < \alpha.$$

This may be explained as follows. Let us call a hand in the interval $[0, u]$ a weak hand. Having a weak hand Player 1 is advised to make a higher bid suggesting opponents he is having a strong hand. Now, if the numbers in $[0,1]$are chosen randomly with equal probability, then u is a probability of having a weak hand – and bluffing. Hence $u = \gamma(1-\gamma)/(1+3\gamma)$ give us the answer how to mix 'false and true' playing the version of poker considered in [1]. It is interesting to note that the maximal value of $u, u_{max} = 0.10511$, is achieved when

$\gamma = 0.23241$. Certainly, $\gamma(1-\gamma)/(1+3\gamma)$ is a right value for u mixing 'a truth and lies in a right proportion', but only in the specific version of poker considered in [1]. To find a general answer we have to consider a general model.

Example. This example is a version of the famous experiment by Amos Tversky and Daniel Kahneman on inductive reasoning and posterior probabilities described in their *Prospect Theory*.

Consider two alternatives called hypothesis H_1 and H_2 having a priori probabilities p_1 and p_2, $p_1 + p_2 = 1$. Let

$$\mathbb{P}\{\,\mathrm{Q}|H_i\,\} = q_i\,, \mathrm{i} = 1{,}2$$

be a conditional probability of an event Q under the hypothesis H_i. How often has Q to be repeated by the propaganda tube in mass media to convince opponent's public opinion that, say, H_1 is more than 95% sure?

From Bayes rule

$$\mathbb{P}\{\,H_1|\mathrm{Q}\,\} = p_1\,(q_1)^n/[p_1\,(q_1)^n + \,p_2\,(q_2)^n] = 1/[1 + \,a\,(b)^n] \geq 0.95$$

where $a = \,p_2/p_1$ and $b = \,q_2/q_1$. It follows that

$$n \geq \,(\ln 19\,a)/(\ln b^{-1}).$$

provided $0 < b < 1$. When $p_1 = 0.05,\ a = 19,\ q_1 = 0.8, b = 0.25$, then for $\mathbb{P}\{\,H_1|\mathrm{Q}\,\}] \geq 0.95$ it is enough to have $n \geq 5$. In the words: in order to shift public opinion belief concerning the hypothesis H_1 from 5 to 95 percent it is enough to repeat in the row Q in mass media five times only (!) provided $q_1/q_2 = 4$.

11 Disinformation in complex games

In order to provoke opponents to make a false move in the main game the player S, we call him the Strategist, can arrange an artificial game, or games. We call this combination of games the 'complex game'. There are many variants of the complex games depending on the number of players involved in the combination, initial knowledge they possess, etc. Consider first the simplest one. Let D and P play the game on the matrixes $\boldsymbol{A}$ and $\boldsymbol{B}$ as it was described in preceding paragraphs. However, D does not know that he is playing with a third player S a game with matrix $\boldsymbol{C}$. Moreover, S was clever enough to arrange the game between D and P, because he wanted D to make a move which is bad in the game D vs C. However, it may appears that there is another player, an super Strategist (SS) playing the game with S, or (S and D), or (S and D and P) who has arranged the whole combination. In this way we arrive to the hierarchy of complex games and Strategists, a theme which was mentioned early in the Introduction. The hierarchy can be described easily in the von Neumann scheme. For instance, the payoffs

$$L_P(s_1, s_2),\ L_D(s_2, s_2),\ L_S(s_2, s_3),\ L_{SS}(s_3, s_4),$$
$$L_P(s_1, s_2),\ L_D(s_1, s_2),\ L_S(s_2, s_3),\ L_{SS}(s_2, s_3, s_4),$$
$$L_P(s_1, s_2),\ L_D(s_1, s_2),\ L_S(s_2, s_3),\ L_{SS}(s_1, s_2, s_3, s_4),$$

fully describe the game structures mentioned above. Although, each row strongly indicates the dominant role of the last player, then looking in the averse direction, i.e., from the row to the structure, there is only a hint that S has arranged the game P vs D and SS has arranged S vs D in the first row. S arranged the games (P vs D) and SS arranged (D vs S) in the second row. In the third row SS has arranged all games.

Despite the problem of one−to−one relations between the structures and the rows, there is the fundamental question of how the structure (or the row) has to be identified in practice by players or observers. Since this work is addressed to the active players rather than passive, and to the players rather than observers, we are mainly interested in the methodology telling how to create the complex games, and less about the identification problems for itself.

What differs political games from non−political is that these games are coupled, i.e., they cannot be divided into a set of two person games analysed separately. Making a political move in the game I vs J, players I and J give signals to other players as well. Moreover, these signals have to be consistent with the previous and with players' politics as a whole. In the consequence, the next moves of I and J have to take these constrains into account. Considered as possible and even reasonable in separate games, say I vs J, some strategies are contradictory in the games I vs K, J vs K, since they generate not coherent and inconsistent political signals. Shortly, the s_i – strategy of player I applies in every game played by I and therefore it must appear in the corresponding payoffs**.**

12 Playing with Alcibiades

In this section we refer to this point in the Introduction where the question "but what if we had a misfortune to confront with Alcibiades, Talleyrand or Julius Caesar?" was posed. By the 'game with Alcibiades' we mean a game with highly intelligent and capable player, whose ability to find strategies and apply them is superior. The game is asymmetric since Alcibiades knows how to implement these strategies contrary to his opponent who does not even if these strategies can be guessed. At a first glance, the opponent's position is hopeless. How to play with a master who, in addition, is informationally privileged? Where could one find his Achilles' heel?

Before we shall begin studying how to play with Alcibiades, let us start with much simpler question – how high is a potential cost of risking the game with Alcibiades? We focus here on the following dominant factor – intellectual superiority of Alcibiades over his opponent, say O.

12.1 Zero-sum games

Example 1. To model this game let us consider a large payoff matrix $\mathcal{A}$ (possibly infinite) and a smaller submatrix A of the bigger $\mathcal{A}$. Assume the players know $\mathcal{A}$ and A. Despite of knowing $\mathcal{A}$, O is not able to guess in details the form of corresponding strategies (i,j) unless the result a_{ij} is an element of A. For instance, building a bridge was a great challenge in ancient times and only Caesar's legions were capable for making this effort in a short time. Hence, the strategies which use such advanced engineering capability were unthinkable and not feasible for his opponents, nevertheless the results of these strategies (if applied) could be easily estimated and the corresponding payoff a_{ij} were known. Thus we assume, Alcibiades can manage in choosing any column of $\mathcal{A}$, whereas only the rows of A are allowed for his opponent O to play. If the pair (i,j) was selected the element a_{ij} of $\mathcal{A}$ has to be paid by Alcibiades to his opponent. Following an idea of R.J.-B. Wets (see [3] and the references herein) we apply the Lagrange multiplier to incorporate these intellectual constraints into the model.

Applying the well-known analogy between games and linear programming (see [2]) we arrive to the problem

$$\max\{\langle J-\mu,U\rangle+\langle J_m-\lambda,V\rangle\,;AU\le J_n,A^cV\le J_r\}\triangleq M_1+M_2$$

where $Q=col(U,V)$ is unnormalized probability distribution vector on columns of $\mathcal{A}$, A^c is a matrix obtained from $\mathcal{A}$ after elimination all rows and columns of A. J_m J_n, J_r, and J are all vectors of the form $J=col(1,1,\dots.)$ of appropriate dimensions and $\Lambda=col(\mu,\lambda)$ is a Lagrange multiplier. Since a normalized probability distribution $q=col(u,v)$, where

$$u=\frac{U}{\langle J,U\rangle+\langle J_m,V\rangle}$$
$$v=\frac{V}{\langle J,U\rangle+\langle J_m,V\rangle}$$

then we should have

$$\Lambda(u,0)=\langle\mu,u\rangle=0,$$

what implies

$$\langle\mu,U\rangle=0.$$

Note that maximization over $Q=col(U,V)$ is equivalent to maximization over U and V separately. Let denote U_* and V_* solutions (assumed to exist) corresponding to M_1 and M_2 respectively. If the Lagrange multiplier is chosen properly, we have $M_2=0$ and

$$\langle J_m-\lambda,V_*\rangle=0.$$

Solving for λ with minimal (Euclidean) norm we obtain

$$\lambda_*(V) = \langle J_m, \frac{V_*}{\|V_*\|}\rangle \langle \frac{V_*}{\|V_*\|}, V\rangle$$

hence

$$\Lambda_*(u,0) = \mu_*(u) = 0 \text{ implies } \mu_* = 0,$$

and

$$\Lambda_*(0,v) = \lambda_*(v) = [\langle J, U_*\rangle + \langle J_m, V_*\rangle]\lambda_*(V)$$

is the incremental cost of violating intellectual constrains (here U_* is the solution corresponding for M_1).

Example 2. Now we want to find the Lagrange multiplier Λ connected with the opponent strategies, i.e., with the rows of $\mathcal{A}$. The corresponding linear programming problem has a form, find

$$\min\{\langle J + \mu, W\rangle + \langle J_m + \lambda, H\rangle; A^T W \geq J_n, (A^c)^T H \geq J_r\} \triangleq N_1 + N_2$$

where P= $col(W,H)$ is unnormalized probability distribution vector on the rows of $\mathcal{A}$. Again, we argue that although all rows are known to the player O, but only those belonging to A can be viewed as strategies. Hence, not all but only those belonging to A are allowed to play. Minimization over P is equivalent to minimization over W and H separately. Let denote W_* and H_* solutions (assumed to exist) corresponding to N_1, N_2 respectively. If the Lagrange multiplier Λ is chosen properly we have $N_2 = 0$ and

$$\langle J_m + \lambda, H_*\rangle = 0.$$

Solving for λ with minimal norm we obtain

$$\lambda_*(H) = -\langle J_m, \frac{H_*}{\|H_*\|}\rangle \langle \frac{H_*}{\|H_*\|}, H\rangle$$

thus

$$\mu_* = 0$$

and

$$\Lambda_*(0,h) = \lambda_*(h) = -[\langle J, W_*\rangle + \langle J_m, H_*\rangle]\lambda_*(H)$$

is the incremental cost of violating intellectual constrains specified above.

Remark. Piecing together the above results one can conclude how to estimate a cost of violating these two constrains appearing jointly.

Remark. Even though we still do not know how to play with Alcibiades we do know the game is worth of playing or not.

12.2 Non-zero-sum games

In order to explain how the concept of intellectual constrains and its pricing can be used in the practice of political and military games let's consider the famous Kennedy vs Khrushchev game during the Cuban Missile Crisis in 1962. We shall focus on particular question: how costly were the intellectual constrains of Khrushchev who underestimated JFK. The analysis should be done from Khrushchev perspective available to him at 1962 before the game begun.

The first ascertaining is that only God knows everything and therefore cannot be cheated, not humans. Even high intelligence cannot create new information but only can use the amount available. Therefore the Achilles heel of the high intelligence are data. Moreover, Alcibiades seems to be more 'sensitive' with respect to data changes then any 'ordinary' player. Indeed, Alcibiades has enormous ability to create unexpected (unpredictable) strategies and this is perhaps his main advantage. Since, strategies are mappings from information into actions, the number n of possible strategies is

$$\mathrm{n} = card(A^{\mathbf{i}})$$

where $A^{\mathbf{i}}$ denote the set of all non-anticipative mappings $\mathbf{i} \to A$, where $\mathbf{i}$ is a set of all available information and A is a set of all possible actions. It follows that increasing (decreasing) information, or actions one increases (decrease) the set of Alcibiades strategies.

It is instructive to compare two numbers: 100^{10} and 10^{100}, where the first is a number of strategies when $card\ (A) = 100$ and $card(\mathbf{i}) = 10$, whereas in the second is opposite; $card(A) = 10$ and $card(\mathbf{i}) = 100$. Since $100^{10} = 10^{20} < < 10^{100}$, we see that increasing card ($\mathbf{i}$) ten times gives 10^{80} times more strategies then increasing card (A) ten times. The comparison partially explains the role played by intelligence agencies in modern political and military games.

By preparing 'new information' one may provoke him to made a false move much worse (for him) comparing with the case when game is played with an ordinary player who is less 'sensitive' with respect to information changes. However, this 'new information' usually requires a new communication channel hence we shall continue this theme in the next paper**.**

Bibliography

[1] von Neumann J., Morgenstern O., *Theory of Games and Economic Behavior*, Princeton 1947.

[2] Luce R. D., Raiffa H., *Games and Decisions*, John Wiley & Sons, Inc., New York 1958.

[3] Wets R. J.-B., *On the Relation between Stochastic and Deterministic Optimization,* in*: Control Theory, Numerical Methods and Computer Systems Modelling*, eds. Bensoussan A. and Lions J.L., "Lectures Notes in Economics and Mathematical Systems", 107, Springer-Verlag, Berlin 1975, 350–361.

[4] Davis M.H.A., Dempster M.A.H., Elliott R.J., *On the Value of Information in Controlled Diffusions Processes*, Liber Amicorum for M. Zakai, 125–138.

[5] Banek T.*, Disinformation – advanced weapon in political and military games: basic ideas – zero sum matrix game*, [in:] *Probability in Action*, Lublin University of Technology, 2017, 9–20.

Tadeusz Banek[1]

Notes on risk minimization

Keywords: risk, risk analysis, risk minimization.

Abstract

Risk is inherently connected with decisions hence risk analysis should be an significant component of decision making processes. The essential part of the risk analysis is estimation and subsequently minimization of risk. This note offers a general approach to modelling, estimation and minimization of risk by using analytical and simulation methods.

1 Introduction

Human activities are risky. Unpredictable changes in dynamic neighbourhood, disturbances, noises, climate changes, hazards events − all are the risk sources. They can cause some deviations from a planned scenario and the aim of **risk analysis** is to identify all possible events which can lead up to this negative aftermath which imply losses. Sometimes it is possible and useful to introduce a quantitative characterization of possible losses as a function of these quantities modelled here by a random variable X and as a function of possible actions u which have to be chosen in order to minimize its effects. Localization of risk sources and estimation of hazards elements and/or events together with their probabilities are the subjects of **risk estimation**. These results are used next for the purpose of **risk minimization**. How it is done constitute the main theme of these notes.

The problem of risk minimization can be seen in a broader sense. For this purpose, let us consider the decision problem of profit maximization with uncertainty. Let income I depend on chosen actions $u = col(u_1, \dots, u_p)$, i.e., $I = I(u)$, and all possible losses are in the form $L = L(u, X(\omega))$, where $X(\omega)$ denotes a random element. It is natural for decision maker (DM), to pose

[1] Faculty of Economical and Technical Sciences, Pope John Paul II State School of Higher Education in Biała Podlaska

Problem 1. Given a set of available decisions U, a profit function $I(u) - L\big(u, X(\omega)\big)$, find

$$\sup_{u\in U}\mathbb{P}\big(\omega; I(u) - L\big(u, X(\omega)\big) > z\big),$$

where z is a highest level of not acceptable profits. Because

$$\mathbb{P}\big(\omega; I(u) - L\big(u, X(\omega)\big) > z\big) = \mathbb{P}\big(\omega; L\big(u, X(\omega)\big) < I(u) - z\big),$$

the problem is equivalent to

$$\sup_{u\in U}\mathbb{P}\big(\omega; L\big(u, X(\omega)\big) < \alpha\big),$$

where $\alpha = I(u) - z$, which is our main model for the risk minimization.

More generally, if a formula for the profit function $Z\big(u, X(\omega)\big)$ is known, then it is more natural to set another problem.

Problem 2. Given $Z\big(u, X(\omega)\big)$, find

$$\sup_{u\in U}\mathbb{P}\big(\omega; Z\big(u, X(\omega)\big) > z\big).$$

As these examples indicate, risk minimization can be a part of more general optimization problems.

2 Statement of the problem

Definition 3. We call $L\colon \mathbb{R}^p \times \mathbb{R}^n \supset U \times D \ni (u, x) \to L(u, x) \in \mathbb{R}_+ \equiv [0, \infty)$, where U, D are non empty open sets, a **loss function**, iff

$$L(u, tx) \geq L(u, sx)$$

for $u \in \mathbb{R}^p$, $x \in \mathbb{R}^n$ and $t \geq s \geq 0$.

Problem 4. Given $\alpha \geq 0$, a loss function L, and $X\colon \Omega \to \mathbb{R}^n$, a vector valued random variable defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$, find

$$\sup_{u\in U}\mathbb{P}\big(\omega; L\big(u, X(\omega)\big) < \alpha\big).$$

Denote by F_X the distribution function of X, and by $q_X(\xi)$ its density (if exists). Then

$$\mathbb{P}\big(\omega; L\big(u, X(\omega)\big) < \alpha\big) = \begin{cases} \displaystyle\int_A F_X(d\xi), & \text{or} \\ \displaystyle\int_A q_X(\xi)d\xi, & \end{cases}$$

where $A = \{\mathbb{R}^n \ni \xi; L(u, \xi) < \alpha\}$.

Example 5. When $n = p = 1$, $D = \{\xi > 0\}$, $X \geq 0$, $c(u) > 0$, *for all* $u \in U$, $L(u,x) = c(u)\xi^2$, then

$$\mathbb{P}(\omega; c(u)X^2(\omega) < \alpha) = \mathbb{P}\left(\omega; X(\omega) < \sqrt{\alpha/c(u)}\right) = F_X\left(\sqrt{\alpha/c(u)}\right),$$

hence

$$\sup_{u\in U}\mathbb{P}(\omega; c(u)X^2(\omega) < \alpha) =$$
$$\sup_{u\in U}F_X\left(\sqrt{\alpha/c(u)}\right) = F_X\left(\sup_{u\in U}\sqrt{\alpha/c(u)}\right), \text{by monotonicity of } F_X$$

However, for the loss functions which are generally not invertible we have to work harder.

Definition 6. We call a pair $\big(\alpha, P(\alpha)\big)$, where $P(\alpha) = \sup_{u\in U}\mathbb{P}(\omega; L\big(u, X(\omega)\big) < \alpha)$, a **risk graph** of (L, F_X).

Example 7. Working devices are subjected to breakdowns. The pause needs T working hours for reparations which are costly. Let KT is a cost of reparation which takes T working hours, where K is the reparation cost per hour. Let $C(T)$ represents a cost of $T-$ hours pause of device's work. Then the total cost of the breakdown is $KT + C(T)$. Denote $X = col(X_1, X_2) = col(T, K)$. If $u-$ workers are employed, then the financial loss function is

$$L(u, X) = X_1X_2 + C(X_1/u),$$

where $C(0) = 0$, $0 \leq s \leq t$, implies $C(sx) \leq C(tx)$, for $x \in \mathbb{R}$, if $L(u,X)$ has to fulfill the formal requirements. This formula reflects a joint effect of randomness represented by X and actions chosen represented here by u. If F_{TK} is a joint distribution function of the random variable $col(T,K)$, then

$$\mathbb{P}\big(\omega; L\big(u, X(\omega)\big) < \alpha\big) = \int_A F_{TK}(d\xi),$$

where

$$A = \{\mathbb{R}^2_+ \ni \xi; \xi_1\xi_2 + C(\xi_1/u) < \alpha\}.$$

3 Development

3.1 Approach via smooth transformations

Let $B(0,r)$ be a centered ball in $\mathbb{R}^n$ of radius r, $f\colon \mathbb{R}^p \times \mathbb{R}^n \to \mathbb{R}^n$, be a smooth mapping such that

$$f\big(u, B(0,r)\big) = A = \{\xi; L(u,\xi) < \alpha\},$$

for some $r = r(\alpha, u)$ dependent on α and u.

Example 8. Let

$$L(u,x) = x^T Q(u)x,$$

$Q = \left(q_{ij}\right) \in M(n,n)$, where $M(n,n)$ is a set of quadratic $n \times n$ matrices. We allow Q to depend on u, i.e., $Q = Q(u) = \left(q_{ij}(u)\right)$. Moreover, if $Q = \left(q_{ij}\right) \in M_s^+(n,n) -$ set of quadratic $n \times n$ symmetric, positive defined matrices, then

$$f(u,x) = \frac{a}{r} Q^{-1/2}(u)x,$$

satisfies

$$\left\|Q^{1/2}(u)f(u,x)\right\|_{\mathbb{R}^n} = \frac{a}{r}\|x\|_{\mathbb{R}^n} \le a,$$

if $\|x\|_{\mathbb{R}^n} \le r$. Thus

$$f\left(u, B(0,r)\right) = \frac{a}{r} Q^{-1/2}(u)B(0,r) = A,$$

for any nonnegative r, a. Note, that

$$\frac{\partial}{\partial \varepsilon}[f(u+\varepsilon v, x)]_{\varepsilon=0} = -\frac{a}{2r} Q^{-3/2}(u)Q_v'(u)x,$$

where

$$\left(Q_v'(u)\right)_{ij} = \langle \nabla q_{ij}(u), v\rangle,$$

and

$$Jf(u,\eta) = \left|\det \frac{\partial f(u,\eta)}{\partial \eta}\right| = \frac{a}{r}\left|\det Q^{-1/2}(u)\right| = \frac{a}{r}|\det Q(u)|^{-1/2}.$$

Thus

$$\frac{\partial}{\partial \varepsilon}[Jf(u+\varepsilon v,\eta)]_{\varepsilon=0} = -\frac{a}{2r}|\det Q(u)|^{-3/2}\sum_{i=1}^{n} \det Q_i(u,v),$$

where $Q_i(u,v)$ is a matrix obtained from $Q(u)$ by $\varepsilon -$ differentiation of its i^{th} column.

Now turning back to the general case, we have

$$\begin{aligned}
&\int_A q_X(\xi)d\xi = \int_{\mathbb{R}^n} \P_A(\xi)q_X(\xi)d\xi \\
&= \int \P_{B(0,r)}(\eta)q_X\left(f(u,\eta)\right)Jf(u,\eta)d\eta \\
&= \int_0^\infty \left(\int_{\mathbb{R}^{n-1}} \P_{S^{n-1}(\tau)\cap B(0,r)}(\eta)q_X\left(f(u,\eta)\right)Jf(u,\eta)d\mathcal{H}^{n-1}(\eta)\right)d\tau \\
&= \int_0^{r(u)} \left(\int_{\mathbb{R}^{n-1}} \P_{S^{n-1}(\tau)}(\eta)q_X\left(f(u,\eta)\right)Jf(u,\eta)d\mathcal{H}^{n-1}(\eta)\right)d\tau \equiv J(u),
\end{aligned}$$

where in the line two we have changed variables putting $\xi = f(u,\eta)$, and in the next lines integration is performed with respect the Hausdorff measure $\mathcal{H}^{n-1}$, and $S^{n-1}(\tau) = \{\mathbb{R}^n \ni \xi; \|\xi\|_{\mathbb{R}^n} = \tau\}$ is a centered sphere of radius τ, and Jf is a Jacobian of f (see [3]) Hence, the necessary condition for $u^{\text{å}}$ to maximize $\mathbb{P}\big(\omega; L\big(u, X(\omega)\big) < \alpha\big)$, is

$$\frac{\partial}{\partial\varepsilon}[J(u+\varepsilon v)]_{\varepsilon=0} = 0.$$

Theorem 9.

$$\begin{aligned}\frac{\partial}{\partial\varepsilon}[J(u+\varepsilon v)]_{\varepsilon=0} &= \langle \nabla r\big(u^{\text{å}}\big), v\rangle \int_{\mathbb{R}^{n-1}} \P_{S^{n-1}\left(r\left(u^{\text{å}}\right)\right)}(\eta) q_X\left(f\big(u^{\text{å}},\eta\big)\right) Jf\big(u^{\text{å}},\eta\big) d\mathcal{H}^{n-1}(\eta)\\ &+ \int_0^{r\left(u^{\text{å}}\right)} \left(\int_{\mathbb{R}^{n-1}} \P_{S^{n-1}(\tau)}(\eta) \left\langle \nabla_u q_X\left(f\big(u^{\text{å}},\eta\big)\right), v\right\rangle Jf\big(u^{\text{å}},\eta\big) d\mathcal{H}^{n-1}(\eta)\right) d\tau\\ &+ \int_0^{r\left(u^{\text{å}}\right)} \left(\int_{\mathbb{R}^{n-1}} \P_{S^{n-1}(\tau)}(\eta) \langle \nabla_u Jf\big(u^{\text{å}},\eta\big), v\rangle q_X\left(f\big(u^{\text{å}},\eta\big)\right) d\mathcal{H}^{n-1}(\eta)\right) d\tau\end{aligned}$$

Proof. Differentiation

$$\frac{\partial}{\partial\varepsilon}\left[\int_0^{r\left(u^{\text{å}}+\varepsilon v\right)} \left(\int_{\mathbb{R}^{n-1}} \P_{S^{n-1}(\tau)}(\eta) q_X\left(f\big(u^{\text{å}}+\varepsilon v,\eta\big)\right) Jf\big(u^{\text{å}}+\varepsilon v,\eta\big) d\mathcal{H}^{n-1}(\eta)\right) d\tau\right]_{\varepsilon=0}$$

gives the result.

Example 10. (continued) Put $r(u) = \alpha$. Substitution

$$\begin{aligned}&\frac{\partial}{\partial\varepsilon}\left[q_X\left(f\big(u^{\text{å}}+\varepsilon v,\eta\big)\right) Jf\big(u^{\text{å}}+\varepsilon v,\eta\big)\right]_{\varepsilon=0}\\ &= -\frac{1}{2}\left[q_X'\left(f\big(u^{\text{å}},\eta\big)\right) Q^{-3/2}(u) Q_v'(u)\eta Jf\big(u^{\text{å}},\eta\big)\right.\\ &\left.+ q_X\left(f\big(u^{\text{å}},\eta\big)\right) |\det Q(u)|^{-3/2} \sum_{i=1}^n \det Q_i(u,v)\right] u^{\text{å}}\end{aligned}$$

into

$$\int_0^{\alpha} \left(\int_{\mathbb{R}^{n-1}} \P_{S^{n-1}(\tau)}(\eta) \left\langle \nabla_u q_X\left(f\big(u^{\text{å}},\eta\big)\right) Jf\big(u^{\text{å}},\eta\big), v\right\rangle q_X\left(f\big(u^{\text{å}},\eta\big)\right) Jf\big(u^{\text{å}},\eta\big) d\mathcal{H}^{n-1}(\eta)\right) d\tau,$$

gives the result.

3.2 Standard approach

Let's define a new random variable $Y = col(Y_1, \dots Y_n)$, such that $Y_1 = X_1$, ..., $Y_{n-1} = X_{n-1}$, $Y_n = L(u,X)$. Then

$$q_Y(\xi) = q_X\left(f\big((\xi)\big)\right)|Jf(\xi)|,$$

where $f = col(f_1,\dots f_n)$, $f_1(\xi) = \xi_1$,..., $f_{n-1}(\xi) = \xi_{n-1}$, $f_n(\xi) = L(u,\xi)$. Consequently

$$Jf(\xi) = \frac{\partial L(u,\xi)}{\partial \xi_n}$$

$$q_Y(\xi) = q_X(\xi_1,\dots,\xi_{n-1},L(u,\xi_1,\dots,\xi_n))\left|\frac{\partial L(u,\xi_1,\dots,\xi_n)}{\partial \xi_n}\right|,$$

and

$$q_{Y_n}(\xi_n) = \int q_Y(\xi)d\xi_1\dots d\xi_{n-1} = \int q_X\left(f\big((\xi)\big)\right)|Jf(\xi)|d\xi_1\dots d\xi_{n-1}$$
$$= \int q_X(\xi_1,\dots,\xi_{n-1},L(u,\xi_1,\dots,\xi_n))\left|\frac{\partial L(u,\xi)}{\partial \xi_n}\right| d\xi_1\dots d\xi_{n-1},$$

hence

$$\sup_{u\in U}\mathbb{P}\big(\omega; L\big(u,X(\omega)\big) < \alpha\big) = \sup_{u\in U}\mathbb{P}(\omega; Y_n(\omega) < \alpha)$$
$$= \sup_{u\in U}\int_{-\infty}^{\alpha} q_{Y_n}(\xi_n)d\xi_n$$
$$= \sup_{u\in U}\int_{-\infty}^{\alpha}\left[\int q_X(\xi_1,\dots,\xi_{n-1},L(u,\xi_1,\dots,\xi_n))\left|\frac{\partial L(u,\xi)}{\partial \xi_n}\right| d\xi_1\dots d\xi_{n-1}\right]d\xi_n,$$

hence for

$$J(u) = \mathbb{P}\big(\omega; L\big(u,X(\omega)\big) < \alpha\big),$$

we have

$$J(u+\varepsilon v) = \mathbb{P}\big(\omega; L\big(u+\varepsilon v,X(\omega)\big) < \alpha\big)$$
$$= \int_{-\infty}^{\alpha}\left[\int q_X(\xi_1,\dots,\xi_{n-1},L(u+\varepsilon v,\xi_1,\dots,\xi_n))\left|\frac{\partial L(u+\varepsilon v,\xi)}{\partial \xi_n}\right| d\xi_1\dots d\xi_{n-1}\right]d\xi_n.$$

Theorem 11. The necessary condition of optimality for u^*, is

$$\int_{-\infty}^{\alpha}\left[\int \partial_L q_X(\xi_1,\dots,\xi_{n-1},L(u^*,\xi))\langle\nabla_u L(u^*,\xi),v\rangle\left|\frac{\partial L(u^*,\xi)}{\partial \xi_n}\right| d\xi_1\dots d\xi_{n-1}\right]d\xi_n +$$
$$\int_{-\infty}^{\alpha}[\int \operatorname{sign}(L_n(u^*,\xi))q_X(\xi_1,\dots,\xi_{n-1},L(u^*,\xi))\langle\nabla_u L_n(u^*,\xi),v\rangle d\xi_1\dots d\xi_{n-1}]d\xi_n = 0,$$

for all $v \in U$, where

$$L_n(u,\xi) = \frac{\partial L(u,\xi)}{\partial \xi_n}.$$

Proof.

$$\frac{\partial}{\partial\varepsilon}=\int_{-\infty}^{\alpha}\left[\int\partial_L q_X(\xi_1,\dots,\xi_{n-1},L(u,\xi))\langle\nabla_u L(u,\xi),v\rangle\left|\frac{\partial L(u,\xi)}{\partial\xi_n}\right|d\xi_1\dots d\xi_{n-1}\right]d\xi_n+$$
$$\int_{-\infty}^{\alpha}\left[\int\operatorname{sign}(L_n(u,\xi))q_X(\xi_1,\dots,\xi_{n-1},L(u,\xi))\langle\nabla_u L_n(u,\xi),v\rangle d\xi_1\dots d\xi_{n-1}\right]d\xi_n.$$

Example 12 (see [2]). Let $L(u,\xi)=\langle u,\xi\rangle^2$, $X\sim N(\mu,Q)$, then

$$\begin{aligned}J(u)&=\mathbb{P}\big(\omega;L\big(u,X(\omega)\big)<\alpha\big)\\&=\frac{1}{\sqrt{(2\pi)^n|Q|}}\int_A\exp\left\{-\frac{1}{2}(\xi-\mu)^TQ^{-1}(\xi-\mu)\right\}d\xi,\\A&=\{\mathbb{R}^n\ni\xi;\langle u,\xi\rangle^2<\alpha\}.\end{aligned}$$

Let

$$\xi=Q^{1/2}\eta+\mu,$$

hence

$$\begin{aligned}J(u)&=\frac{1}{\sqrt{(2\pi)^n}}\int_{A(u)}\exp\left\{-\frac{1}{2}\|\eta\|^2\right\}d\eta,\\A(u)&\triangleq\left\{\mathbb{R}^n\ni\eta;-\sqrt{\alpha}-\langle u,\mu\rangle<\left\langle Q^{1/2}u,\eta\right\rangle<\sqrt{\alpha}-\langle u,\mu\rangle\right\}.\end{aligned}$$

and

$$\begin{aligned}J(u)&=\frac{1}{\sqrt{2\pi}}\int_{a(u)}^{b(u)}\exp\left\{-\frac{t^2}{2}\right\}dt,\\a(u)&=-\left(\sqrt{\alpha}+\langle u,\mu\rangle\right)/\left[Q^{1/2}u\right]_n,\\b(u)&=\left(\sqrt{\alpha}-\langle u,\mu\rangle\right)/\left[Q^{1/2}u\right]_n,\end{aligned}$$

where $\left[Q^{1/2}u\right]_n$ is the last coordinate of $Q^{1/2}u$.

3.3 Markowitz models

Since the middle of 20th century when Harry Markowitz published his famous paper "Portfolio Selection" [1], the Markowitz Model (MM) begun its carrier in the financial world. There are several version of MM developed by his followers (Roy, Tobin, Sharpe) and even today MM is still an active research area. Using our notation and convention, let $u=col(u_1,\dots,u_n)$ denotes financial investment in risky assets, i.e. $\sum_{i=1}^n u_i=K$, where K is a total amount of money which has to be invested in stock. Let $X=col(X_1,\dots,X_n)$ is return's vector, hence $\langle u,X\rangle$ is a total investment profit. Elementary computations give

$$\mathbb{E}\langle u, X\rangle = \langle u, \mathbb{E}X\rangle = \langle u, m\rangle$$
$$var(\langle u, X\rangle) = \mathbb{E}[\langle u, X\rangle - \mathbb{E}\langle u, X\rangle]^2$$
$$= \mathbb{E}[\langle u, X\rangle - \langle u, m\rangle]^2$$
$$= \mathbb{E}[\langle u, X - m\rangle]^2$$
$$= \mathbb{E}u^T[X - m][X - m]^T u$$
$$= u^T(\mathbb{E}[X - m][X - m]^T)u = u^T Qu,$$

where $Q = cov(X, X)$, and $\mathbb{E}$ is expectation with respect $\mathbb{P}$. Markowitz posed the following

Problem 13. Given K and z, find

$$\min\{u^T Qu; \langle J, u\rangle = K, \langle m, u\rangle = z\},$$

where $J = col(1, \ldots, 1)$, or equivalently, find

$$\min\{\|y\|^2; \langle Q^{1/2}J, y\rangle = K, \langle Q^{1/2}m, y\rangle = z\}.$$

Solution 14. It is immediate that the minimal norm $y \in span(Q^{1/2}J, Q^{1/2}m)$, hence

$$y^* = aQ^{1/2}J + bQ^{1/2}m,$$

where a, b solves

$$a\langle QJ, J\rangle + b\langle QJ, m\rangle = K,$$
$$a\langle Qm, J\rangle + b\langle Qm, m\rangle = z.$$

If $\langle QJ, J\rangle\langle Qm, m\rangle - \langle QJ, m\rangle^2 \neq 0$, then there is a unique pair a^*, b^* which satisfies above equations. Thus

$$u^* = Q^{-1/2}y^* = a^*J + b^*m.$$

We are going now to describe the Markowitz problem using the loss functions and the smooth transformation approach. For

$$L(u, X) = \|u\|^2\left\|Q^{-1/2}[X - m]\right\|^2,$$

we have

$$\mathbb{P}\big(\omega; L\big(u, X(\omega)\big) < \alpha\big) = \int_A F_X(d\xi),$$
$$A = \left\{\mathbb{R}^n \ni \xi; \left\|Q^{-1/2}[X - m]\right\|^2 < \alpha/\|u\|^2\right\}.$$

If $X \sim N(m, Q)$, then $F_X(d\xi) = q_X(\xi)d\xi$, where

$$q_X(\xi) = (2\pi)^{-n/2}|Q|^{-1/2}\exp\left\{-\frac{1}{2}\left\|Q^{-1/2}[\xi - m]\right\|^2\right\}.$$

Let $\eta = Q^{-1/2}[\xi - m]$, then

$$\mathbb{P}\big(\omega; L\big(u, X(\omega)\big) < \alpha\big) = \int_A F_X(d\xi) =$$
$$= (2\pi)^{-n/2}|Q|^{-1/2} \int_A \exp\left\{-\tfrac{1}{2}\left\|Q^{-1/2}[\xi - m]\right\|^2\right\} d\xi =$$
$$= (2\pi)^{-n/2} \int \P_{B(0,\alpha/\|u\|^2)}(\eta) \exp\left\{-\tfrac{1}{2}\|\eta\|^2\right\} d\eta \triangleq \theta_n(\alpha, \|u\|).$$

Since $\theta_n(\alpha,\cdot)$, is decreasing on $(0,\infty)$, hence

$$\sup_{u\in U}\mathbb{P}\big(\omega; L\big(u, X(\omega)\big) < \alpha\big) = \sup_{u\in U}\theta_n(\alpha, \|u\|) = \theta_n(\alpha, \operatorname{arginf}\{\|u\|; u \in U\}).$$

But $U = \{\mathbb{R}^p \ni u; \langle J, u\rangle = K, \langle m, u\rangle = z\}$, hence u^* must be of the Markowitz form $u^* = a^*J + b^*m$.

4 Multivariate loss functions

We call $L\colon \mathbb{R}^p \times \mathbb{R}^n \ni (u,x) \to L(u,x) \in \mathbb{R}^m_+ \equiv [0,\infty)^m$ a **loss function**, iff

$$L_i(u, tx) \geq L_i(u, sx), \qquad \text{for } i = 1, \ldots, m,$$

$u \in \mathbb{R}^p$, $x \in \mathbb{R}^n$, and $t \geq s$.

Problem 15. Given L and $\alpha \in \mathbb{R}^m_+$, find

$$\sup_{u\in U}\mathbb{P}\big(\omega; L_1\big(u, X(\omega)\big) < \alpha_1, \ldots, L_m\big(u, X(\omega)\big) < \alpha_m\big).$$

Denote by F_X the distribution function of X, and by $q_X(\xi)$ its density (if exists). Then

$$\mathbb{P}\big(\omega; L_1\big(u, X(\omega)\big) < \alpha_1, \ldots, L_m\big(u, X(\omega)\big) < \alpha_m\big) = \begin{cases} \int_A F_X(d\xi), & \text{or} \\ \int_A q_X(\xi)d\xi, \end{cases}$$

where $A = \{\mathbb{R}^n \ni \xi; L_1(u,\xi) < \alpha_1, \ldots, L_m(u,\xi) < \alpha_m\}$.

Definition 16. We call a sequence $\big(\alpha_1, \ldots, \alpha_m, P(\alpha_1, \ldots, \alpha_m)\big)$, where

$$P(\alpha_1, \ldots, \alpha_m) = \sup_{u\in U}\mathbb{P}\big(\omega; L_1\big(u, X(\omega)\big) < \alpha_1, \ldots, L_m\big(u, X(\omega)\big) < \alpha_m\big),$$

a **risk graph** of (L, F_X).

Example 17. Let $L_1\big(u, X(\omega)\big) = L_1\big(u, X_1(\omega)\big)$ and $L_2\big(u, X(\omega)\big) = L_2\big(u, X_2(\omega)\big)$, where X_1, X_2, are independent random variables with distributions F_1, F_2, respectively.

Then

$$\sup_{u\in U}\mathbb{P}\big(\omega; L_1\big(u, X(\omega)\big) < \alpha_1, L_2\big(u, X(\omega)\big) < \alpha_2\big)$$
$$= \sup_{u\in U}\big[\mathbb{P}\big(\omega; L_1\big(u, X_1(\omega)\big) < \alpha_1\big) \cdot \mathbb{P}\big(\omega;, L_2\big(u, X_2(\omega)\big) < \alpha_2\big)\big]$$
$$= \sup_{u\in U}\Big(\int_{A_1} F_1(d\xi_1)\Big)\Big(\int_{A_2} F_2(d\xi_2)\Big),$$

where $A_i = \{\mathbb{R} \ni \xi_i; L_i(u, \xi_i) < \alpha_i\}$, $i = 1,2$.

4.1 Resulting loss function

4.1.1 Cascade paths

If the risky actions are in the '**cascade**' **path**, then it may be reasonable to introduce a weighted sum of loss functions

$$S(u, X) = \lambda_1 L_1(u, X) + \ldots + \lambda_m L_m(u, X)$$
$$= \langle \boldsymbol{\Lambda}, \boldsymbol{L}(\boldsymbol{u}, \boldsymbol{X})\rangle, \quad \boldsymbol{\Lambda} = \boldsymbol{col}(\boldsymbol{\lambda_1}, \ldots, \boldsymbol{\lambda_m}).$$

Then

$$\mathbb{P}\big(\omega; S\big(u, X(\omega)\big) < \alpha\big) = \mathbb{P}\big(\omega; \big\langle \Lambda, L\big(u, X(\omega)\big)\big\rangle < \alpha\big)$$
$$= \int_A F_X(d\xi_1 \ldots d\xi_n),$$
$$A = \{\xi \in \mathbb{R}^n; \langle \Lambda, L(u, \xi)\rangle < \alpha\},$$

hence

$$\sup_{u\in U}\mathbb{P}\big(\omega; S\big(u, X(\omega)\big) < \alpha\big) = \sup_{u\in U}\int_A F_X(d\xi_1 \ldots d\xi_n).$$

When $\lambda_1 = \ldots = \lambda_m = 1$, then we have an ordinary sum of loss functions, and $A = \{\mathbb{R}^n \ni \xi; \langle J, L(u, \xi)\rangle < \alpha\}$, $J = col(1, \ldots, 1)$, in this case.

4.1.2 Cascade paths

If the risky actions are **parallel**, then it may be reasonable to introduce

$$M(u, X) \triangleq \max\big(L_1(u, X), \ldots, L_m(u, X)\big)$$

as a resulting loss function.

4.2 Conditional risk

Sometimes it is necessary to include an additional information into the model. Operational risk models for instance, by definition include all information available up to the moments when the planned actions have to be executed. Information means here a knowledge that some events has occurred or not. In the language of probability theory

it has a form of sub-sigma field of $\mathcal{F}$, say G. Then the **conditional risk model** consists in solving

Problem 18. Given $G \subset \mathcal{F}$, L and $\alpha \geq 0$, find

$$\sup_{u(\omega)\in\mathbb{U}} \mathbb{P}\big(\omega; L\big(u, X(\omega)\big) < \alpha|\mathrm{G}\big).$$

where $\mathbb{P}(\cdot\,|\mathrm{G})$, is a conditional probability. Contrary to the previous problems, here we look for a $\mathrm{G}-$measurable random variable $u(\omega) \in \mathbb{R}^p$ and the class of such variables is denoted here by U.

If $F_X(\xi|\mathrm{G})$ is a conditional distribution function of X, and $q_X(\xi|\mathrm{G})$ its density (if exists), then

$$\mathbb{P}\big(\omega; L\big(u, X(\omega)\big) < \alpha|\mathrm{G}\big) = \begin{cases} \int_A F_X(d\xi|\mathrm{G}), & \text{or} \\ \int_A q_X(\xi|\mathrm{G})d\xi, \end{cases}$$

where $A = \{\mathbb{R}^n \ni \xi; L(u,\xi) < \alpha\}$.

4.3 Operational risk

If the action $u \in \mathbb{R}^p$ is subjected to imperfection or randomly perturbed, then the joint effect of these imperfection is called the **operational risk.** In the simplest case disturbances add to actions, i.e.,

$$v(\omega) = u + w(\omega),$$

where u is a chosen action, $w(\omega)$ models imperfections and disturbances, $v(\omega)$ is a **resulting** action. Let $Y(\omega) = col\big(X(\omega), w(\omega)\big)$. Then

$$K\big(u, Y(\omega)\big) \triangleq L\big(u + w(\omega), X(\omega)\big),$$

is a new loss function provided that $L(u, X)$ is. Due this modification of the loss function the operation risk can be included into considerations for ordinary loss functions.

5 Risk in net's operations

There is sometimes a need to consider a risk of more advanced, multistage, activity which is conducted in many different parallel paths. The paths crossed at some stages and the next actions can be initiated when the previous have already been completed. The paths structure is called a **net**. At each stage there is associated an elementary risk which we have considered in previous sections. The problem which we are going to consider in this section is how to model the resulting risk of the whole project which is mapped in the net. We have considered a sequence of actions realized on the path and called such structure a cascade. The resulting loss function of a cascade is a sum of the

elementary risk functions. For parallel paths starting from the initial point **S** and ending at the point **E**, it is natural to define the resulting loss function M as

$$M(u,x) \triangleq \max\big(L_1(u,x),\dots,L_m(u,x)\big),$$

where L_i, $i=1,\dots,m$, is the loss functions of the $i^{th}-$ path.

In general, the loss functions reflects three kind of losses; financial measured monetary, temporal measured at time units, and product's quality loss which results in demand's decreasing. It is natural to accept the following rules:

- the financial losses always add despite the place in net where they occur - thus they are always in cascade;
- the temporal losses are in cascade only when they are in cascade path. If the paths are parallel, then the temporal losses are parallel too;
- the quality losses can belong to the both classes thus requires individual case studies.

Example 19 (Example 4 continued)

$$L_1(u,X) = X_1/u, \text{temporal losses component},$$

$$L_2(u,X) = X_1X_2 + C(X_1/u), \text{financial losses component}.$$

6 Generalizations

In the more realistic and advanced cases the distribution function F_X of X, (thus the X itself), can depend of action chosen u. This will be reflected in the notations F_X^u, q_X^u, X^u, if needed. Hence, the problems considered previously, takes now the corrected form; given L and $\alpha \in \mathbb{R}_+^m$, find

$$\sup_{u\in U}\mathbb{P}\big(\omega; L_1\big(u,X^u(\omega)\big) < \alpha_1,\dots,L_m\big(u,X^u(\omega)\big) < \alpha_m\big),$$

where

$$\mathbb{P}\big(\omega; L_1\big(u,X^u(\omega)\big) < \alpha_1,\dots,L_m\big(u,X^u(\omega)\big) < \alpha_m\big) = \begin{cases} \int_A F_X^u(d\xi), & \text{or} \\ \int_A q_X^u(\xi)d\xi, \end{cases}$$

where $A = \{\mathbb{R}^n \ni \xi; L_1(u,\xi) < \alpha_1,\dots,L_m(u,\xi) < \alpha_m\}$.

Example 20. In the Example 4 we have assumed that the number T of working hours needed to complete the reparation does not depend on the number of workers who do the job. However, everyday practice shows that, in contrary, it strongly does. The generalized version reads as follows. The pause needs T_u working hours for reparations if u workers are employed. Let KT_u is a cost of reparation which takes T_u working hours, where K is the reparation cost per hour. Let $C(T_u)$ represents a cost of T_u hours pause of

device's work. Then the total cost of the breakdown is $KT_u + C(T_u)$, provided u workers are employed. Denote $X^u = col(X_1^u, X_2) = col(T_u, K)$. Then the loss function is

$$L_1(u, X^u) = X_1^u, \text{temporal losses component},$$

$$L_2(u, X^u) = X_1^u X_2 + C(X_1^u), \text{financial losses component}.$$

This formula reflects a joint effect of randomness represented by X^u and actions chosen represented here by u. If F_{TK}^u is a joint distribution function of the random variable $col(T_u, K)$, then

$$\mathbb{P}(\omega; L_1(u, X^u) < \alpha_1, L_2(u, X^u) < \alpha_2) = \int_A F_{TK}^u(d\xi),$$

where $A = \{\mathbb{R}^2 \ni \xi; \xi_1 < \alpha_1, \xi_1\xi_2 + C(\xi_1) < \alpha_2\}$.

6.1 Approach via smooth transformations

It follows that theory developed in Section 2.1 needs some modifications. Indeed, in the general case, we have

$$\begin{aligned}
&\int_A q_X^u(\xi)d\xi = \int_{\mathbb{R}^n} \P_A(\xi) q_X^u(\xi)d\xi \\
&= \int \P_{B(0,r)}(\eta) q_X^u\big(f(u,\eta)\big) Jf(u,\eta)d\eta \\
&= \int_0^\infty \left(\int_{\mathbb{R}^{n-1}} \P_{S^{n-1}(\tau)\cap B(0,r)}(\eta) q_X^u\big(f(u,\eta)\big) Jf(u,\eta) d\mathcal{H}^{n-1}(\eta) \right) d\tau \\
&= \int_0^{r(u)} \left(\int_{\mathbb{R}^{n-1}} \P_{S^{n-1}(\tau)}(\eta) q_X^u\big(f(u,\eta)\big) Jf(u,\eta) d\mathcal{H}^{n-1}(\eta) \right) d\tau \\
&\equiv J(u),
\end{aligned}$$

where in the line 2 we have changed variables putting $\xi = f(u,\eta)$, where $f: \mathbb{R}^p \times \mathbb{R}^n \to \mathbb{R}^n$ is a smooth mapping such that

$$f\big(u, B(0,r)\big) = A = \{\mathbb{R}^n \ni \xi; L_1(u,\xi) < \alpha_1, \ldots, L_m(u,\xi) < \alpha_m\},$$

for some $r = r(\alpha, u)$ dependent on α and u. In the next lines integration ([3]) is performed with respect to the Hausdorff measure $\mathcal{H}^{n-1}$, and $S^{n-1}(\tau) = \{\mathbb{R}^n \ni \xi; \|\xi\|_{\mathbb{R}^n} = \tau\}$ is a centred sphere of radius τ, and Jf is a Jacobian of f. Hence, the necessary condition for $u^{\text{å}}$ to maximize $\mathbb{P}\big(\omega; L\big(u, X(\omega)\big) < \alpha\big)$, is

$$\frac{\partial}{\partial \varepsilon}[J(u + \varepsilon v)]_{\varepsilon=0} = 0.$$

Theorem 21.

$$\frac{\partial}{\partial\varepsilon}[J(u+\varepsilon v)]_{\varepsilon=0}$$
$$= \langle \nabla r(u^{\mathring{a}}), v\rangle \int_{\mathbb{R}^{n-1}} \P_{S^{n-1}(r(u^{\mathring{a}}))}(\eta) q_X\left(f(u^{\mathring{a}},\eta)\right) Jf(u^{\mathring{a}},\eta) d\mathcal{H}^{n-1}(\eta)$$
$$+ \int_0^{r(u^{\mathring{a}})} \left(\int_{\mathbb{R}^{n-1}} \P_{S^{n-1}(\tau)}(\eta) \left\langle \nabla_u q^{u^{\mathring{a}}}\left(f(u^{\mathring{a}},\eta)\right), v\right\rangle Jf(u^{\mathring{a}},\eta) d\mathcal{H}^{n-1}(\eta)\right) d\tau$$
$$+ \int_0^{r(u^{\mathring{a}})} \left(\int_{\mathbb{R}^{n-1}} \P_{S^{n-1}(\tau)}(\eta) \langle \nabla_u Jf(u^{\mathring{a}},\eta), v\rangle q_X^{u^{\mathring{a}}}\left(f(u^{\mathring{a}},\eta)\right) d\mathcal{H}^{n-1}(\eta)\right) d\tau$$

Proof. Differentiation

$$\frac{\partial}{\partial\varepsilon}\left[\int_0^{r(u^{\mathring{a}}+\varepsilon v)} \left(\int_{\mathbb{R}^{n-1}} {}_{S^{n-1}(\tau)}(\eta) q_X^{u^{\mathring{a}}+\varepsilon v}\left(f(u^{\mathring{a}}+\varepsilon v,\eta)\right) Jf(u^{\mathring{a}}+\varepsilon v,\eta) d\mathcal{H}^{n-1}(\eta)\right) d\tau \right]_{\varepsilon=0},$$

gives the result.

6.2 Standard approach

Let us define a new random variable $Y = col(Y_1,\dots Y_n)$, such that $Y_1 = X_1$, ..., $Y_{n-m} = X_{n-m}$, $Y_{n-m+1} = L_1(u,X)$,...,$Y_n = L_m(u,X)$. Then

$$q_Y(\xi) = q_X\left(f((\xi))\right)|Jf(\xi)|,$$

where $f = col(f_1,\dots f_n)$, $f_1(\xi) = \xi_1$,..., $f_{n-m}(\xi) = \xi_{n-m}$,...,$f_{n-m+1}(\xi) = L_1(u,\xi)$,..., $f_n(\xi) = L_m(u,\xi)$. Consequently for $Y(m,n) = col(Y_{n-m+1},\dots,Y_n)$ we have

$$q_{Y(m,n)}(\xi_{n-m+1},\dots,\xi_n) = \int q_Y(\xi) d\xi_1 \dots d\xi_{n-m} = \int q_X\left(f((\xi))\right)|Jf(\xi)| d\xi_1 \dots d\xi_{n-m}$$
$$= \int q_X(\xi_1,\dots,\xi_{n-m}, L_1(u,\xi),\dots,L_m(u,\xi))\left||Jf(\xi)|\right| d\xi_1 \dots d\xi_{n-m}$$

and

$$\sup_{u\in U} \mathbb{P}\left(\omega; L_1\left(u,X(\omega)\right) < \alpha_1,\dots,L_m\left(u,X(\omega)\right) < \alpha_m\right)$$
$$= \sup_{u\in U} \mathbb{P}(\omega; Y_1(\omega) < \alpha_1,\dots,Y_m(\omega) < \alpha_m)$$
$$= \sup_{u\in U} \int_{-\infty}^{\alpha_1} \dots \int_{-\infty}^{\alpha_m} q_{Y(m,n)}(\xi_{n-m+1},\dots,\xi_n) d\xi_{n-m+1} \dots d\xi_n$$

$$= \sup_{u\in U} \int_{-\infty}^{\alpha_1} \dots \int_{-\infty}^{\alpha_m} \Big[\int q_X(\xi_1, \dots, \xi_{n-m}, L_1(u,\xi), \dots, L_m(u,\xi)) \\ |Jf^u(\xi)| d\xi_1 \dots d\xi_{n-m} \Big] d\xi_{n-m+1} \dots d\xi_n,$$

hence for

$$J(u) = \mathbb{P}\big(\omega; L_1\big(u, X(\omega)\big) < \alpha_1, \dots, L_m\big(u, X(\omega)\big) < \alpha_m\big),$$

we have

$$J(u+\varepsilon v) = \mathbb{P}\big(\omega; L_1\big(u+\varepsilon v, X(\omega)\big) < \alpha_1, \dots, L_m\big(u+\varepsilon v, X(\omega)\big) < \alpha_m\big) = \\ \int_{-\infty}^{\alpha_1} .. \int_{-\infty}^{\alpha_m} \Big[\int q_X(\xi_1, .., \xi_{n-m}, L_1(u+\varepsilon v, \xi), .., \quad L_m(u \\ + \varepsilon v, \xi)) |Jf^{u+\varepsilon v}(\xi)| d\xi_1 .. d\xi_{n-m} \Big] d\xi_{n-m+1} .. d\xi_n.$$

Theorem 22. The necessary condition of optimality for u^* is

$$\int \partial_L q_X(\xi_1, \dots, \xi_{n-1}, L(u^*,\xi)) \langle \nabla_u L(u^*,\xi), v\rangle \left| \frac{\partial L(u^*,\xi)}{\partial \xi_n} \right| d\xi_1 \dots d\xi_{n-1}$$

$$+ \int \operatorname{sign}(L_n(u^*,\xi)) q_X(\xi_1, \dots, \xi_{n-1}, L(u^*,\xi)) \langle \nabla_u L_n(u^*,\xi), v\rangle d\xi_1 \dots d\xi_{n-1} = 0,$$

where

$$L_n(u,\xi) = \frac{\partial L(u,\xi)}{\partial \xi_n}.$$

Proof.

$$\frac{\partial}{\partial \varepsilon}[J(u+\varepsilon v)]_{\varepsilon=0} = \int \partial_L q_X(\xi_1, \dots, \xi_{n-1}, L(u,\xi)) \langle \nabla_u L(u,\xi), v\rangle \left| \frac{\partial L(u,\xi)}{\partial \xi_n} \right| d\xi_1 \dots d\xi_{n-1} \\ + \int \operatorname{sign}(L_n(u,\xi)) q_X(\xi_1, \dots, \xi_{n-1}, L(u,\xi)) \langle \nabla_u L_n(u,\xi), v\rangle d\xi_1 \dots d\xi_{n-1}.$$

7 Average risk minimization

Sometime it is useful to consider

Problem 23. Given a loss function L, *and* $X: \Omega \to \mathbb{R}^n$, a vector valued random variable defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$, find

$$\inf_{u\in U} \mathbb{E}\big[L\big(u, X(\omega)\big)\big].$$

where $\mathbb{E}$ stands for expectation against measure $\mathbb{P}$.

In our notations

$$\mathbb{E}\big[L\big(u,X(\omega)\big)\big]=\int_{\mathbb{R}^n}L(u,\xi)q_X(\xi)d\xi.$$

Example 24 (see [4]) (Games against Nature) Consider two person, zero-sum game with matrix payoff $A=\big(a_{ij}\big)$, $i=1,\dots,n$, $j=1,\dots,m$ which models possible losses a_{ij} when the player N (Nature) applies strategy i^{th} *row* against defender's strategy j^{th} column. We assume that *N* has chosen the probability distribution $p=col(p_1,\dots,p_n)$, as strategy for playing the rows $r_1,\dots,r_n$ *of A*, see figure below

$$[N\backslash D][q_1,\ \dots\ ,q_m]$$
$$\begin{bmatrix}p_1\\ \dots\\ p_n\end{bmatrix}\begin{bmatrix}a_{11} & \dots & a_{1m}\\ . & \dots & .\\ a_{n1} & \dots & a_{nm}\end{bmatrix}.$$

Thus

$$\mathbb{E}\big[L\big(u,X(\omega)\big)\big]=\sum_{1\le i\le n}\sum_{1\le j\le m}p_ia_{ij}q_j,$$

where the probability vector $q=col(q_1,\dots,q_m)$ is a strategy for playing columns of A by defender. Consequently,

$$\inf_{u\in U}\mathbb{E}\big[L\big(u,X(\omega)\big)\big]=\inf_q\sum_{1\le i\le n}\sum_{1\le j\le m}p_ia_{ij}q_j=\inf_q\langle w,q\rangle,$$

where

$$w=col(w_1,\dots,w_m),$$
$$w_j=\sum_{1\le i\le n}p_ia_{ij}.$$

Since $q_j\ge 0$, $\sum_{1\le j\le m}q_j=1$, hence the best q^* must be in the form $q_k=1$, for some $1\le k\le m$, and $q_j=0$, for $j\ne k$. Certainly, $w_k\le w_j$, for all $j\ne k$.

8 Decisions with risk

8.1 Projects selection

We shall consider in this section the following decision problem: one has to select one project between many alternatives. With each of them there is associated a loss function L_i, $i=1,\dots,d$, and the risk graphs $\big(\alpha_i,P(\alpha_i)\big)$, of $\big(L_i,F_{X^i}\big)$, where X^i is a random variable associated with i^{ih} project. How one has to select the 'right' project? If there is the one, common for all projects a level of risk tolerance, say $\aleph$, then one should

select a project, say p, such that $\alpha_p = \min\{\alpha_i; P(\alpha_i) = \aleph\}$. However, when these projects are distributed between different areas of public administration, such as health, security, transport, etc., then there no exists one, level of risk tolerance which is commonly accepted and this rule has to be modified. Let $\aleph_i$, $i = 1, \ldots, d$, is a risk tolerance level in i^{th} area of project implementation. Then the previous rule can be generalized to

$$\alpha_p = \min\{\alpha_i; P(\alpha_i) = \aleph_i\}.$$

8.2 Losses minimization under risk level constraints

If decision maker is interested with loss minimization, then he should consider if there exists a risk tolerance level which cannot be exceeded (see [2] for extended discussion). Once this level $r \in [0,1]$, is chosen, then it is natural to define a set

$$C_{r,\alpha} = \{U \ni u_r; \mathbb{P}(\omega; L(u, X(\omega)) > \alpha) \leq r\},$$

of admissible strategies such that probability of losses bigger than α, is less than r. Formally we can state

Problem 25. Given U, F_X, find a strategy u_r such that

$$\mathbb{P}\left(\omega; L(u_r, X(\omega)) \leq L(u, X(\omega))\right) > 1 - r.$$

for $u \in C_{r,\alpha}$.

Proposition 26

$$u_r = \arg\inf\{F^{-1}_{L(u)}(1-r); u \in C_{r,\alpha}\},$$

where

$$F_{L(u)}(\alpha) = \int_{-\infty}^{\alpha} q_{L(u)}(\alpha) d\xi_n,$$

$$q_{L(u)}(\xi_n) = \int q_X(\xi_1, \ldots, \xi_{n-1}, L(u, \xi_1, \ldots, \xi_n)) \left|\frac{\partial L(u,\xi)}{\partial \xi_n}\right| d\xi_1 \ldots d\xi_{n-1}.$$

Proof. Since $q_{L(u)} \geq 0$, $F^{-1}_{L(u)}$ can be defined as

$$F^{-1}_{L(u)}(a) = \inf\{s; F_{L(u)}(s) < a\}.$$

From section entitled 'Standard approach'

$$\mathbb{P}(\omega; L(u, X(\omega)) < \alpha) = F_{L(u)}(\alpha) \geq 1 - r,$$

hence $\alpha \geq F^{-1}_{L(u)}(1-r)$, i.e., minimizing $L(u, X(\omega))$, under fixed probability constrain implies minimization of $F^{-1}_{L(u)}(1-r)$ with respect to $u \in C_{r,\alpha}$.

Crucial for implementing numerical simulations is the following'

Corollary 27 *If* $(u,\alpha)\to F_{L(u)}(\alpha)$ belongs to $C(\mathbb{R}^p\times\mathbb{R})$, then minimizing $F^{-1}_{L(u)}(1-r)$ is equivalent to minimization of $F_{L(u)}(1-r)$.

Proof. We have

$$F^{-1}_{L(u_r)}(1-r)=\inf_{u\in C_{r,\alpha}}F^{-1}_{L(u)}(1-r)=\inf_{u\in C_{r,\alpha}}\inf\{s;F_{L(u)}(s)<1-r\}$$

$$=\inf\Big\{s;\inf_{u\in C_{r,\alpha}}F_{L(u)}(s)<1-r\Big\}\,(\text{bycontinuityof}F_L)$$

$$=\inf\{s;F_{L(u^{r,\alpha})}(s)<1-r\}$$

$$=F^{-1}_{L(u^{r,\alpha})}(1-r)u_r=u^{r,\alpha},$$

where $u^{r,\alpha}$ is minimizing strategy for F_L.

9 Numerical simulations

Theory developed in the previous sections allows to make risk minimization given loss function $L(u,X)$ and probability distribution F_X of random element X. However, this approach has some defects. In order to use differential calculus for optimization one must be sure that all functions appearing in the model are smooth. But, as some examples show dependence F^u on u does not, and not-smooth optimization seems to be inevitable. On the other hand, when the model is smooth, then standard variational methods of optimization lead to analytical formulae which are complex and difficult to apply. It looks therefore reasonable to consider parallel approach based on numerical simulations.

9.1 Monte-Carlo method

For the clarity of presentation we consider first the problem of probability calculation via the M-C method under simplifying assumptions

9.1.1 X_1,...,X_n are stochastically independent

We assume firstly that X^u does not depend on u, and simply write X, instead of X^u. Secondly, we assume that coordinates of X, i.e., the random variables X_1,...,X_n, are stochastically independent, hence $F_X=F_1\otimes\ldots\otimes F_n$, where $F_i=F_{X_i}$. Similarly for densities $q_X=q_1\otimes\ldots\otimes q_n$, where $q_i=q_{X_i}$. Thus

$$\mathbb{P}\big(\omega;L_1\big(u,X(\omega)\big)<\alpha_1,\ldots,L_m\big(u,X(\omega)\big)<\alpha_m\big)$$

$$=\int_A\prod_{i=1}^n q_i(\xi_i)d\xi_i,$$

$$A=\{\mathbb{R}^n\ni\xi;L_1(u,\xi)<\alpha_1,\ldots,L_m(u,\xi)<\alpha_m\}.$$

Now we introduce a new probability space $S = \left([0,1)^n, \sqsubseteq(\mathcal{B}), \Lambda^n_{[0,1)}\right)$, where $\sqsubseteq(\mathcal{B})$ is a sigma field of Borel sets on $\mathbb{R}^n$, restricted to $[0,1)^n$, and $\Lambda^n_{[0,1)}$ is a n-dimensional Lebesgue measure restricted to $[0,1)^n$. The generic elements of $[0,1)^n$ we denote by $w = (w_1, \ldots, w_n)$. On S define random variables $U = (U_1, \ldots, U_n)$, by the formulae

$$U_i(w) = F_i^{-1}(w_i),$$

where

$$F_i^{-1}(a) = \inf\{s; F_i(s) < a\}.$$

Now

$$\begin{aligned} F_U(\xi) &= \Lambda^n_{[0,1)}(w; U_1(w) < \xi_1, \ldots, U_n(w) < \xi_n) \\ &= \Lambda^n_{[0,1)}\big(w; w_1 < F_1(\xi_1), \ldots, w_n < F_n(\xi_n)\big) \\ &= \textstyle\prod_{i=1}^n F_i(\xi_i) = F_X(\xi) = \mathbb{P}(\omega; X_1(\omega) < \xi_1, \ldots, X_n(\omega) < \xi_n), \end{aligned}$$

what implies

$$\begin{aligned} \mathbb{P}\big(\omega; L_1\big(u, X(\omega)\big) < \alpha_1, \ldots, L_m\big(u, X(\omega)\big) < \alpha_m\big) \\ = \Lambda^n_{[0,1)}\big(w; L_1\big(u, U(w)\big) < \alpha_1, \ldots, L_m\big(u, U(w)\big) < \alpha_m\big) \\ = \Lambda^n_{[0,1)}\Big(w; L_1\Big(u, F_1^{-1}(w_1), \ldots, F_n^{-1}(w_n)\Big) \\ < \alpha_1, \ldots, L_m\Big(u, F_1^{-1}(w_1), \ldots, F_n^{-1}(w_n)\Big) < \alpha_m\Big). \end{aligned}$$

Consequently, by the Central Limit Theorem the probability

$$\mathbb{P}\big(\omega; L_1\big(u, X(\omega)\big) < \alpha_1, \ldots, L_m\big(u, X(\omega)\big) < \alpha_m\big)$$

is approximately equal to

$$\simeq \frac{1}{N}\sum_{j=1}^{N} \P_\Phi(w),$$

where

$$\begin{aligned} \Phi = \Big\{[0,1)^n \ni w; L_1\Big(u, F_1^{-1}(w_1), \ldots, F_n^{-1}(w_n)\Big) \\ < \alpha_1, \ldots, L_m\Big(u, F_1^{-1}(w_1), \ldots, F_n^{-1}(w_n)\Big) < \alpha_m\Big\} \end{aligned}$$

and therefore can be obtained by independent simulations of w_i, $i = 1, \ldots, n$, according to the uniform distribution on $[0,1)$, and summing $\sum_{j=1}^N \P_\Phi(w)/N$.

Final remark in this section; when X^u is $u-$ dependent, all calculations have to be done for each fixed u, and then

$$\sup_{u\in U}\mathbb{P}\big(\omega; L_1\big(u, X(\omega)\big) < \alpha_1, \ldots, L_m\big(u, X(\omega)\big) < \alpha_m\big)$$
$$\simeq \frac{1}{N}\sup_{u\in U}\sum_{j=1}^{N} \P_{\Phi(u)}(w),$$

where

$$\Phi(u) = \Big\{[0,1)^n \ni w; L_1\Big(u, F_{u1}^{-1}(w_1), .., F_{un}^{-1}(w_n)\Big) < \alpha_1, .., L_m\Big(u, F_{u1}^{-1}(w_1), .., F_{un}^{-1}(w_n)\Big) < \alpha_m\Big\}$$

and

$$F_{ui}(\alpha) = \mathbb{P}(\omega; X_i^u(\omega) < \alpha).$$

Remark 28. When X_1,...,X_n are stochastically dependent, then there no more exist F_i, $i = 1, .., n$ such that $F_X = F_1 \otimes \ldots \otimes F_n$. However, the approach presented in this section can be applied as a first step approximation. For this purpose, define $\tilde{F}_i(\xi_i)$, $i = 1, \ldots, n$ as marginal distributions *of* F_X, *i.e.*,

$$\tilde{F}_i(\xi_i) = \int_{\mathbb{R}^{n-1}} F_X(\xi_1, \ldots, \xi_n) \prod_{j\neq i} d\xi_j.$$

Besides $F_X \neq \tilde{F}_1 \otimes \ldots \otimes \tilde{F}_n$, all calculations which were done above can be repeated and the final error become as smaller as stochastically dependences are weaker.

9.1.2 $X = f(U)$

We shall consider now the case with a given smooth, invertible mapping $f: \mathbb{R}^n \to \mathbb{R}^n$, such that

$$X = f(U),$$

where $U = \mathbf{col}(U_1, \ldots, U_n)$, and $U_1, \ldots, U_n$ is a sequence of independent random variables uniformly distributed on $[0,1)$. Because of the inverse assumption

$$U = f^{-1}(X) = \mathbf{col}\big(\phi_1(X), \ldots, \phi_n(X)\big)$$

there exists a collection $\phi_1(\cdot), \ldots, \phi_n(\cdot)$, of scalar functions ϕ_i on $\mathbb{R}^n$, such that $\phi_1(X), \ldots, \phi_n(X)$, are stochastically independent. Uniform distributions of $\phi_i(X)$, implies that as basic probability space we may choose $S = ([0,1)^n, \sqsubseteq(\mathcal{B}), \Lambda^n_{[0,1)})$ from the previous subsection. From probability theory we know, that for any bounded $h: \mathbb{R}^n \to \mathbb{R}$,

$$\mathbb{E}\big[h\big(f(U)\big)\big] = \int h\big(f(\eta)\big) q_U(\eta) d\eta = \int h(\xi) q_U\big(f^{-1}(\xi)\big) |Jf^{-1}(\xi)| d\xi$$
$$= \mathbb{E}[h(X)] = \int h(\xi) q_X(\xi) d\xi,$$

hence

$$q_X(\xi) = q_U\big(f^{-1}(\xi)\big)|Jf^{-1}(\xi)| = \prod_{i=1}^{n} q_i\big(f^{-1}(\xi)\big)|Jf^{-1}(\xi)|,$$

by independence of U_i, and where $q_i = q_{U_i}$, $q_i(a) = \P_{[0,1)}(a)$. This means that for $X = f(U)$, we have

$$\mathbb{P}\big(\omega; L_1\big(u, X(\omega)\big) < \alpha_1, \ldots, L_m\big(u, X(\omega)\big) < \alpha_m\big)$$

$$= \int_A q_X(\xi)d\xi = \int_A \prod_{i=1}^{n} q_i\big(f^{-1}(\xi)\big)|Jf^{-1}(\xi)|d\xi = \int_A \prod_{i=1}^{n} \P_{[0,1)}\big(f^{-1}(\xi)\big)|Jf^{-1}(\xi)|d\xi,$$

$$A = \{\mathbb{R}^n \ni \xi; L_1(u,\xi) < \alpha_1, \ldots, L_m(u,\xi) < \alpha_m\}.$$

By using the last integral expression one can simulate numerically the inquired probability in the following procedure:

1. select randomly an independent sequence $u_1(1), \ldots, u_n(1)$, from $[0,1)$ according to the uniform distribution,
2. compute $\xi^1 = \mathbf{col}\big(\xi_1(1), \ldots; , \xi_n(1)\big)$,
   where $\xi_1(1) = f\big(u_1(1), \ldots, u_n(1)\big), \ldots, \xi_n(1) = f\big(u_1(1), \ldots, u_n(1)\big)$;
3. check if $L_1(u, \xi^1) < \alpha_1, \ldots, L_m(u, \xi^1) < \alpha_m$;
4. if so take a number one, if not take zero;
5. repeat step 1 with a sequence $u_1(2), \ldots, u_n(2)$;
6. repeat step 2 with a sequence $\xi^2 = \mathbf{col}\big(\xi_1(2), \ldots, \xi_n(2)\big)$,
   where $\xi_1(2) = f\big(u_1(2), \ldots, u_n(2)\big), \ldots, \xi_n(2) = f\big(u_1(2), \ldots, u_n(2)\big)$;
7. repeat step 3 by checking if $L_1(u, \xi^2) < \alpha_1, \ldots, L_m(u, \xi^2) < \alpha_m$;
8. repeat step 4;
9. repeat steps 1-4 in feedback N times.

Use these computations in the sum

$$\mathbb{P}\big(\omega; L_1\big(u, X(\omega)\big) < \alpha_1, \ldots, L_m\big(u, X(\omega)\big) < \alpha_m\big)$$

$$\simeq \frac{1}{N}\sum_{i=1}^{N} \P_A\big(\xi^i\big)\big|Jf^{-1}\big(\xi^i\big)\big|.$$

9.1.3 General case

We shall consider now the case when F_X, and F_U are given and we are looking for a smooth mapping $f: \mathbb{R}^n \to \mathbb{R}^n$, such that

$$X \overset{D}{=} f(U),$$

where $\overset{D}{=}$ means distributional equality, i.e., $F_X = F_{f(U)}$, and where U is a random vector uniformly distributed on $[0,1)^n$. Let $f^{-1}(\xi) = \phi(\xi) = \mathbf{col}\big(\phi_1(\xi), \ldots, \phi_n(\xi)\big)$. Then

$$F_X(\xi) = \mathbb{P}(\omega; X_1(\omega) < \xi_1, \dots, X_n(\omega) < \xi_m)$$
$$= \mathbb{P}\big(\omega; U_1(\omega) < \phi_1(\xi), \dots, U_n < \phi_n(\xi)\big)$$
$$= F_U\big(\phi_1(\xi), \dots, \phi_n(\xi)\big) = \prod_{i=1}^{n} F_{U_i}\big(\phi_i(\xi)\big).$$

Problem 29. Given F_X, find $f\colon \mathbb{R}^n \to \mathbb{R}^n$ such that

$$F_X(\xi) = \prod_{i=1}^{n} F_{U_i}\big(\phi_i(\xi)\big) = \prod_{i=1}^{n} \phi_i(\xi),$$

for $0 \le \phi_i(\xi) < 1$, $i = 1, \dots, n$.

Definition 30. Let us denote by $F_1, F_{12}, \dots, F_{1,\dots,n-1}$, 'cascade' marginals *of* F_X, defined as

$$F_1(\xi_1) = \lim_{\xi_2,\dots,\xi_n \to \infty} F_X(\xi),$$
$$F_{1,2}(\xi_1, \xi_2) = \lim_{\xi_3,\dots,\xi_n \to \infty} F_X(\xi),$$
$$\dots$$
$$F_{1,\dots,n-1}(\xi_1, \dots, \xi_{n-1}) = \lim_{\xi_n \to \infty} F_X(\xi).$$

We shall find a solution map f in the special, 'cascade' form

$$\xi_1 = f_1(u_1),$$
$$\xi_2 = f_2(u_1, u_1),$$
$$\dots$$
$$\xi_n = f_n(u_1, \dots, u_n).$$

Theorem 31.

$$f_1(u_1) = F_1^{-1}(u_1),$$
$$f_2(u_1, u_2) = F_{1,2}^{-1}(F_1^{-1}(u_1), u_1 u_2),$$
$$f_3(u_1, u_2, u_3) = F_{1,2,3}^{-1}\Big(F_1^{-1}(u_1), F_{1,2}^{-1}(F_1^{-1}(u_1), u_1 u_2 u_3)\Big),$$
$$\dots$$
$$f_n(u_1, \dots, u_n) = F_X^{-1}\Big(F_1^{-1}(u_1), F_{1,2}^{-1}(F_1^{-1}(u_1), \dots, u_1 u_2 u_3)\Big)$$

where $F_{1,\dots,k}^{-1}$, $k = 1, \dots, n-1$, is an inverse of $F_{1,\dots,k}(\xi_1, \dots, \xi_{k-1}, \cdot)$, with respect to the last variable when $\xi_1, \dots, \xi_{k-1}$, are fixed. Similarly, F_X^{-1} is an inverse of $F_X(\xi_1, \dots, \xi_{n-1}, \cdot)$.

Proof. To prove the first equality, note

$$F_1(\xi_1) = F_{X_1}(\xi_1) = \mathbb{P}(X_1 < \xi_1) = \mathbb{P}(f_1(U_1) < \xi_1) = \mathbb{P}\big(U_1 < \phi_1(\xi_1)\big) = \phi_1(\xi_1),$$

hence

$$\xi_1 = \phi_1^{-1}(u_1) = F_1^{-1}(u_1) = f_1(u_1).$$

Similarly for the second

$$F_{1,2}(\xi_1, \xi_2) = F_{X_1X_2}(\xi_1, \xi_2) = \mathbb{P}(X_1 < \xi_1, X_2 < \xi_2)$$
$$= \mathbb{P}\big(U_1 < \phi_1(\xi_1), U_2 < \phi_2(\xi_1, \xi_2)\big) = \phi_1(\xi_1)\phi_2(\xi_1, \xi_2),$$

hence

$$\phi_2(\xi_1, \xi_2) = \frac{F_{1,2}(\xi_1, \xi_2)}{\phi_1(\xi_1)}, F_{1,2}(\xi_1, \xi_2) = u_1u_2,$$
$$\xi_2 = F_{1,2}^{-1}(\xi_1, u_1u_2) = F_{1,2}^{-1}(F_1^{-1}(u_1), u_1u_2) = f_2(u_1, u_2).$$

Finally for the third

$$F_{1,2,3}(\xi_1, \xi_2, \xi_3) = F_{X_1X_2X_3}(\xi_1, \xi_2, \xi_3) = \mathbb{P}(X_1 < \xi_1, X_2 < \xi_2, X_3 < \xi_3)$$
$$= \mathbb{P}\big(U_1 < \phi_1(\xi_1), U_2 < \phi_2(\xi_1, \xi_2), U_3 < \phi_3(\xi_1, \xi_2, \xi_3)\big)$$
$$= \phi_1(\xi_1)\phi_2(\xi_1, \xi_2)\phi_3(\xi_1, \xi_2, \xi_3),$$

hence

$$u_3 = \phi_3(\xi_1, \xi_2, \xi_3) = \frac{F_{1,2,3}(\xi_1, \xi_2, \xi_3)}{\phi_1(\xi_1)\phi_2(\xi_1, \xi_2)} = \frac{F_{1,2,3}\big(F_1^{-1}(u_1), F_{1,2}^{-1}\big(F_1^{-1}(u_1), u_1u_2\big), \xi_3\big)}{u_1u_2}$$

and consequently

$$\xi_3 = F_{1,2,3}^{-1}\big(F_1^{-1}(u_1), F_{1,2}^{-1}(F_1^{-1}(u_1), u_1u_2), u_1u_2u_3\big) = f_3(u_1, u_2, u_3).$$

Next steps follow by induction.

To complete the proof, we have to show that the mapping f is the solution of our problem, i.e., $F_X = F_{f(U)}$. Note that

$$F_{f(U)}(\xi) = \mathbb{P}(f_1(U_1) < \xi_1, f_2(U_1, U_2) < \xi_2, f_3(U_1, U_2, U_3) < \xi_3, \dots)$$
$$= \mathbb{P}\big(U_1 < \phi_1(\xi_1), U_2 < \phi_2(\xi_1, \xi_2), U_3 < \phi_3(\xi_1, \xi_2, \xi_3), \dots U_n < \phi_n(\xi)\big)$$
$$= \mathbb{P}\left(U_1 < F_1(\xi_1), U_2 < \frac{F_{1,2}(\xi_1, \xi_2)}{F_1(\xi_1)}, U_3 < \frac{F_{1,2,3}(\xi_1, \xi_2, \xi_3)}{F_1(\xi_1)\frac{F_{1,2}(\xi_1, \xi_2)}{F_1(\xi_1)}}, \dots, U_n < \phi_n(\xi)\right)$$
$$= F_1(\xi_1)\frac{F_{1,2}(\xi_1, \xi_2)}{F_1(\xi_1)}\frac{F_{1,2,3}(\xi_1, \xi_2, \xi_3)}{F_1(\xi_1)\frac{F_{1,2}(\xi_1, \xi_2)}{F_1(\xi_1)}} \cdots \frac{F_X(\xi)}{F_{1,\dots,n-1}(\xi_1, \dots, \xi_{n-1})} = F_X(\xi).$$

This completes the proof, since equality $F_X = F_{f(U)}$ holds for any ordering of $\xi_1, \ldots, \xi_n$ variables.

We complete this section by noting, that once the mapping f is found, then the simulations can be realized according to the procedure described in the end of the previous section.

10 Selecting F_X from expert's distribution and independent data

We are going to consider the problem of finding F_X. We assume that a subjective distribution F_X^S obtained by experts is given. In addition, we have an access to new observations x_n, $n = 1, \ldots, l$, which were unavailable for the experts. In order to state this problem generally, let

$$(x_1, \ldots, x_n) \to F_{x_1,\ldots,x_n}(\xi) = \frac{1}{n}\sum_{i=1}^{n} ¶_{(-\infty,\xi]}(x_i),$$

be an empirical distribution of X, computed from the first n observations.

Problem 32. Given $F_{x_1,\ldots,x_n}$ and x_{n+1}, find $F_{x_1,\ldots,x_{n+1}}$

Solution 33.

$$F_{x_1,\ldots,x_{n+1}}(\xi) = \frac{1}{n+1}\left[nF_{x_1,\ldots,x_n}(\xi) + ¶_{(-\infty,\xi]}(x_{n+1})\right].$$

Proof. Indeed

$$\frac{1}{n+1}\left[nF_{x_1,\ldots,x_n}(\xi) + ¶_{(-\infty,\xi]}(x_{n+1})\right] =$$

$$\frac{1}{n+1}\left[\sum_{i=1}^{n} ¶_{(-\infty,\xi]}(x_i) + ¶_{(-\infty,\xi]}(x_{n+1})\right] = \frac{1}{n+1}\sum_{i=1}^{n+1} ¶_{(-\infty,\xi]}(x_i) = F_{x_1,\ldots,x_{n+1}}(\xi).$$

Thus the initial problem can be solved in two steps; (1) accept F_X^S as $F_{x_1,\ldots,x_m}(\xi)$ for some m, (2) compute $F_{x_1,\ldots,x_n}(\xi)$, for $n = m+1, m+2, \ldots, m+l$, using the above formula.

There is enormous number of papers dealing with risk analysis and estimation at various levels of generality and scope. We give below only those which are closely related

Bibliography

[1] Markowitz H., *Portfolio selection*, "J. Finance", 8, 1952, 77–91.
[2] Banek T., *Rachunek ryzyka*, WSZiA w Zamościu, Lublin 2000.
[3] Federer H., *Geometric Measure Theory*, Springer, 1991.
[4] Luce R.D., Raiffa H., *Games and Decisions*, John Wiley & Sons, Inc., New York 1958.

Iwona Malinowska, Małgorzata Murat[1]

Comparative study of different ARIMA models for forecasting monthly meteorological data

Keywords: time series, meteorology, ARIMA, Fourier series, forecast, model selection,

Abstract

The aim of this article is to model and forecast weather monthly time series with different Auto Regressive Integrated Moving-Average methodology with R software. Monthly air temperature, wind speed and precipitation data were used from January 1st ,1980 to December 31st , 2010, the data is derived from measurements of Lleida (Spain) station. It will be demonstrated that obtained models are able to capture the dynamics in the data and produce sensible forecasts.

1 Introduction

In recent years a lot of different statistical models were establish and developed to predict time series. Basic approaches to construct sensible forecasts, which can capture regular pattern and dynamic of data are methods based on differentiation and methods based on decomposition. In this article we focus on methods based on differentiation such as seasonal and non-seasonal Auto Regressive Integrated Moving-Average (SARIMA, ARIMA) and Autoregressive Integrated Moving-Average with external regressors in the form of Fourier terms (ARIMAF).

In last decades, ARIMA models have been widely used for various applications such as medicine, business, economics, finance and engineering. Many scientists use ARIMA models to understand the phenomena like temperature, precipitation and wind speed. Muhammet [20] used ARIMA method to predict the temperature and precipitation in Afyonkarahisar Provincei, Turkey until the year of 2025. Balyani et al. [3] selected ARIMA as the optimal model of temperature in a 50-year time period (1955-2005) for Shiraz, south of Iran, while Babazadeh et al. [2] forecasted monthly air temperature of India using seasonal autoregressive integrated moving average (SARIMA) model. Khedhiri [15] studied the statistical properties of historical temperature data of Canada for the period (1913-

[1] Department of Mathematics, Faculty of Electrical Engineering and Computer Science, Lublin University of Technology, e-mails:i.malinowska@pollub.pl, m.murat@pollub.pl.

2013) and determined seasonal ARIMA model for the series and predicted future temperature records. El-Mallah and Elsharkawy [8] showed that linear ARIMA model and quadratic ARIMA model had the best overall performance in making short-term predictions of annual temperature in Libya. Hipel and McLeod in their book [4] showed that many different ARIMA models can be used to make sensible forecast of precipitation data. More ARIMA models for precipitation time series one can find in Said et al. [22], Khadar Babu et al. [14], Kwon et al. [17], Soltani et al. [23]. Forecasting with use of ARIMA models of wind speed is provided for example by Torres [24], Lalarukh, Yasmin [18], Cadenas, Rivera [7].

By referring to mentioned above studies for weather parameters forecasting the best model using statistical methodology could vary by changing the data. So, it is important to assess all the time series models for any area and any weather parameters for choosing the best model for our purpose. So the aim of this paper is to examine statistical properties of monthly air temperature, precipitation and wind speed from Lleida station located in Spain, develop predictive models and use them to forecast monthly values up to six years ahead. Lleida has a semi-arid climate with Mediterranean-like precipitation patterns (annual average of 369 millimetres), foggy and mild winters and hot and dry summers (Köppe-Geiger classification: BSk) and represents Mediterranean south climates. Its latitude is 41°42' (°N), longitude 10°6' (°E) and altitude 337 metres. The descriptive statistics of the studied time series are presented in Table 1.

The paper is organized as follows. In Section 2, a brief account of methods based on differentiation that is ARIMA, SARIMA and ARIMAF models are given. Methods of model selections are provided in Section 3. In Section 4 the detailed analysis of time series of monthly air temperature, precipitation and wind speed of Lleida is conducted to construct models, which will generate sensible the six year ahead forecast. Section 5 offers the concluding remarks.

Table 1. Descriptive statistics of the whole monthly 31 years' meteorological time series from Spain (ES). Mean, min, max, standard deviation (Std) and median have units corresponding to the units of meteorological variable, skewness and kurtosis are non-dimensional.

Variable	Mean	Min	Max	Std	Median	Skewness	Kurtosis
Air temperature (°C)	15.0	0.2	27.7	7.1	14.6	0.0	1.7
Wind speed (m/s)	2.6	0.8	5.5	0.7	2.5	0.6	3.8
Precipitation (mm/day)	0.9	0.0	4.2	0.9	0.6	1.3	4.4

Source: data from the ECA&D project website www.ecad.eu.

2 Methodology

ARIMA model was popularized by Box and Jenkins [4]. It is a combination of three mathematical models. It uses auto-regressive, integrated, moving-average (ARIMA) models for time series data. An ARIMA (p, d, q) model can account for temporal dependence in several ways. Firstly, the time series is deference to render it stationary, by taking d differences. If $d = 0$, the observations are modelled directly, and if $d \neq 0$, the differences between consecutive observations are modelled. Secondly, the time dependence of the stationary process $\{X_t\}$ is designed, by including p auto-regressive. The equation for p is that:

$$X_t = a + \sum_{i=1}^{p} \phi_i x_{t-i} + \varepsilon_t,$$

where a is the constant, ϕ_i is the parameter of the model, x_t is the value that observed at t and ε_t stands for random error. Thirdly, q is the moving-average terms, in addition to any time-varying covariates. It takes the observation of previous errors. The equation is

$$X_t = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i},$$

where θ_i is the parameter of the model. Finally, combining these two models we get ARMA model. So the general form of the ARMA models is given by

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-i}, \tag{1}$$

where Y_t is a stationary stochastic process, c is the constant, ε_t is the error or white noise disturbance term, ϕ_i means auto-regression coefficient and θ_j is the moving average coefficient. For a seasonal time series, these steps can be repeated according to the period of the cycle, whether time interval. Usually ARIMA models are described using the backward operator B defined as

$$B^k(X_t) = X_{t-k}, \qquad t > k;\ t, k \epsilon N.$$

Using following notation

$$\phi(z) = 1 - \sum_{i=1}^{p} \phi_i z^i, \quad \phi_p \neq 0,$$

$$\theta(z) = 1 - \textstyle\sum_{i=1}^{q} \theta_i z^q, \quad \theta_q \neq 0,$$

the relation (1) can be written respectively as

$$\phi(B)(1-B)^d y_t = c + \theta(B)\varepsilon_t.$$

The seasonal ARIMA process noted as SARIMA(p, d, q)(P, D, Q)$_m$ is given by

$$\Phi(B^m)\phi(B)(1-B^m)^D(1-B)^d y_t = c + \Theta(B^m)\theta(B)\varepsilon_t,$$

where m is the seasonal period, $\Phi(z)$ and $\Theta(z)$ are polynomials of orders P and Q respectively, each containing no roots inside the unit circle and m is a number of periods per season. If $c \neq 0$, there is an implied polynomial of order $d + D$ in the forecast function [6].

The main task in SARIMA forecasting is selecting an appropriate model order, that is the values p, q, P, Q, D, d. If d and D are known, we can select the orders p, q, P and Q via of the chosen forecast error measure. Sometimes SARIMA model does not tend to give good result for time series with a period greater than 200. In such situation the simplest approach is a regression with ARIMA errors, where the order of the ARIMA model and the number of Fourier terms is selected by minimizing the RMSE, MAE or MASE. In such models external regressors in the form of Fourier terms are added to an ARIMA(p,d,q) model to account for the seasonal behaviour. We can consider ARIMA models with repressors as a regression model which includes a correction for autocorrelated errors that is we can add ARIMA terms to the regression model to eliminate the autocorrelation and further reduce the forecast error measure. To do this we re-fit the regression model as an ARIMA(p,d,q) model with regressors, and specify the appropriate AR(p) or MA(q) terms to fit the pattern of autocorrelation we observed in the original residuals. To be more precise, we consider the following model

$$Y_t = c + \sum_{l=1}^{K}\left[\alpha_l sin\frac{2\pi lt}{m} + \beta_l cos\frac{2\pi lt}{m}\right] + N_t,$$

where N_t is an ARIMA process, α_l and β_l are Fourier coefficients and m is a length of period. The value of K is chosen by minimizing forecast error measures. For the purpose of this article this process will be noted as ARIMF(p,d,q)[K]. According to Hyndman [13] the main advantages of this approach are: it allows any length seasonality for data with more than one seasonal period, Fourier terms of different frequencies can be included, the seasonal pattern is smooth for small values of K and the short-term dynamics are easily handled with a simple ARMA error. The only real disadvantage (compared to a seasonal ARIMA model) is that the seasonality is assumed to be fixed (the pattern is not allowed to change over time), but in our situation, seasonality is remarkably constant (compare Figure 1).

3 Model selection

To improve the model accuracy wide range of error measures is used in addition to ACF and PACF. Commonly used measure are mean absolute error (MAE) and root mean square error (RMSE). MAE is defined by the formula

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|e_t|,$$

where n is the number of periods of time and $e_t = y_t - f_t$ is the forecast error between the actual value y_t and the forecasted value f_t. The MAE is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. RMSE given as follows

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}e_t^2}$$

is a next frequently used measure of accuracy forecast. The RMSE is the square root of the average the squared values of the differences between forecast and the corresponding observation. MAE and RMSE have the same units of measurement and depend on the units in which the data are measured.

Hyndman and Koehler [11] proposed the mean absolute scaled error (MASE) to comparing forecast accuracy. Their idea that is suitable in all situations is by scaling the error based on the in-sample MAE from the naive (random walk) forecast method. Using the naive method, the one-period-ahead forecasts is generated from each data point in the sample. Thus, a scaled error is defined as

$$q_t = \frac{e_t}{\frac{1}{N-1}\sum_{j=2}^{N}|y_j - y_{j-1}|},$$

where N is the length of the training data set. For one-step-ahead forecasts MASE is calculated as

$$MASE = \frac{1}{m}\sum_{t=1}^{m}|q_t|,$$

where m is the number of one-step-ahead forecasts. For h-step-ahead forecasts we calculate MASE as

$$MASE = \frac{1}{h}\sum_{t=1}^{h}|q_t|.$$

When MASE < 1, the proposed method gives, on average, smaller errors than the one-step-ahead forecast errors from the naive method. Conversely, it is greater than one, if the forecast is worse than the average naive forecast computed on the training data. The only circumstance under which this measure would be infinite or undefined is when all historical observations are equal (see Hyndman R. J. [10] and Hyndman R. J., Koehler A. B. [11]). The MASE is independent of the scale of the data, so can be used to compare forecasts for data sets with different scales. When comparing forecasting methods, the method with the lowest MASE is the preferred method.

After choosing the adequacy model, the accuracy of the model will be determined by looking at 3 diagnostic methods which are: standardized residuals, ACF of residuals and p-values for Ljung-Box statistics. The standardized residuals must be stationary (the variance near to zero), ACF of residuals has no spikes, the Ljung-Box p-values must be above 0.05. The Ljung-Box test was proposed by Ljung and Box [19] and is based on the statistic

$$Q^* = T(T+2)\sum_{k=1}^{h}\frac{r_k^2}{T-k},$$

where T is the length of the time series, r_k is the k-th autocorrelation coefficient of the residuals, and h is the number of lags to test. Large values of Q^* indicate that there are significant autocorrelations in the residual series. It can be tested against a χ^2 distribution with $h - K$ degrees of freedom, where K is the number of parameters estimated in the model. This test is a diagnostic tool used to test the lack of fit a time series model and is applied to the residuals of time series after fitting on model to the data.

4 Analysis and results

A good fitting of the model with the historical data does not necessarily mean good forecasting. This problem can be overcome by measuring true out of sample forecast accuracy. For this purpose the total data are divided into a learning set and a test set. Then, the learning set is used to estimate parameters of a model and the test set is used to assess the predictability accuracy of the fit. In our project, the collected 372 months from January 1980 to December 2010 data set was divided into a set of observations from January 1980 to December 2004 – the learning set and a set of observations from January 2005 to December 2010 – the test set. The learning set was only used in the model fitting, so obtained forecasts are genuine forecast made without using the values of the observations belonging to the test set and the accuracy measures are computed on the basis of the test set only.

To detect possible presence of seasonality and trend we inspect the plots of the observed data side by side with the plots of autocorrelation function (ACF)

and partial autocorrelation function (PACF) with addition of time series decomposition into its constituent components, which are: usually trend component, irregular component, and if it is a seasonal time series, a seasonal component. The visual analysis ACF and PACF plots in Figure 1 suggest that the air temperature and wind speed Lleida time series show seasonal character. We observed a slow decay of the ACF and PACF at multiple lags of 12, which are significant. Quite different situation takes place in case of precipitation data, where time series courses, ACF and PACF plots do not indicate any seasonal character and any trend.

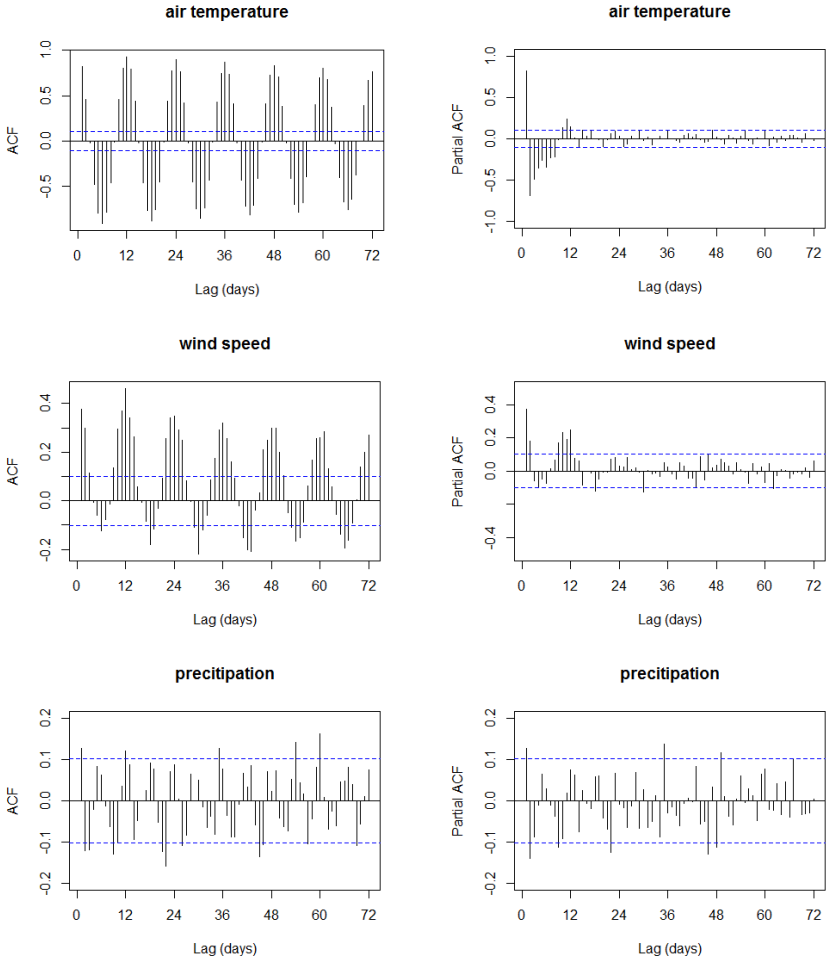


Figure 1. Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots for time series from Lleida stations.

Source: data from the ECA&D project website www.ecad.eu, plots prepared with the R software.

Additionally, the air temperature time series decomposition into its constituent components given in Figure 2 indicates regular fluctuations, which are repeated from year to year with about the same timing and lever intensity. The same behaviour one can observe for the wind speed time series decomposition in Figure 3. This analysis supports the above assertion of seasonality in the data, hence, the need for seasonal differencing with period of 12. The decomposition of the precipitation time series in Figure 4 does not show any regular changes in seasonal component.

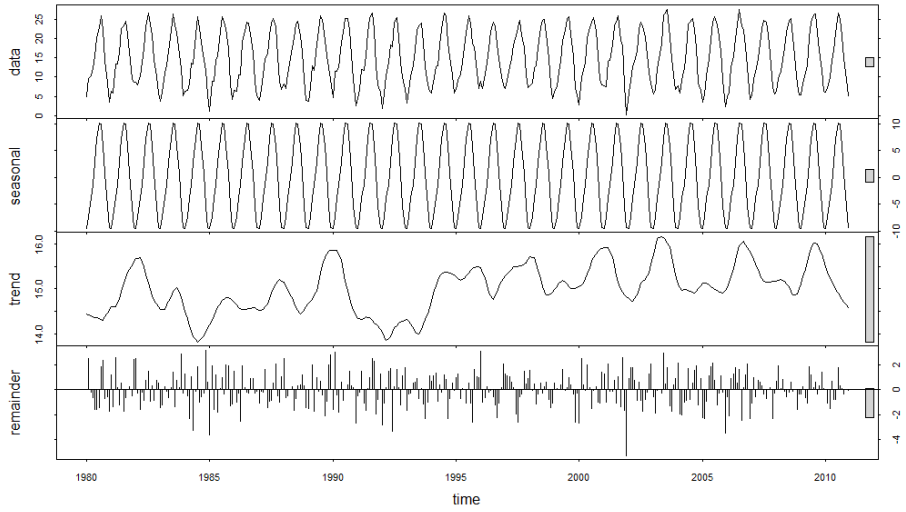


Figure 2. Air temperature time series graphs with random, seasonal and trend components in Lleida station)

Source: data from the ECA&D project website www.ecad.eu, plots prepared with the R software.

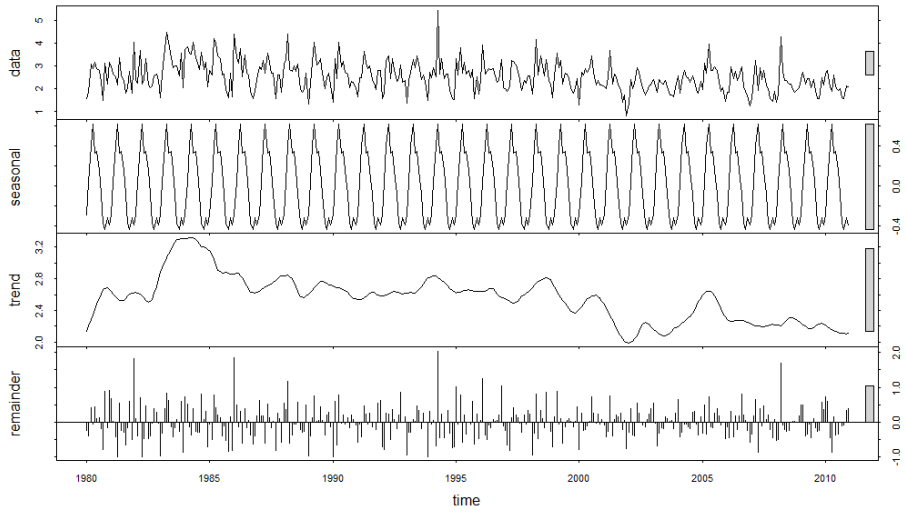


Figure 3. Wind seed time series graphs with random, seasonal and trend components in Lleida station.

Source: data from the ECA&D project website www.ecad.eu, plots prepared with the R software.

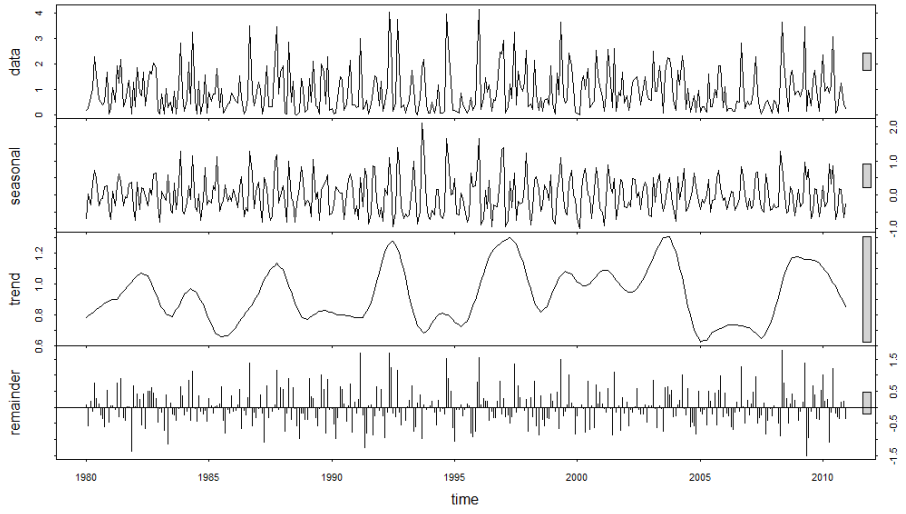


Figure 4. Precipitation time series graphs with random, seasonal and trend components in Lleida station.

Source: data from the ECA&D project website www.ecad.eu, plots prepared with the R software.

This investigation in plots, ACF, PACF and decomposition of considered time series force us to consider SARIMA $(p,d,q)(P,D,Q)_m$ models with m=12, D=1, Q=0, P=0, d=1 or d=0, and p, q changing from 0 to 3. The best parameters among considered 32 SARIMA models for air temperature and wind speed were chosen by minimizing the forecast RMSE and MASE. In order to establish ARIMAF models parameters we tried d=0 and d=1, p and q between 0 and 3 while the number of Fourier terms K varied between 1 and 10. So for both of considered as seasonal series we tested 320 cases. The models which are the most closed to the actual data were chosen by minimizing forecast RMSE and MASE. The best model parameters with their forecast errors are presented in Table 2.

The precipitation data are modelled with ARIMA models. We chose the best parameters by minimizing forecast RMSE and MASE among 32 different models considering p, q changing from 0 to 3 and d equal 0 or 1. Obtained models with their forecast errors are presented also in Table 2.

One can observe that for air temperature and wind speed time series the smallest forecast RMSE and MASE were produced by SARIMA model with the same parameters, while selection of ARIMAF parameters for the best forecast depends on assumed error.

Table 2. The statistical models parameters and errors forecast for chosen models with the smallest RMSE and MASE.

	Model	MAE	RMSE	MASE
Air temperature	SARIMA(2,0,3)$(0,1,0)_{12}$	1.4290	**1.8237**	**0.8907**
	ARIMAF(3,1,3)[5]	1.0929	**1.3572**	0.6812
	ARIMAF(1,1,0)[3]	1.0869	1.3638	**0.6775**
Wind speed	SARIMA(3,0,3)$(0,1,0)_{12}$	0.4039	**0.5015**	**0.7595**
	ARIMAF(1,1,3)[3]	0.3438	**0.4518**	0.6465
	ARIMAF(3,1,3)[3]	0.3435	0.4550	**0.6459**
Precipitation	ARIMA(3,0,3)	0.5998	**0.7908**	**0.7101**

Source: data from the ECA&D project website www.ecad.eu, own calculations.

.All models were diagnosed by Ljung-Box test. The p-values is greater than the usually chosen critical level of 0.05 (see Table 3) except ARIMAF(1,1,0)[3] model. Then ARIMAF(1,1,0)[3] model will not be utilized to perform air temperature forecasting. The Ljung-Box test applied to other models from Table 2 is no significant and therefore we do not reject the null hypothesis in all cases. This indicates, that the residuals of those fitted models are white noise, and for that reason the models fit the series quite well, the parameters of the models are significant and the residuals are uncorrelated.

Table 3. p-values for Ljung-Box test of chosen models.

	Model	p-value for Ljung-Box test
Air temperature	SARIMA(2,0,3)$(0,1,0)_{12}$	0.9426
	ARIMAF(3,1,3)[5]	0.5075
	ARIMAF(1,1,0)[3]	**0.0130**
Wind speed	SARIMA(3,0,3)$(0,1,0)_{12}$	0.8572
	ARIMAF(1,1,3)[3]	0.9821
	ARIMAF(3,1,3)[3]	0.9983
Precipitation	ARIMA(3,0,3)	0.4120

Source: data from the ECA&D project website www.ecad.eu, own calculations.

Moreover, residuals appear to be randomly, scattered, no evidence exists that the error terms are correlated with one another as well as no evidence of existence of an outlier, what is shown in their plots and ACF plots. In Figures 5a, 5b and 5c there are plots of residuals and ACF residuals plots for models with the smallest RMSE. Thus the residuals plots and its ACF plots collaborate the conclusion of the Ljung-Box test and we can use models listed in Table 2, except ARIMAF(1,1,0)[3], to make forecasts.

ARIMAF(3,1,3)[5]

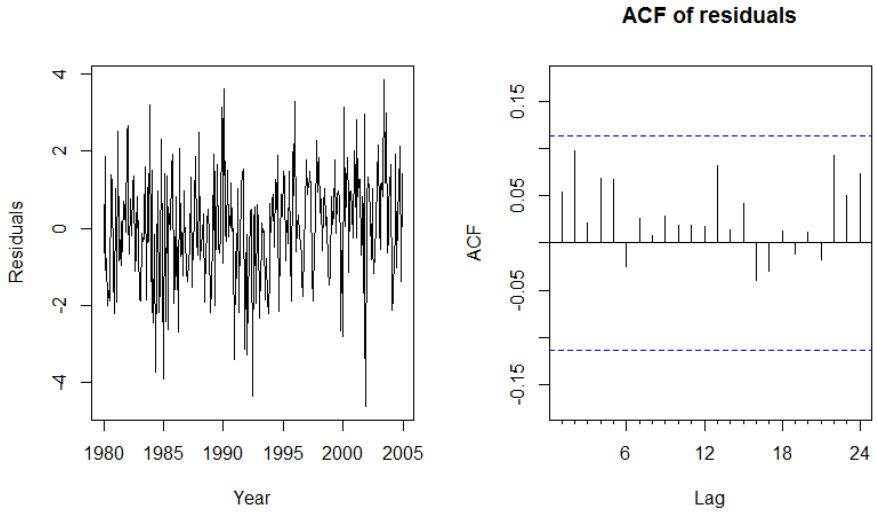


Figure 5a. Plots of residuals and their ACF plots for models with smallest RMSE – ARIMAF(3,1,3)[5].

Source: data from the ECA&D project website www.ecad.eu, plots prepared with the R software.

ARIMAF(1,1,3)[3]

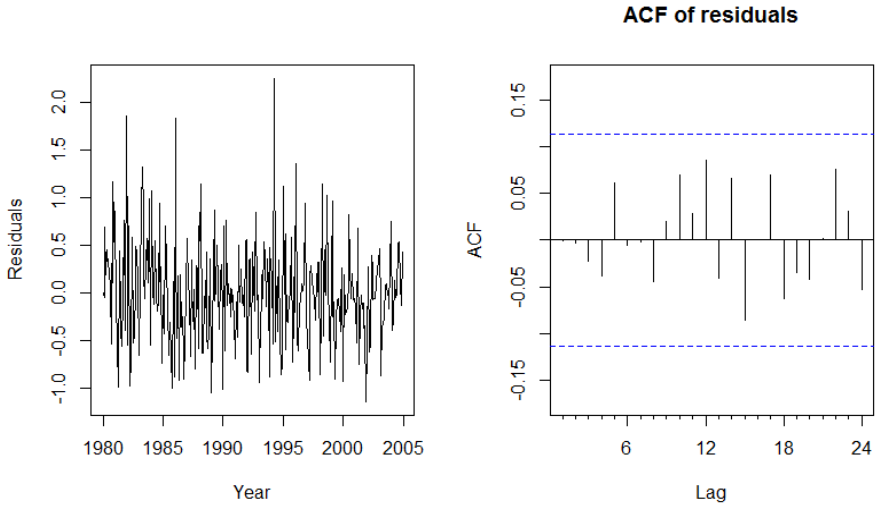


Figure 5b. Plots of residuals and their ACF plots for models with smallest RMSE – ARIMAF(1,1,3)[3].

Source: data from the ECA&D project website www.ecad.eu, plots prepared with the R software.

ARIMA(3,0,3)

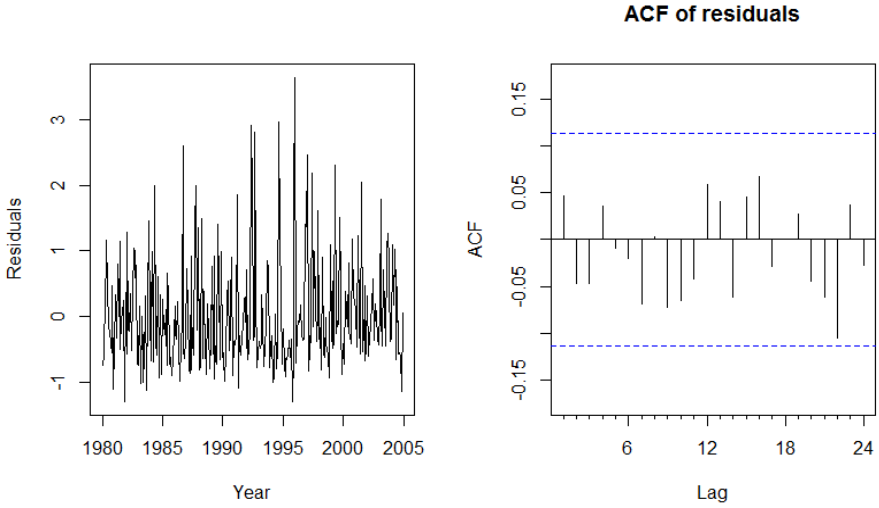


Figure 5c. Plots of residuals and their ACF plots for models with smallest RMSE – ARIMA(3,0,3).

Source: data from the ECA&D project website www.ecad.eu, plots prepared with the R software.

Figure 6 compares the results obtained the SARIMA(2,0,3)(0,1,0)$_{12}$ and ARIMAF(3,1,3)[5] models with the real data for air temperature in Lleida. In this figure it can be observed how both models follow the same tendency of real data, however a higher difference between the real data and data predicted with SARIMA model is noticed. This observation is corroborated by calculated forecast errors measure presented in Table 2.

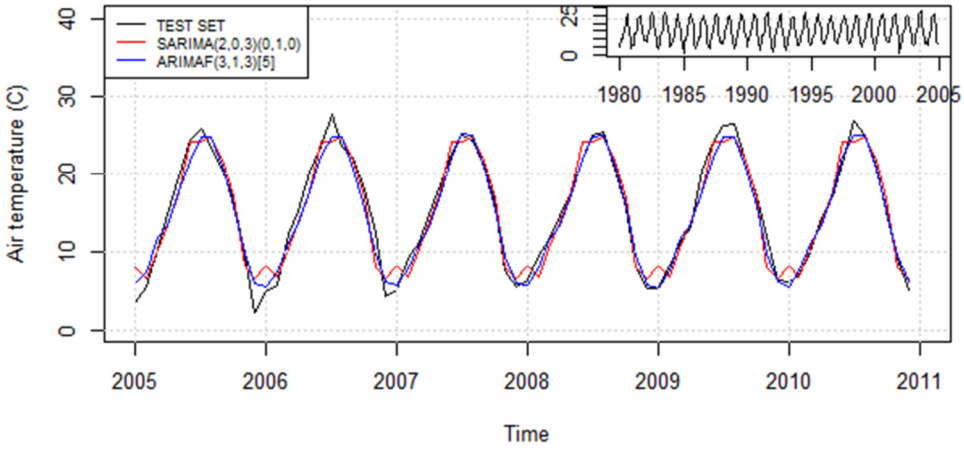


Figure 6. Real data and forecast plots for monthly air mean temperature time series from Lleida. The larger plots contain test set and forecast whereas the smaller inside plots presents learn set.

Source: data from the ECA&D project website www.ecad.eu, plots prepared with the R software.

Forecast results using SARIMA and ARIMAF models for monthly wind speed are presented in Figure 7. From Table 2 we can see that the RMSE errors of ARIMAF(1,1,3)[3] are smaller compared with SARIMA(3,0,3)(0,1,0)$_{12}$ model, showing that ARIMAF(1,1,3)[3] model is identified as the best fitted time series model for wind speed and is the best way of representing the observed wind speed pattern.

The statistical ARIMA model structured as (3, 0, 3) appeared to be the best fitting for the precipitation forecasting purpose at Lleida. The forecast and test set are plotted and compared in Figure 8.

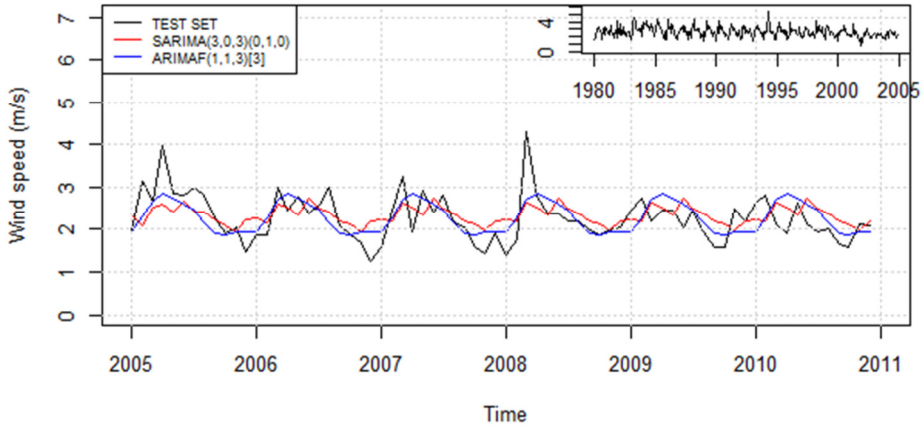


Figure 7. Real data and forecast plots for monthly wind speed time series from Lleida. The larger plots contain test set and forecast whereas the smaller inside plots presents learn set.

Source: data from the ECA&D project website www.ecad.eu, plots prepared with the R software.

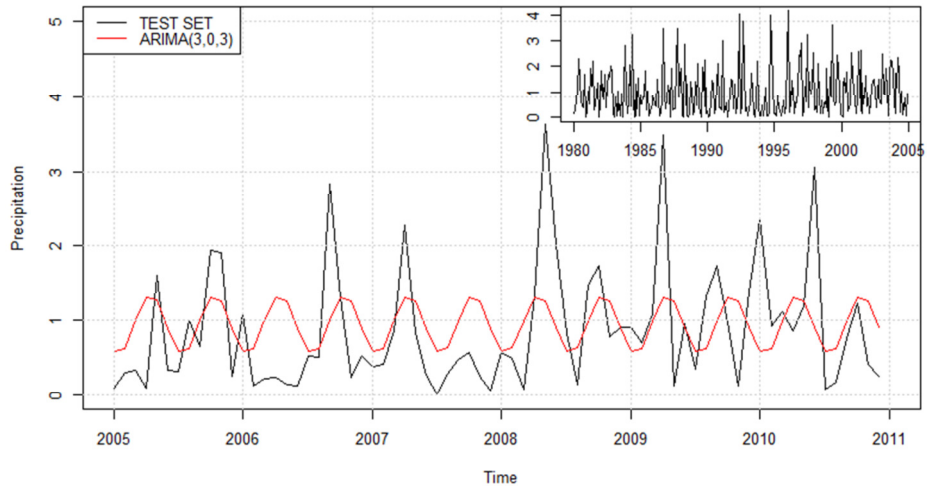


Figure 8. Real data and forecast plots for monthly precipitation time series from Lleida. The larger plots contain test set and forecast whereas the smaller inside plots presents learn set.

Source: data from the ECA&D project website www.ecad.eu, plots prepared with the R software.

5 Conclusions

The dynamics of meteorological time series for Lleida (Spain) station located in Mediterranean south climate zone were analysed via three ARMA models. The results indicated that the studied meteorological quantities possess specific time and space dynamics, which can be attributed to climatic conditions. Above studies for weather parameters forecasting show that the best model using statis-

tical methodology could vary by changing the forecast error. We observed that for air temperature and wind speed time series selection of parameters of ARIMAF model were depended of assumed forecast error, while the smallest forecast RMSE and MASE were created by SARIMA model with the same model parameters. It mean that for obtaining a reasonable and sensible forecast, more than one forecast error measure should be used in practice. The empirical study of two seasonal real data sets highlights the importance of considering the seasonality in forecasting of air temperature and wind speed in Lleida, contrasting to forecasting precipitation. In addition, the conducted research clearly suggests that ARIMAF models gives better prediction for seasonal time series. The result indicate, that the use of ARIMA models to weather time series analysis is a valuable tool to get information about analysed data structures and their components, being a good basis for successful future forecast.

Acknowledgements

We acknowledge the data providers in the ECA&D project – for Lleida, the Agencia Estatal de Meteorología (AEMET) (Klein Tank et al. 2002). Data and metadata are available at www.ecad.eu

Bibliography

[1] Anitha K., Boiroju N. K., Reddy P. R., *Forecasting of monthly mean of maximum surface air temperature in India*, "International Journal of Statistika and Mathematika", 9(1), 14–19, 2014.

[2] Babazadeh H., Shamsina S. A., *Modeling climate variables using time series analysis in arid and semi-arid regions*, "African Journal of Agricultural Research", 9(26), 2018–2027, 2014.

[3] Balyani Y., Niya G. F., Bayaat A., *A study and prediction of annual temperature in Shiraz using ARIMA model*, "J. of Geographic Space", 12, 2014.

[4] Box G., and Jenkins G., *Time Series Analysis: forecasting and control (1st ed.)*, Holden-Day, San Francisco 1970.

[5] Box G. E. P. and Tiao G. C., *Intervention Analysis with Applications to Economic and Environmental Problems*, "JASA", 70, 70–79, 1975.

[6] Box G., Jenkins G. and Reinsel G., *Time series analysis (4th ed.*), Wiley, New Jersey. 2008.

[7] Cadenas E., Rivera W., *Wind speed forecasting in three different regions of Mexico, using a hybrid ARIMA-ANN model*, "Renewable Energy", 35, 2732–2738, 2010.

[8] El-Mallah E. S. and Elsharkawy S. G., *Time-Series Modeling and Short Term Prediction of Annual Temperature Trend on Coast Libya Using the Box-Jenkins ARIMA Model*, "Advances in Research", 6(5), 1–11, 2016.

[9] Hipel K. W., McLeod A. E., *Time series modeling of water resources and environmental system*s, Elsevier, Amsterdam, 1994.

[10] Hyndman R. J., *Another look at forecast-accuracy metrics for intermittent demand. Foresight*, "The Int. J. of Applied Forecasting",4(4), 43–46, 2006.
[11] Hyndman R. J., Koehler A. B., *Another look at measures of forecast accuracy*, "International Journal of Forecasting", 22(4), 679–688, 2006.
[12] Hyndman R., Koehler A., Ord J. and Snyder R., *Forecasting with Exponential Smoothing: The State Space Approach*, Springer-Verlag, Berlin, 2008.
[13] Hyndman R., *Forecasting with long seasonal periods* from http://robjhyndman.com/hyndsight/longseasonality/, 2010.
[14] Khadar Babu S. K., Karthikeyan K., Ramanaiah M. V., Ramanah D., *Prediction of rain-fall flow time series using auto-regressive model,* Adv Appl Sci Res 2(2), 128–133, 2011.
[15] Khedhiri S., *Forcasting temperature record in PEI*, "Canada Lett. Spat Resource Sci.", 2014.
[16] Klein Tank A. M. G, Wijngaard J.B., Können G. P, Böhm R and others, *Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment*, "Int. J. Climatol." 22, 1441−1453, 2002.
[17] Kwon H. H., Lall U., Khalil A. F., *Stochastic simulation model for nonstationary time series using an autoregressive wavelet decomposition: applications to rainfall and temperature*, "Water Resour. Res.", 43(5), 1–15, 2007.
[18] Lalarukh K, Yasmin Z. J., *Time series models to simulate and forecast hourly averaged wind speed in Quetta, Pakistan*, "Solar Energy", 61(1), 23–32, 1997.
[19] Ljung; G. M. and Box G. E. P., *On a Measure of a Lack of Fit in Time Series Models*, "Biometrika", 65(2), 297–303, 1978.
[20] Muhammet B., *The analyse of precipitation and temperature in Afyonkarahisar (Turkey) in respect of Box-Jenkins technique*, "The Journal of Academic Social Science Studies", 5(8), 196–212, 2012.
[21] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org., 2014.
[22] Said S. M., Manjang S., Tjaronge M. W., Thaha M. A., *Arima Application as an Alternative Method of Rainfall Forecasts In Watershed Of Hydro Power Plant*, "International Journal of Computational Engineering Research", 3(9), 68–73, 2013.
[23] Soltani S., Modarres R., Eslamian S. S., *Arima Application as an Alternative Method of Rainfall Forecasts In Watershed Of Hydro Power Plant*, "Int. J. Climatol.", 27, 819–829, 2007.
[24] Torres J.L., García A., De Blas M., De Francisco A., *Forecast of hourly average wind speed with ARMA models in Navarre*, "Solar Energy", 79, 65–77, 2005.

Ewa Łazuka[1]

Construction of an optimal bonus-malus system

Keywords: motor insurance, insurance premium, bonus-malus system.

Abstract

The paper is devoted to the construction of an optimal bonus-malus system which will provide the insurer with financial balance in the portfolio and will be fair for the insureds. The first part describes a model of number of claims based on the negative binomial distribution. The next section presents the test used to compare the empirical and the theoretical probability distributions. The following part covers methods of estimating net premiums. Another section is devoted to the Bayesian methods. In the last part of the work, which is based on survey data, calculations that determined the value of premium in the optimal bonus-malus system were carried out.

1 Introduction

The bonus-malus system is one of the elements of tariffication in motor insurance which makes the insurance premiums dependent on the current history of insurance. The most common factor taken into consideration is the number of claims in the previous year of insurance. The insured without the history of claims is classified to the basic class and then in subsequent periods of insurance they move to a specific tariff class in accordance to the number of claims. Despite criticism of the bonus-malus systems, they are widely used around the world and are an important element of tariffication.

The aim of this work is to construct an optimal bonus-malus system for a sample obtained from a survey carried out among car owners. In the survey the respondents were asked to answer three questions about their reported damages (caused by their own cars) in 2016 and their impact on the increase of their insurance premiums in the following year.

To describe number of claims in the portfolio of the insureds, a model based on the negative binomial distribution was used. The article covers comparing the theoretical and empirical distributions, using the χ^2 test as the test of goodness of fit. Further part of the work is a description of methods of estimating motor

[1] Department of Applied Mathematics, Faculty of Fundamentals of Technology, Lublin University of Technology, e-mail: e.lazuka@pollub.pl.

insurance premiums, with emphasis on *a posteriori* tariffs. The problem of Bayesian methods was discussed, focusing on the method of estimating premium rates in bonus-malus systems. The Bayesian premium rate estimator using the quadratic loss function and based on the number of claims and the equivalent net premium principle was reviewed. In the last part of the article an optimal bonus-malus system was designed. Presenting estimator values, parameters a and b, expected sample sizes, χ^2 test and the premium rate was possible thanks to the use of Excel spreadsheet.

2 Model describing the process of occurrence of claims

In motor insurance risk models it is assumed that the number of claims in a given unit of time is a discrete random variable and that the distribution of claims in the portfolio for each insured is of the same type. According to the actuarial literature, Poisson distribution and the negative binomial distribution are most commonly used in claims distribution models. The Poisson model assumes that the risk portfolio is homogeneous. However, in the real world, the risk portfolio is usually heterogeneous. Therefore, we will use the model based on the negative binomial distribution to describe the number of claims in the insureds' portfolio.

The negative binomial distribution is a mixture of Poisson distribution and gamma distribution (see e.g. [6]). Let the random variable N have the Poisson distribution with the parameter $\lambda > 0$ described by the equation:

$$P(N = k) = f_k(\lambda) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0,1, \ldots \tag{1}$$

Let λ be the realization of the random variable Λ, which has the gamma distribution with parameters (a, b) and the density function described by the equation:

$$g(\lambda) = \begin{cases} \dfrac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} & \text{for } \lambda > 0, \\ & \text{for } \lambda \leq 0, \end{cases} \tag{2}$$

where $a > 0$ and $b > 0$. As a result of the composition of Poisson distribution with gamma distribution, we obtain the distribution of number of claims described by the random variable X with the probability distribution defined as follows:

$$\begin{aligned} P(X = k) = p_k &= \int_0^\infty f_k(\lambda) g(\lambda) d\lambda = \int_0^\infty \frac{\lambda^k}{k!} e^{-\lambda} \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} d\lambda = \\ &= \frac{b^a}{\Gamma(a)k!} \int_0^\infty \lambda^{a+k-1} e^{-(b+1)\lambda} d\lambda = \\ &= \frac{b^a}{\Gamma(a)k!} \frac{\Gamma(a+k)}{(1+b)^{a+k}} \int_0^\infty \frac{(1+b)^{a+k} \lambda^{a+k-1} e^{-(1+b)\lambda}}{\Gamma(a+k)} d\lambda. \end{aligned}$$

The last integral's value is 1 since it is the intergal of gamma distribution density function with parameters $(a+k, 1+b)$. Hence:

$$P(X=k)=p_k = \frac{b^a}{\Gamma(a)k!}\frac{\Gamma(a+k)}{(1+b)^{a+k}} = a+k-1_k\ b^a\left(\frac{1}{1+b}\right)^{a+k} = \\ = a+k-1_k\left(\frac{b}{1+b}\right)^a\left(\frac{1}{1+b}\right)^k. \tag{3}$$

This is the formula of probability of the random variable X with the negative binomial distribution (see e.g. [10]). Assuming that $p=\frac{b}{1+b}$ and $q=\frac{1}{1+b}$, we obtain:

$$P(X=k)=p_k=\binom{a+k-1}{k}p^a q^k, \tag{4}$$

where

$$\binom{a+k-1}{k}=\frac{\Gamma(a+k)}{\Gamma(a)\Gamma(k+1)}=\frac{\Gamma(a+k)}{\Gamma(a)k!}.$$

The expected value and variance of the negative binomial distribution with parameters (a,q) described by equations (3) and (4) are respectively:

$$EX=\frac{a(1-p)}{p}=\frac{a}{b}, \qquad D^2X=\frac{a(1-p)}{p^2}=\frac{a}{b}\left(1+\frac{1}{b}\right). \tag{5}$$

The moment generating function of X is given by the equation:

$$M_X(t)=\left(\frac{p}{1-(1-p)e^t}\right)^a, \qquad t<-\ln(1-p). \tag{6}$$

The most common distribution used for describing the number of claims in heterogeneous portfolios is the negative binomial distribution. It is used to model the claims distribution, as $EX<D^2X$. The greater the difference between the expected value and the variance, the greater is the heterogeneity of the risk in the portfolio.

3 Adjusting the theoretical distribution to the empirical distribution of the number and value of claims

The initial choice of theoretical claims distribution can be based on the calculated sample moments and the frequency coefficients. Let $X_1, X_2, \ldots, X_n$ be a simple sample of independent random variables that have the same discrete distribution and let N_k be the number of observations X_i in the sample, for which $X_i=k$, where $k=0,1,2,\ldots$. Then, the r-th sample raw moments have the form:

$$M_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r, \qquad r = 1,2,3,\ldots \tag{7}$$

and for aggregated data:

$$M_r = \frac{1}{n}\sum_{k=0}^{m-1} k^r N_k, \qquad r = 1,2,3,\ldots \tag{8}$$

where $n = \sum_{k=0}^{m-1} N_k$ and m is the number of classes.
First three sample central moments can be expressed in terms of the raw moments:

- $\bar{X} = M_1$,
- $S^2 = M_2 - M_1^2$,
- $C_3 = M_3 - 3M_2M_1 + 2M_1^3$.

Frequency coefficients are described by the equation:

$$T_k = (k+1)\frac{N_{k+1}}{N_k}, \quad k = 0,1,2,\ldots. \tag{9}$$

In the initial stage of research, the choice of theoretical distribution of claims is usually reduced to the evaluation of adjusting the empirical data to one of the distributions from the so-called $(a, b, 0)$ class. These distributions are discrete and their probabilities $p_k = P(X = k)$ fulfill the recurrence relation:

$$p_k = \left(a + \frac{b}{k}\right)p_{k-1}, \quad k = 1,2,3,\ldots \tag{10}$$

The only distributions belonging to this class are:

- Poisson distribution with parameter $\lambda > 0$, for which:

$$(a, b) = (0, \lambda),$$

- binomial distribution with parameters (l, q) where $l \in \mathbb{Z}_+$, $q \in (0,1)$, for which:

$$(a, b) = \left(-\frac{q}{p}, (l+1)\frac{q}{p}\right),$$

- negative binomial distribution with parameters (r, q) where $r > 0$, $q \in (0,1)$, for which:

$$(a, b) = (q, (r-1)q).$$

For distributions from the $(a, b, 0)$ class the function T_k may be written as a linear function:

$$T_k \approx (k+1)\frac{p_{k+1}}{p_k} = ak + (a+b), \quad k = 0,1,2,\ldots. \tag{11}$$

Function T_k describes a line of slope a.

- If $a = 0$ and $\bar{X} = S^2$, then the number of claims distribution may be the Poisson distribution,
- If $a < 0$ and $\bar{X} > S^2$, then the number of claims distribution can be modelled by the binomial distribution,
- If $a > 0$ and $\bar{X} < S^2$, then the number of claims distribution may be the negative binomial distribution.

If the function described by the equation (9) increases faster than linear, the skewness of the distribution shall be considered. Let $T = 3S^2 - 2\bar{X} + 2\frac{(S^2-\bar{X})^2}{\bar{X}}$.

- If $C_3 = T$, then the negative binomial distribution should describe the number of claims in the portfolio well.
- If $C_3 < T$, then the generalized Poisson-Pascal distribution or the Poisson-Inverse Gaussian can be used to model the distribution of number of claims.
- If $C_3 > T$, then Neyman type A, Polya-Aeppli, Poisson-Pascal or negative binomial distribution can be used (see e.g. [12]).

4 Pearson's χ^2 test

Verifying the goodness of fit of theoretical and empirical distributions is carried out by means of so-called tests of goodness of fit. They help to verify the null hypothesis, which assumes that the analysed random variable has a distribution belonging to a certain distribution family.

Definition 3.1. ([2]) The test of goodness of fit is used to verify a simple or complex hypothesis concerning the fit between the distribution of the set of values in the sample and the theoretical distribution, i.e. a hypothesis of the form: $H_0: F \in \mathcal{F}$ where F is the cumulative distribution of the studied feature in the population and $\mathcal{F}$ is a certain class of distributions.

The test most commonly used to verify the goodness of fit of theoretical and empirical distributions is the Pearson's χ^2 test (see e.g. [11]). It can only be used when the following assumptions are satisfied:

- number of classes k is not less than 5,
- sample size $n \geq 10k$,
- all empirical numbers are always greater than 5 (if the class size does not meet the condition, the adjacent classes should be merged).

For a large sample $n \geq 200$ we set the number of classes according to Table 1.

Table 1. Number of classes for χ^2 test

Number of observations	Number of classes
<400	15-19
401-600	20-24
601-800	25-27
801-1000	27-30
1001-1500	30-35
1501-2000	35-40

Source: own elaboration.

Let a studied feature of the population have cumulative distribution F (continuous or not). Given a large sample $(X_1, X_2, \dots, X_n)$ randomly and independently drawn from the population, on the basis of the results from it, it is necessary to verify the hypothesis that the feature has a distribution of type $\mathcal{F}$, i.e. $H_0: F \in \mathcal{F}$ where $\mathcal{F}$ is a certain class of distributions. The alternative hypothesis takes the form $H_1: F \notin \mathcal{F}$.

Let n denote the size of a random sample from the population. In order to verify the hypothesis H_0, test results should be divided into k mutually disjoint classes of sizes n_i, where $n_1 + n_2 + \cdots + n_k = n$. Thanks to the properties of the theoretical distribution we can calculate the probabilities p_i that the analyzed random variable with the given distribution takes values belonging to i-th class $(i = 1,2, \dots, k)$. The probabilities p_i must satisfy the condition $p_i = \sum_{i=1}^{k} p_i = 1$. By multiplying p_i by n we obtain the expected numbers np_i.

As a measure of divergence between the observed numbers $n_1, n_2, \dots, n_k$ and expected numbers $np_1, np_2, \dots, np_k$ we use the statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{n_i^2}{np_i} - n. \tag{13}$$

Assuming the truth of hypothesis H_0, the statistic defined by the formula (12) has an asymptotic distribution χ^2 with $k - l - 1$ degrees of freedom, where l is the number of parameters of the theoretical distribution. Because of the asymptotic distribution of the statistic, this test can be used when the sample is large (see e.g. [2]).

The value of χ^2 is 0 if and only if all the observed numbers are equal to expected numbers. Thus, the greater the divergence, the greater the value of χ^2 statistic. We reject the null hypothesis if the value of the χ^2 statistic calculated from the sample belongs to the critical area determined by our assumed significance level α.

5 Methods of estimating the premium rate in motor insurance

The premium is a fee that an individual exposed to risk pays to the insurer for taking over some of the risk involved in their activity. The premium is calculated by the insurer according to a principle that random aggregate claims of the insured clients in a given portfolio shall be compensated by pre-determined fees (premiums).

In the process of calculating premiums, three golden rules of insurance must be taken into account (see e.g. [7]). The first principle determines the need for balance between the insurance fund and the benefits and compensations paid by the insurance company. The second principle postulates the necessity to maintain a proportional relationship between the premium and the sum insured. The higher the sum insured, the higher the insurance premium. The third rule of equivalence of premiums and benefits defines the need to maintain an adequate relationship between the premium and degree of risk, which means that premiums shall be differentiated in different types of insurance.

A key part of an insurance premium is the net premium that covers the compensation and benefits. The net premium plus other expenses (administrative costs, insurer's profit, taxes etc.) is the gross premium. The net premium is calculated on the basis of the anticipated number and size of claims and then increased by so-called security surcharge to cover any fluctuations in loss ratio over time.

We denote by $\Pi(X)$ the premium that the insurer charges to cover a risk X. When we refer to the risk X, what we mean is that claims from this risk are distributed as the random variable X. In the individual risk model, as well as in the collective risk model, the random variable of interest S denotes the total claims on a portfolio of insurance contracts. In the individual risk model S is modelled as the sum of all claims on the policies, which are assumed independent. The model that is often used to approximate the individual model is the collective risk model. In this model, an insurance portfolio is viewed as a process that produces claims over time. We calculate the distribution of the total claim amount in a certain time period, but now we regard the portfolio as a collective that produces claims at random points in time. We write $S = X_1 + X_2 + \cdots + X_K$, where K denotes the number of claims and X_i is the i-th claim, and by convention, we take $S = 0$ if $K = 0$. The number of claims K is a random variable, and we assume that the individual claims X_i are independent and identically distributed. We also assume that K and X_i are independent.

Basic methods of calculating the net premium are: equivalence of premium principle (pure risk premium principle), expected value principle and standard deviation principle (see e.g. [3]). The choice of a premium principle depends heavily on the importance attached to its properties. There is no premium principle which is uniformly best.

Equivalence of premium principle is based on the assumption of balance of collected net premiums and the expected amount of compensation. Therefore:

$$\Pi(S) = E(S). \tag{13}$$

This principle is applied in homogeneous non-life insurance groups, in which we assume that:

- the probability of an insurance case for every risk in the portfolio is the same,
- every risk in the portfolio is insured for the same amount,
- every insurance case causes total or partial damage,
- it is possible to specify the intensity indicator, which shows how much of the insured sum represents the insurance claim in a given risk group.

The equivalence of net premium principle in the long run leads to the insurer's ruin and therefore it is modified by introducing a safety factor.

Expected value principle includes the safety factor introduced due to the uncertainty of the premium calculation, therefore:

$$\Pi(S) = (1 + \theta)E(S). \tag{14}$$

Parameter $\theta > 0$ called the safety factor is determined by the insurer for each risk respectively. The expected value principle does not include the variability of the random variable, so its subsequent modifications were designed.

Standard deviation principle corrects premiums by factors that include the variability of the random variable S, therefore:

$$\Pi(S) = E(S) + \beta\sigma(S), \beta > 0 \tag{15}$$

In very large portfolios and with full risk information, the equivalence of premium principle should be applied. The insurer's goal should be to determine the premium that is adequate to the risks represented by an individual insured and that ensures the solvency of the insurance company. Too high premium could result in loss of customers and consequently lead to the insurer's ruin.

6 The essence of the bonus-malus system

Determining premium rates in motor insurance when the bonus-malus system is applied can be divided into two stages. First, for each insured a basic premium is determined by including them in a specific tariff group. This is the *a priori* pricing, which includes factors describing the driver and the vehicle details. However, there are some individual driver's characteristics that we cannot determine *a priori*, such as: reflexes, law compliance, knowledge of traffic regulations, stress behaviour or influence of alcohol. Those factors have a significant impact on the number of accidents. Therefore, the second stage is the *a posteriori* pricing, through which the basic premium is adjusted to the individual history of the insured's claims (see e.g. [9]). Non-accident drivers are rewarded with

a premium reduction (bonus) and the drivers that caused one or more accidents are penalized with a higher premium (malus). The *a posteriori* pricing is designed to estimate the individual risk, so that each insured in the long term pays a premium corresponding to their frequency of claims. Such pricing is called the bonus-malus system.

The bonus-malus system in motor insurance is defined as a system for setting an individual premium, taking into account the number of claims reported by the insured in the past. The premium rate depends on the value of damages reported in the previous insurance periods, on the basis of which the insured person is classified into the relevant tariff class (see e.g. [8]).

Definition 5.1. ([5]) The insurer's system of determining individual net premiums can be called the bonus-malus system if:

- the insureds in a given tariff group (portfolio) are assigned to a finite number of tariff classes C_i for $i = 1,2, \dots, s$ in such a way that their annual premium depends only on the tariff class in which they are located,
- the tariff class in the current insurance period (usually one year) depends on the class in which the insured was classified in the previous insurance period (in the previous year) and the number of claims in the previous insurance period.

The bonus-malus system is based on the following assumptions:

- not all the insureds cause the same amount of damage on average per year,
- insurance periods are of equal length,
- the amount of damages caused by the insured does not depend on their size,
- the distribution of number of damages caused by the insured in one year does not change over time,
- the distribution of size of a single damage is constant over time and is the same for all insureds,
- the insureds remain in the same bonus-malus class throughout the entire single insurance period.
- The bonus-malus system is described by three elements:
- initial class, to which all new insureds are classified,
- basic premium rates vector $\bar{b} = (b_1, b_2, b_3, \dots, b_s)$ expressed as a percentage of the basic premium,
- transition rules describing moving from one class to another depending on the number of claims in the previous period.

Tariff classes in bonus-malus systems differ in the premium rate, expressed as a percentage of the basic premium. Classes with premium rate less than 100% of basic premium are classes in which there is a 'bonus' and classes with premium rate greater than 100% of basic premium are the 'malus' classes.

Insurance companies usually describe transition rules by using tables. A sample presentation of the bonus-malus system is shown in table 2.

Table 2. Bonus-malus system in comprehensive cover.

Number of observations	Number of classes
<400	15-19
401-600	20-24
601-800	25-27
801-1000	27-30
1001-1500	30-35
1501-2000	35-40

Source: own elaboration.

The bonus-malus system presented in table 2 contains $s = 5$ classes, for which the premium rate has the following form:

$$\bar{b} = [1.1 \quad 1 \quad 0.8 \quad 0.7 \quad 0.6].$$

With such a bonus-malus structure, its two functions are assumed: tariff and preventive. The first one makes it possible to match the premium rate to individual risk better. The second one causes reduction in the number and size of claims through higher or lower premium rates. In case of markets where various bonus-malus systems coexist, the selective function is also fulfilled. This means that a properly designed system attracts good drivers and discourages bad ones (see e.g. [1]).

Every bonus-malus system consists of classes with specified premium rates, the initial class and transition rules. Thus, to each bonus-malus class a premium rate is assigned, which is a percentage of net premium and is also called the bonus-malus coefficient. The design of bonus-malus systems is primarily the estimation of premium rates in each class of the system as well as determining the transition rules.

The 'optimal bonus-malus system' is a system that fulfils the expectations of both the insurance company and the insured. It is therefore a system that provides the insurer with a financial balance in the portfolio and at the same time is fair to the insured. It can be determined with the use of Bayesian methods (see e.g. [4,5]). In the construction of optimal bonus-malus systems, Bayesian individual risk parameters will be used to estimate net premiums. In this case, the premium rate is the quotient of the Bayesian premium (determined on the basis of individual history of the insured's claims) and the premium for the single policy determined for the whole portfolio, called the collective premium. The optimal bonus-malus system constructed this way is fair to the insured, since each insured pays a premium proportional to the number of their claims in the past. It is also financially balanced, since the average bonus-malus premium for a policy from the portfolio is equal to the average premium calculated without the use of the bonus-malus system.

7 Bayesian estimators of bonus-malus premium rates

In this part of the work the Bayesian estimator of risk structure parameter (called the Bayesian premium with quadratic loss function) will be calculated (see e.g. [2, 3, 12]). The estimator of bonus-malus premium rate will be derived, depending on the individual number of claims. This can be done by calculating the collective premium while taking into account the expected number of claims in the portfolio (in this case $EX = 1$ and the random variable K has the expected value EK).

In motor insurance, the individual net premium in insurance period $t + 1$ is determined on the basis of the equation:

$$\Pi(X, K) = (EX)(EK)b_{t+1}, \tag{16}$$

where

- $\Pi(X, K)$ – individual net premium in insurance period $t + 1$,
- EX – expected value of a single claim in the portfolio,
- EK – expected value of number of claims for a single policy in the portfolio,
- b_{t+1} – premium rate in insurance period $t + 1$.

We assume that random variables of number and value of claims are independent.

The premium rate is the quotient of the *a posteriori* premium, called the Bayesian premium, and the collective premium for the portfolio, called the *a priori* premium:

$$b_{t+1} = \frac{P^B}{P^K}100\% = \frac{\Pi(X, K)}{(EX)(EK)}100\%, \tag{17}$$

where

- P^B – Bayesian premium,
- P^K – collective premium.

In order to construct the bonus-malus system, the insurer must specify the percentages b_{t+1} and set the transition rules.

Let us assume the following notations:

- K_j – random variable for number of claims in j-th year,
- $(k_1, k_2, \dots, k_t)$ – vector of numbers of claims observed over the past t years,
- $\lambda_{t+1}(k_1, k_2, \dots, k_t)$ – unknown loss parameter in year $t + 1$ for the policy described by the observation vector $(k_1, k_2, \dots, k_t)$.

Suppose the distribution of claims in the portfolio is negative binomial defined by (3). The claim frequency parameter λ has the *a priori* gamma distribution (2) with parameters a and b. Assuming a quadratic loss function, the Bayesian estimator of the parameter λ is the conditional expected value of the *a posteriori* distribution and has the form:

$$\hat{\lambda}_B = \lambda_{t+1}(k_1, \dots, k_t) = E_\lambda[\lambda|k_1, \dots, k_t] = \int_0^\infty \lambda dF(\lambda|k_1, \dots, k_t) \tag{18}$$

where $E_\lambda[\lambda|k_1, \dots, k_t]$ is the conditional expected value of the *a posteriori* distribution of parameter λ and $F(\lambda|k_1, \dots, k_t)$ is the cumulative distribution of the random variable λ for observed values $(k_1, k_2, \dots, k_t)$ (see e.g. [5]). Using the Bayes' theorem we obtain:

$$\begin{aligned} dF(\lambda|k_1, \dots, k_t) = & \frac{P(k_1,\dots,k_t|\lambda)dF(\lambda)}{\int_0^\infty P(k_1,\dots,k_t|\lambda)dF(\lambda)} = \frac{\frac{\lambda^k e^{-t\lambda}}{\prod_{j=1}^t (k_j!)} \frac{b^a e^{-b\lambda}\lambda^{a-1}}{\Gamma(a)}}{\int_0^\infty \frac{\lambda^k e^{-t\lambda}}{\prod_{j=1}^t (k_j!)} \frac{b^a e^{-b\lambda}\lambda^{a-1}d\lambda}{\Gamma(a)}} d\lambda = \\ = & \frac{\frac{\lambda^k e^{-t\lambda}}{\prod_{j=1}^t (k_j!)} \frac{b^a e^{-b\lambda}\lambda^{a-1}d\lambda}{\Gamma(a)}}{\frac{b^a}{\prod_{j=1}^t (k_j!)\Gamma(a)} \int_0^\infty \lambda^{a+k-1} e^{-(b+t)\lambda}d\lambda} = \frac{\lambda^{a+k-1} e^{-(b+t)\lambda}d\lambda}{\int_0^\infty \lambda^{a+k-1} e^{-(b+t)\lambda}d\lambda} = \\ = & \frac{(b+t)^{a+k}\lambda^{a+k-1} e^{-(b+t)\lambda}d\lambda}{(b+t)^{a+k}\int_0^\infty \lambda^{a+k-1} e^{-(b+t)\lambda}d\lambda} = \frac{(b+t)^{a+k}\lambda^{a+k-1} e^{-(b+t)\lambda}d\lambda}{\int_0^\infty (b+t)^{a+k}\lambda^{a+k-1} e^{-(b+t)\lambda}d\lambda} = \\ = & \frac{\hat{b}^{\hat{a}}\lambda^{\hat{a}-1} e^{-\hat{b}\lambda}d\lambda}{\int_0^\infty \hat{b}^{\hat{a}}\lambda^{\hat{a}-1} e^{-\hat{b}\lambda}d\lambda} = \frac{\hat{b}^{\hat{a}}\lambda^{\hat{a}-1} e^{-\hat{b}\lambda}}{\Gamma(\hat{a})} d\lambda, \end{aligned}$$

where $P(k_1, \dots, k_t|\lambda) = P(k_1|\lambda) \dots P(k_t|\lambda) = \frac{\lambda^{k_1} e^{-\lambda}}{k_1!} \dots \frac{\lambda^{k_t} e^{-\lambda}}{k_t!} = \frac{\lambda^k e^{-t\lambda}}{\prod_{j=1}^t (k_j!)}$.

Thus, the *a posteriori* distribution of the parameter λ is a gamma distribution with parameters $\hat{a} = a + k$ and $\hat{b} = b + t$, where $k = \sum_{i=1}^t k_i$.

The Bayesian estimator of the parameter λ is the expected value of the gamma distribution with parameters $\hat{a}$ and $\hat{b}$ and has the form:

$$\hat{\lambda}_B = \lambda_{t+1}(k_1, \dots, k_t) = \frac{\hat{a}}{\hat{b}} = \frac{a+k}{b+t}. \tag{19}$$

The predictive distribution of the number of claims in year $t+1$ is the negative binomial distribution with the probability distribution function defined by the formula:

$$P(K_{t+1} = k|K_1 = k_1, \dots, K_t = k_t) = \frac{\Gamma(\hat{a}+t)}{t!\,\Gamma(\hat{a})} \left(\frac{\hat{b}}{1+\hat{b}}\right)^{\hat{a}} \left(\frac{1}{1+\hat{b}}\right)^t. \tag{20}$$

Standard premium rates in motor insurance are estimated on the basis of the number of claims reported by the insured over the period $1, \dots, t$. Then the equation (16) can be represented as follows:

$$\Pi(X, K) = (EX)(EK)b_{t+1}(k_1, \dots, k_t). \tag{21}$$

Assuming that the distribution of the number of claims is the negative binomial distribution, we obtain $EK = \frac{a}{b}$. Further assuming that $EX = 1$ and that in this case $\Pi(X, K) = \Pi(K)$, the equation (21) has the form:

$$\Pi(K) = \frac{a}{b} b_{t+1}(k_1, \dots, k_t). \tag{22}$$

Thus, the premium rate in year $t + 1$ for the insured who reported k claims over the past t years should be:

$$b_{t+1}(k_1, \dots, k_t) = \frac{b}{a} \Pi(K) 100\%. \tag{23}$$

Thus, assuming the negative binomial distribution of number of claims, using the quadratic loss function in the Bayesian estimation, applying the equivalence of premium principle in calculating the individual net premium and taking into account the formula (19), the premium rate in year $t + 1$ for the insured who reported k claims over the past t years can be calculated with the formula:

$$b_{t+1}(k_1, \dots, k_t) = \frac{b(a + k)}{a(b + t)} 100\%. \tag{24}$$

8 Construction of an optimal bonus-malus system

In February 2017 a study among car owners was conducted. Its aim was to gain information on the number of policies, from which k claims were reported, where $k = 0,1,2,$ ... (see Table 3). 500 people participated in the online survey. The respondents were asked to answer three questions regarding their claims (for damages caused by their own cars) reported to insurance companies in 2016 and causing an increase of their premium rate in the following year.

Table 3. Observed numbers (empirical)

k	0	1	2	3	4 and more
Number of policies with k claims	378	81	24	10	7

Source: own elaboration.

An Excel spreadsheet was used to determine the values of descriptive statistics and estimators. For $k \geq 4$ a simplification $k = 4$ was made in calculations, i.e. it was assumed that $\sum_{k \geq 4} N_k = N_4$, which had no significant impact on final results since the respondents rarely reported more than 4 claims. Let $n = \sum_{k=0}^{4} N_k = 500$ be the number of risks from observations. For out data we obtain:

$$\bar{X} = M_1 = \frac{1}{n}\sum_{k=0}^{4} kN_k = 0{,}3740,$$

$$M_2 = \frac{1}{n}\sum_{k=0}^{4} k^2 N_k = 0{,}7580,$$

$$M_3 = \frac{1}{n}\sum_{k=0}^{4} k^3 N_k = 1{,}9820,$$

$$C_3 = M_3 - 3M_2M_1 + 2{M_1}^3 = 1{,}2362,$$

$$S^2 = M_2 - M_1^2 = 0{,}6181.$$

Using the formulas (5), we can calculate the values of parameters a and b:

$$\bar{X} = \frac{a}{b},$$

$$S^2 = \frac{a}{b}\left(1 + \frac{1}{b}\right),$$

$$S^2 = \frac{a}{b}\left(1 + \frac{1}{b}\right) = \bar{X}\left(1 + \frac{1}{b}\right) = \bar{X} + \frac{\bar{X}}{b},$$

$$b = \frac{\bar{X}}{S^2 - \bar{X}},$$

$$a = \bar{X}b = \frac{\bar{X}^2}{S^2 - \bar{X}}.$$

Therefore:

$$b = \frac{\bar{X}}{S^2 - \bar{X}} = 1.5322,$$

$$a = \frac{\bar{X}^2}{S^2 - \bar{X}} = 0.5730.$$

Thus $T_k = (a + b) + ak = (0.5730 + 1.5322) + 0{,}5730k$, where $k = 0,1,2,\ldots$. The slope a is positive and $\bar{X} < S^2$, so the distribution of number of claims may be negative binomial.

The statistical test procedure starts with the null hypothesis:

H_0: *The random variable for number of claims has the negative binomial distribution*.

Assuming the truth of the null hypothesis, we calculate the expected numbers (theoretical) np_k. Using the characteristics of the theoretical distribution, we can

calculate the probabilities p_k that the random variable for number of claims with negative binomial distribution is of values from class k, where $k = 1,2,\dots,m-1$ and m is the number of classes. Of course numbers p_k must satisfy the condition $\sum_{k=0}^{m-1} p_k = 1$. Then, by multiplying p_k by n we obtain the expected numbers:

$$p_k = a+k-1_k \left(\frac{b}{1+b}\right)^a \left(\frac{1}{1+b}\right)^k,$$

$$p_0 = a+0-1_0 \left(\frac{b}{1+b}\right)^a \left(\frac{1}{1+b}\right)^0 = \left(\frac{b}{1+b}\right)^a,$$

$$\begin{aligned} p_{k+1} = {} & a+k_{k+1} \left(\frac{b}{1+b}\right)^a \left(\frac{1}{1+b}\right)^{k+1} = a+k_{k+1} \left(\frac{b}{1+b}\right)^a \left(\frac{1}{1+b}\right)^k \left(\frac{1}{1+b}\right) = \\ = {} & \left(\frac{1}{1+b}\right) a+k_{k+1} \left(\frac{b}{1+b}\right)^a \left(\frac{1}{1+b}\right)^k = \\ = {} & \left(\frac{1}{1+b}\right) \frac{(a+k)!}{(a-1)!(k+1)!} \left(\frac{b}{1+b}\right)^a \left(\frac{1}{1+b}\right)^k = \\ = {} & \left(\frac{1}{1+b}\right) \frac{(a+k)(a+k-1)!}{(k+1)k!(a-1)!} \left(\frac{b}{1+b}\right)^a \left(\frac{1}{1+b}\right)^k = \\ = {} & \left(\frac{1}{1+b}\right) \left(\frac{a+k}{k+1}\right) \frac{(a+k-1)!}{k!(a-1)!} \left(\frac{b}{1+b}\right)^a \left(\frac{1}{1+b}\right)^k = \\ = {} & \frac{(a+k)}{(1+b)(k+1)} a+k-1_k \left(\frac{b}{1+b}\right)^a \left(\frac{1}{1+b}\right)^k = \frac{(a+k)}{(1+b)(k+1)} p_k. \end{aligned}$$

Thus, the probability that k claims will be reported can be calculated:

$$p_0 = \left(\frac{b}{1+b}\right)^a = 0.7499,$$

$$p_1 = \frac{a}{(1+b)} p_0 = 0.1697,$$

$$p_2 = \frac{a+1}{2(1+b)} p_1 = 0.0527,$$

$$p_3 = \frac{a+2}{3(1+b)} p_2 = 0.0178,$$

$$p_4 = 1 - (p_0 + p_1 + p_2 + p_3) = 0.0099.$$

Table 4 contains the calculated expected numbers np_k.

Table 4. Expected numbers (theoretical).

k	0	1	2	3	4 and more
np_k	374.95	84.85	26.35	8.9	4.95

Source: own elaboration.

To evaluate the adequacy of the theoretical distribution to the empirical distribution, we will use the Pearson's χ^2 test, since the following assumptions are satisfied:

- $k = 5$ – number of classes is not less than 5,
- $n = 500 \geq 10k = 50$,
- empirical numbers are greater than 5.

Let us assume that the significance criterion will be the value of probability of the null hypothesis less than 0.05, i.e. $\alpha = 0.05$. The value of the χ^2 statistic is:

$$\chi^2 = \sum_{k=0}^{4} \frac{(n_k - np_k)^2}{np_k} = 1.1591.$$

We have $k = 5$ classes and 2 necessary parameters of the distribution ($l = 2$). Therefore, the number of degrees of freedom is $df = k - l - 1 = 5 - 2 - 1 = 2$. Thus $\chi^2_\alpha = 5.9915$ and $\chi^2 = 1.1591 < \chi^2_\alpha = 5.915$.

Therefore, there is no reason to reject the null hypothesis that the random variable for the number of claims has a negative binomial distribution.

We assume that the distribution of number of claims is the negative binomial distribution, $EX = 1$ and $\Pi(X, K) = \Pi(K)$. Using the formula (24), we can calculate premium rates in year $t + 1$ for the insured who reported k claims over the past t years. The calculated premium rates of the optimal bonus-malus system are presented in Table 5.

Table 5. Premium rates of the optimal bonus-malus system.

t	$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k \geq 4$
0	100				
1	61	166	272	377	483
2	43	119	195	270	346
3	34	93	152	211	270
4	28	76	124	173	221
5	23	64	105	146	187
6	20	56	91	127	162

Source: own elaboration.

9 Summary

The main aim of the article was to construct an optimal bonus-malus system. The chosen model of number of claims, estimating the premium rate in motor insurance as well as the loss function made it possible to design a system meet-

ing all the expectations of both the insurance company and the surveyed insureds.

To calculate the net premium rates, the Bayesian estimator of individual risk parameter was necessary. In this case, the premium rate is the quotient of the Bayesian premium by the individual premium for a single policy determined for the whole portfolio. An optimal bonus-malus system constructed this way is fair for the insured since each of them pays a premium proportional to the number of their claims in the past. Moreover, it is financially balanced since the average premium for a policy from the portfolio determined with the use of the bonus-malus system is equal to the average premium calculated without the use of the bonus-malus system.

The bonus-malus systems used in the insurance market are constantly being modified and the theoretical methods of constructing those systems are an important problem in the field of actuarial mathematics.

Bibliography

[1] Cieślik B., *System bonus-malus jako narzędzie konkurencji na rynku ubezpieczeń komunikacyjnych*, Poltext, Warszawa 2013.
[2] Domański Cz., Pruska K., *Nieklasyczne metody statystyczne*, Polskie Wydawnictwo Ekonomiczne, Warszawa 2000.
[3] Kowalczyk P., Poprawska E., Ronka-Chmielowiec W., *Metody aktuarialne*, PWN, Warszawa 2006.
[4] Lemaire J., *Automobile insurance. Actuarial Models*, Kluwer, Boston 1985.
[5] Lemaire J., *Bonus-Malus Systems in Automobile Insurance*, Kluwer, Boston 1995.
[6] Kaas R., Goovaerts M., Dhaene J., Denuit M., *Modern Actuarial Risk Theory*, Kluwer Academic Publishers, New York, Boston, Dordrecht, London, Moscow 2001.
[7] Monkiewicz J., *Podstawy ubezpieczeń – Vol. 1. Mechanizmy i funkcje*, Poltext, Warszawa 2000.
[8] Łazuka E., Stępkowska K., *Analiza modyfikacji systemów bonus-malus w ubezpieczeniach komunikacyjnych AC na przykładzie wybranego zakładu ubezpieczeń*, „Wiadomości Ubezpieczeniowe", Polska Izba Ubezpieczeń, No. 1/2014, 91-114.
[9] Ostasiewicz W., *Składki i ryzyko ubezpieczeniowe: modelowanie stochastyczne*, Wydawnictwo Akademii Ekonomicznej, Wrocław 2004.
[10] Otto W., *Ubezpieczenia majątkowe – część I – Teoria ryzyka*, WNT, Warszawa 2004.
[11] Stanisz A., *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny – Vol. 1. Statystyki podstawowe*, StatSoft, Kraków 2006.

[12] Szymańska A., *Statystyczna analiza systemów bonus-malus w ubezpieczeniach komunikacyjnych*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź 2014.

Przemysław Kowalik[1]

Selection of independent variables in econometric models as a binary programming problem and its application to spreadsheet-based calculations

Keywords: econometric model, independent variables, dependent variables, correlation, binary programming, spreadsheets.

Abstract

Selecting independent variables in an econometric model can be performed by using one of many criteria or methods. No matter what method of selection is used, some combination of variables out of k potential independent variables is selected. If the number "1" is assigned to each of the selected variables and the number "0" each of the non-selected ones and the criterion of optimal selection is a single function (maximized or minimized) depending on input data, then the selection process can be considered as a binary programming problem. More precisely, zero-one combinations of independent variables can be considered as vectors of binary variables ("variables" in the sense of the optimization theory, not econometrics). There exists at least one method of selection of independent variables (by M. Rocki) explicitly based on binary linear programming. However, at least two other methods (developed by Z. Hellwig and Z. Pawłowski) can be reformulated as binary non-linear programming problems. Those reformulations are not innovative in any theoretical aspect of the methods nevertheless they may be very useful for practitioners in various fields when performing calculations using spreadsheet programs.

1 Introduction

Selection of independent (or explanatory) variables in an econometric model is performed in order to reject variables irrelevant to that model. Many methods of selection of variables have been developed. Whatever is the mathematical idea behind any specific method, some very general, common description of all the methods exists. Let us consider k potential independent variables. Assuming that each variable can be rejected or not, there are $L = 2^k$ combinations of rejected and not rejected variables. The decision applied to each variable can be described numerically by using 0 for rejection and 1 for acceptance. The combination of k zeroes obviously stands for rejection of all the variables, or, in other

[1] Department of Quantitative Methods in Management, Faculty of Management, Lublin University of Technology, e-mail: p.kowalik@pollub.pl.

words, acknowledging the fact all the input data are not valid to create an econometric model. Such a zero-one (binary) notation leads directly to a formulation of the variable selection problem as "find an optimal k-element combination of zeroes and ones from a set of $L = 2^k$ combinations". Finding an optimal k-element combination of zeroes and ones can be slightly reformulated. Let us assign to each potential independent variable a binary (zero-one valued) decision variable. The earlier usage of the word "variable" obviously refers to its meaning in econometrics whereas the latter to its mean in optimization theory (mathematical programming). Now, the variable selection problem can be expressed as "find an optimal k-element combination of binary decision variables c_j, $j = 1,2,\ldots,k$". If the criterion of optimality of selection of independent variables can be expressed as a single formula (either maximized or minimized), depending on input data and binary decision variables i.e. a function of binary decision variables c_j, then binary programming formulation of the problem may be applied.

Such an explicit binary programming approach to selection of independent variables was proposed by M. Rocki in 1980 ([12], [13] and [14]). The method uses binary linear programming and is based on an older method of selection (introduced in years 1968-1969), known as the method of capacity of information bearers, also called by the name of its author Z. Hellwig, the Hellwig method ([2,3]).

It turned out that the Hellwig method itself can be easily reformulated into a binary linear programming problem. Apparently such a reformulation does not make much sense since because the method is, by the definition, nothing more than checking all the possible combinations in order to find the one for which so-called integral capacity of information bearer is maximal. An important practical reason for this reformulation exists however, namely a convenient implementation of the Hellwig method in spreadsheets. Spreadsheet software, like Microsoft Office Excel, OpenOffice/LibreOffice Calc or WPS Spreadsheets is commonly used by researchers and, probably much more often, by business/public sector practitioners for many purposes including constructing econometric models. However, this software, whereas usually includes many built-in "statistical" features, is not user-friendly in some widely used econometrical or statistical applications. The above statement applies also to the Hellwig method. Some spreadsheet implementations of the Hellwig method based on explicit enumeration of all the possible "0-1" combinations were developed ([1,4,7,8]). Whereas generating all the possible "0-1" combinations of potential independent variables in a spreadsheet file is relatively easy, the exponential growth of the number of combinations results in poor performance of the software due to large sizes of files. This is why a different approach to the implementation of the Hellwig method in spreadsheets, based on binary non-linear programming was

successfully developed ([5,6]) and applied to some real-world calculations ([9,16]).

Another method of selection of independent variables which can be reformulated as a binary non-linear programming problem is the one by Z. Pawłowski (presented for the first time in 1981 in [11]) which consists in maximization of the coefficient of multiple correlation for all the combinations of a fixed number of independent variables.

This paper in sections 2 and 3 contains a review of existing achievements of binary programming approach to selection of independent variables. In section 4, a proposal of binary programming reformulation of the Pawłowski method is presented.

The paper does not estimate or compare considered methods of selection of independent variables regarding their quality from the "statistical" point of view because its main goal is a practical aspect of necessary calculations, not their statistical "background". Calculating correlation matrices in spreadsheet environment, which it is not directly supported by built-in spreadsheet functions is not considered here (see e.g. [5,6] for details).

2 The Hellwig method as a binary non-linear programming problem

The method of capacity of information bearers (also called the method of optimal choice of predictors or, after its author, the Hellwig method) is of one methods of selecting independent variables for an econometric model. As many others, the method consists in selecting such variables which are strongly correlated with the dependent variables and, simultaneously, weakly correlated with other independent variables ([2,3]). The selection of variables requires "testing" all of $L = 2^k - 1$ combinations of k potential independent variables ("zero" combination i.e. rejecting all the variables is not considered). The following notation (based on [10]) will be used:

- l – number of a combination ($l = 1,2,\ldots,L$);
- k_l – number of variables in the l^{th} combination;
- j – number of a variable in the l^{th} combination ($j = 1,2,\ldots,k_l$);
- r_j – correlation of the j^{th} independent variable with the dependent variable;
- r_{ij} – correlation of the j^{th} independent variable with other independent variables included in the l^{th} combination $i = 1,2,\ldots,k_l,\ i \neq j$;

The *individual capacity of information bearer* (later referred to as *individual capacity of information*) h_{lj} for the j^{th} independent variable ($j = 1,2,\ldots,k_l$) in the l^{th} combination $l = 1,2,\ldots,L$ is defined as

$$h_{lj} = \frac{r_j^2}{1 + \sum_{i=1, i \neq j}^{k_l} |r_{ij}|}.$$

The *integral capacity of information bearer* (later referred to as *integral capacity of information*) for the l^{th} combination is the sum of the abovementioned individual capacities of information bearers for the l^{th} combination:

$$h_l = \sum_{j=1}^{k_l} h_{lj}$$

The combination of independent variables for which the maximum of integral capacity of information is attained is chosen to be included in the econometric model.

Both individual and integral capacities of information are normalized i.e. they are included in the [0,1] interval. They increase as the independent variables are strongly correlated with the dependent variable and the independent variables are weakly correlated one to another.

Individual capacities of information for the l^{th} combination can also be expressed by formulas in which correlations are not indexed by combination-dependent indices but by "general" *indices i and j* varying from 1 to *k*

$$h_j = \frac{c_j r_j^2}{\sum_{i=1}^{k} c_i |r_{ij}|}, \quad j = 1,2,\ldots,k.$$

The number „1" from the definition was replaced with $|r_{ii}|$ (obviously equal to 1 for any *i*). The main difference to compare with the original definition by Hellwig is that selection of correlations which "occur" in an individual capacity of information is done by multiplication by 0-1 coefficients of the combination, not by the combination-dependent index of summation (see [6] for details) This is why the above formula is "universal" i.e. it stands for all the individual capacities of information for the j^{th} independent variable.

An immediate consequence of the above concept of indexing is considering c_j - coefficients of 0-1 combinations as k binary variables (the word „variable" is used here in the same meaning as in the deterministic optimization theory).

The problem of finding the maximum of the integral capacities of information bearer can now be expressed as the following binary non-linear programming problem of k variables $c_j (j = 1,2,\ldots,k)$.

$$\sum_{j=1}^{k} \frac{c_j r_j^2}{\sum_{i=1}^{k} c_j |r_{ij}|} \rightarrow \max$$

subject to

$\sum_{i=1}^{k} c_j \geq 1$ – rejecting all the potential independent variables is not allowed

c_j – binary (an independent variable X_j not selected (0)/ selected (1)).

The above formulation of the Hellwig method turned out to be easy to implement in spreadsheet software (Microsoft Excel, LibreOffice/OpenOffice Calc). The implementation uses built-in optimization software (called Solver, which is one common name for different optimization software) as a computational engine that performs the search for the best combination of independent variables. Application of optimization features of spreadsheet software to calculations necessary for the Hellwig methods seems to be useful in university education and data processing in scientific research or business practice.

3 The Rocki method as a binary linear programming problem

In 1980 M. Rocki introduced a method of selection of independent variables algorithm based on binary linear programming [12,13,14]. The statistical idea of his method was similar to that of the Hellwig method. His main aim was to reduce the number of necessary calculations to compare with the Hellwig method. Analogically to the Hellwig method, the Rocki method prefers selecting such variables which are strongly correlated with the dependent variable and, simultaneously, weakly correlated with other independent variables. The method is also based on formulas similar to that used in the Hellwig method. Using the notations defined in the previous section the Rocki method can be expressed as follows:

$$\sum_{j=1}^{k} |r_j| c_j \rightarrow \max$$

subject to

$$\sum_{j=1}^{k} |r_{ij}| c_j \leq r^* + k(1 - c_i), \quad i = 1,2,\dots,k$$

$$c_j - \text{ binary}, \quad j = 1,2,\dots,k$$

where r^* is some threshold value.

Some extension of the Rocki model [13,15], providing elimination of so-called catalytic variables can be done by adding the following linear constraints

$$|r_{ij}|c_j + 1 - c_i \geq 0 \quad i = 1,2,\dots,k-1, \quad j = 1,2,\dots,k$$

$$|r_{ij}|c_j \leq \frac{r_i}{r_j}c_i + 1 - c_i \quad i = 1,2,\dots,k-1, \quad j = 1,2,\dots,k$$

Since it is a standard binary linear programming problem, a spreadsheet implementation can be done easily.

4 The Pawłowski method as a binary non-linear programming problem and its spreadsheet implementation

In 1981 Z. Pawłowski in [11] proposed a method of selection of independent variables which is based on maximization of the coefficient of multiple correlation for all the combinations of a fixed number of independent According to the original formulation also the fraction of variance unexplained (FVU) must not be larger than some threshold level. For further considerations some additional notations must be introduced (based on [10]).

Let

$$\mathbf{R}_0 = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_k \end{bmatrix} \quad \mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1k} \\ r_{21} & r_{22} & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k1} & \cdots & r_{kk} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{bmatrix}$$

be a vector of correlations of independent variables with the dependent variable and a matrix of correlations of independent variables, respectively.

Let us define

$$\mathbf{W} = \begin{bmatrix} 1 & \mathbf{R}_0^T \\ \mathbf{R}_0 & \mathbf{R} \end{bmatrix} = \begin{bmatrix} 1 & r_1 & r_2 & \dots & r_k \\ r_1 & 1 & r_{12} & \cdots & r_{1k} \\ r_2 & r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_k & r_{k1} & r_{k2} & \cdots & 1 \end{bmatrix}$$

The coefficient of multiple correlation is then $R = \sqrt{1 - \frac{det\,\mathbf{W}}{det\,\mathbf{R}}}$ and FVU is $\varphi^2 = 1 - R^2 = \frac{det\,\mathbf{W}}{det\,\mathbf{R}}$.

Let coefficients of combinations of independent variables be the following vector

$$\mathbf{c} = [c_1 \quad c_2 \quad \dots \quad c_k].$$

The main problem of a spreadsheet implementation of the problem under consideration is similar to that of the implementation of the Hellwig method.

Namely, it is necessary to express the optimization criterion as an objective function i.e. a formula directly dependent on 0-1 coefficients of combinations.

In order to express values of the abovementioned determinants as depending on combinations of independent variables it is necessary to modify the matrices $\mathbf{R}$ and $\mathbf{W}$ to make them depend on 0-1 coefficients c_j of the combination. The modifications are the following

$$\mathbf{R}(\mathbf{c}) = \begin{bmatrix} 1 & c_1c_2r_{12} & \cdots & c_1c_kr_{1k} \\ c_2c_1r_{21} & 1 & \cdots & c_2c_kr_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ c_kc_1r_{k1} & c_kc_2r_{k2} & \cdots & 1 \end{bmatrix}$$

$$\mathbf{W}(\mathbf{c}) = \begin{bmatrix} 1 & c_1r_1 & c_2r_2 & \dots & c_kr_k \\ c_1r_1 & 1 & c_1c_2r_{12} & \cdots & c_1c_kr_{1k} \\ c_2r_2 & c_2c_1r_{21} & 1 & \cdots & c_1c_kr_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_kr_k & c_kc_1r_{k1} & c_kc_2r_{k2} & \cdots & 1 \end{bmatrix}.$$

Let r^* be a threshold value for the fraction of variance unexplained and p the number of independent variables to be selected. Considering the coefficients c_j of combinations to be binary variables in an optimization problem, the Pawłowski method can be expressed as

$$\mathbf{R}(\mathbf{c}) = \sqrt{1 - \frac{det\ \mathbf{W}(\mathbf{c})}{det\ \boldsymbol{R}(\mathbf{c})}} \rightarrow \max$$

subject to

$$\varphi(c)^2 = 1 - \mathbf{R}(c)^2 = \frac{det\ \mathbf{W}(\mathbf{c})}{det\ \mathbf{R}(\mathbf{c})} \leq r^*$$

$$c_1 + c_2 + \cdots + c_k = p, \qquad c_1, c_2, \dots, c_k \quad \text{binary.}$$

The objective function can be replaced with an equivalent but simpler one, so finally the optimization problem for the Pawłowski method is

$$\frac{det\ \mathbf{W}(\mathbf{c})}{det\ \boldsymbol{R}(\mathbf{c})} \rightarrow \min$$

subject to

$$\frac{det\ \mathbf{W}(\mathbf{c})}{det\ \mathbf{R}(\mathbf{c})} \leq r^*$$

$$c_1 + c_2 + \cdots + c_k = p, \qquad c_1, c_2, \dots, c_k \quad \text{binary.}$$

A possible spreadsheet implementation must be based directly on "encoding" the $\mathbf{W}(\mathbf{c})$ matrix (and, simultaneously, $\mathbf{R}(\mathbf{c})$ which is "embedded" in the latter one) as spreadsheet formulas. The idea of an implementation will be explained by a small-size example (based on [10]) and illustrated by screenshots.

Example. Find the best combination (in the sense of the Pawłowski method) of two independent variables out of four potential independent variables (the input data are given below)

$$\mathbf{R}_0 = \begin{bmatrix} 0.7 \\ 0.9 \\ 0.1 \\ 0.5 \end{bmatrix} \qquad \mathbf{R} = \begin{bmatrix} 1 & 0.8 & 0.2 & 0.4 \\ 0.8 & 1 & 0.1 & 0.6 \\ 0.2 & 0.1 & 1 & 0.3 \\ 0.4 & 0.6 & 0.3 & 1 \end{bmatrix} \qquad r^* = 0.2 \quad p = 2$$

A possible layout of the above data in the spreadsheet (together with necessary formulas) is shown on Figure 1.

	A	B	C	D	E	F	G	H	I
1			**W and R matrices**						
2		1	0.7	0.9	0.1	0.5			
3		0.7	1	0.8	0.2	0.4		det W	0.03448
4		0.9	0.8	1	0.1	0.6		det R	0.1872
5		0.1	0.2	0.1	1	0.3			
6		0.5	0.4	0.6	0.3	1		R	0.9032
7	"dummy" var.		combination coefficients - variables				sum	(det W)/(det R)=1-R^2	0.184188
8	***1***		**1**	**1**	**1**	**1**	4		
9	aux. numbers		**W(c) and R(c) matrices**					threshold	0.2
10	**1**	1	0.7	0.9	0.1	0.5		number of ind. var.	2
11	**2**	0.7	1	0.8	0.2	0.4			
12	**3**	0.9	0.8	1	0.1	0.6			
13	**4**	0.1	0.2	0.1	1	0.3			
14	**5**	0.5	0.4	0.6	0.3	1			

Figure 1. A screenshot of the example implemented in a spreadsheet (compatibility provided for Microsoft Excel 2007+ and LibreOffice Calc 5.0+).

Source: own elaboration.

Input data are located in the following cells

- B2 – number 1 (the top-left element of the matrix **W**);
- C2:F2– correlations of the independent variables with the dependent variable;
- C3:F6– matrix correlations of all independent variables;
- B8 – number 1 (the "dummy" variable with the fixed value of 1, used to simplify some formulas below);
- C8:F8 – numbers 1 (the initial values of the coefficients of combination – i.e. the variables in the optimization problem);
- I8 – the threshold value r^* for the fraction of variance unexplained ;
- I9 – *p,* the number of independent variables to be selected.;

- A10:A14 – auxiliary numbers 1,2,3,4,5 used to simplify the formula for the $\mathbf{W}(\mathbf{c})$ matrix.

The formulas are placed as follows

- B3:C6 {=TRANSPOSE(C2:F2)} – a transpose of the array of correlations of all independent variables with the dependent variable, necessary to construct the $\mathbf{W}$ and the $\mathbf{W}(\mathbf{c})$ matrices (an array formula);
- B10:F14 {=IF(A10:A14=TRANSPOSE(A10:A14),1,B2:F6*B8:F8* TRANSPOSE(B8:F8))| – $\mathbf{W}(\mathbf{c})$ in B10:F14 which also includes "embedded" $\mathbf{R}(\mathbf{c})$ in C11:F14 (an array formula);
- G8 =SUM(C8:F8) – sum of the variables
- I3 =MDETERM(B10:F14) – det $\mathbf{W}(\mathbf{c})$
- I4 =MDETERM(C11:F14) – det $\mathbf{R}(\mathbf{c})$
- I6 =(1-I3/I4)^0.5 – $R(\mathbf{c})$
- I7 =I3/I4 – $\varphi(\mathbf{c})^2$

The array formula in B10:F14 is the main "trick" of the implementation. The condition (referring to auxiliary numbers 1, 2, 3, 4, 5 in both the "explicit" location A10:A14 and the "virtual" transposed location TRANSPOSE(A10:A14)) provides that the IF function returns 1's as entries of the main diagonal i.e. B10, C11, D12, E13, F14. Elsewhere, non-diagonal entries of the "original" $\mathbf{W}$ matrix (B2:F6) are multiplied by products of pairs of variables $c_i c_j$ (B8:F8* TRANSPOSE(B8:F8)) what results in returning zeroes instead of original correlations whenever at least one of c_i and c_j is zero.

Tests performed on numerical data in various versions of Microsoft Excel showed that the previously formulated binary non-linear programming problem requires a slight change in order to obtain a correct solution. Not only initial values of the variable cells must be all equal to one (as visible on Figure 1), but also the constraint

$$c_1 + c_2 + \cdots + c_k = p$$

must be changed to

$$c_1 + c_2 + \cdots + c_k \leq p.$$

Finally, the Excel Solver settings should be as shown on Figure 2. The screenshot is made in Excel 2007 and it is valid for all the versions up to 2007. All the options of Solver are left at their default values. The settings for the newer Solver interface (Excel 2010 and later) are similar but they were skipped because of the size of the screenshot. Also LibreOffice Calc can be used for selection of independent variables with the Pawłowski method.

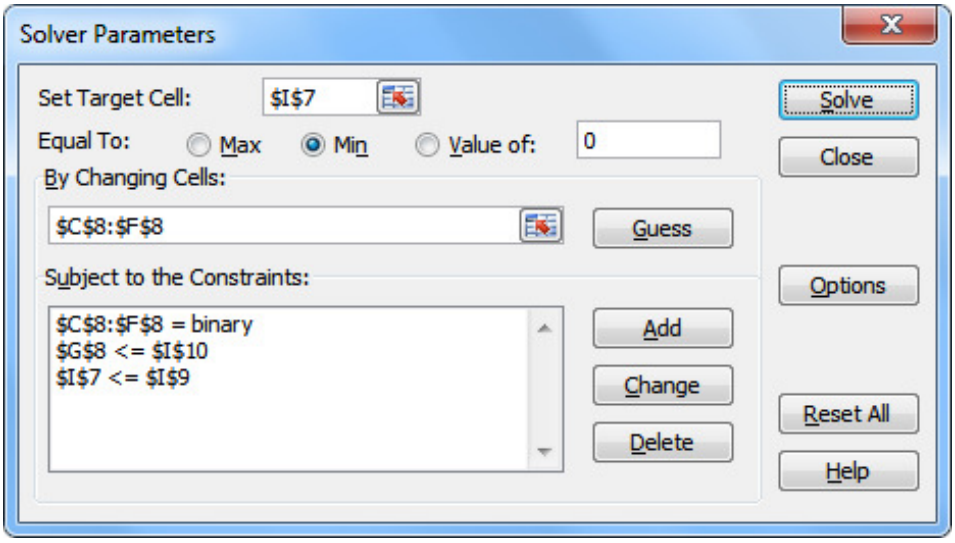


Figure 2. A screenshot of the Solver settings for the example implemented in a spreadsheet (Microsoft Excel 2007 and earlier, easily adaptable to a newer interface in Microsoft Excel 2010 and later).

Source: own elaboration, numerical data for matrices **W** and **R** are taken from Example 5 in [10].

The result of the calculations is shown on Figure 3.

	A	B	C	D	E	F	G	H	I
1			**W** and **R** matrices						
2		1	0.7	0.9	0.1	0.5			
3		0.7	1	0.8	0.2	0.4		det W	0.12
4		0.9	0.8	1	0.1	0.6		det R	0.64
5		0.1	0.2	0.1	1	0.3			
6		0.5	0.4	0.6	0.3	1		R	0.9014
7		"dummy" var.	combination coefficients - variables				sum	(det W)/(det R)=1-R^2	0.1875
8		***1***	**0**	**1**	**0**	**1**	2		
9	aux. numbers		**W(c)** and **R(c)** matrices					threshold	0.2
10	**1**	1	0	0.9	0	0.5		number of ind. var.	2
11	**2**	0	1	0	0	0			
12	**3**	0.9	0	1	0	0.6			
13	**4**	0	0	0	1	0			
14	**5**	0.5	0	0.6	0	1			

Figure 3. The result of selection of independent variables with the Pawłowski method performed by using Microsoft Excel or LibreOffice Calc with their Solver add-ins.

Source: own elaboration, numerical data for matrices **W** and **R** are taken from Example 5 in [10].

The result is identical with that obtained in [10] i.e. variables 2 and 4 are selected. However, the example in [10] does not include a threshold value r^* so the value of $r^* = 0.2$ in the example presented in this paper was added just to provide compatibility with original Pawłowski's concept.

If no integer/binary non-linear optimization feature is available in a spreadsheet programme (or it is inefficient in dealing with an optimization problem resulting from the Pawłowski method, like e.g. WPS Spreadsheets 10.2.0.5934, a part of WPS Office 2016), then there is still an option of using direct enumerating of all the 0-1 combinations. This option can be implemented analogically to similar implementations of the Hellwig method ([4,7,8]). Its main disadvantage is the fact that it basically limited to 10-15 potential independent variables due to spreadsheet efficiency issues. Details of the direct enumerating of all the 0-1 combinations are beyond the scope of this paper.

5 Summary

One of practical aspects of applying econometric tools to real-world problems is dealing with large amounts of numerical data what makes it necessary to engage computers and appropriate software into calculations. Obviously the cost of software as well as the quality of its usage (including training) cannot be neglected. This is why the choice of software is an important issue, and that choice may be either introducing new software or exploiting some features of the software already used. Spreadsheet implementations of methods of selection of independent variables in econometric models were developed because of popularity of that kind of software in business, public administration, education and scientific research. Existing successful spreadsheet implementations of some methods of selection of independent variables based on binary programming show that it can be worth considering creating such implementations also for other methods.

Bibliography

[1] Anholcer M., Gaspars-Wieloch H., Owczarkowski A., *Ekonometria z Excelem: przykłady i zadania*, Wydawnictwo Uniwersytetu Ekonomicznego, Poznań 2010.

[2] Hellwig Z., *On the Optimal Choice of Predictors*, [in:] Study VI, Toward a System of Quantitative Indicators of Components of Human Resources Development, UNESCO, Paris 1968.

[3] Hellwig Z., *Problem optymalnego wyboru predykant*, "Przegląd Statystyczny", R.XVI, 3–4, 1969.

[4] Kowalik P., *On an implementation of the method of capacity of information bearers (the Hellwig method) in spreadsheets*, Probability in Action, eds. Tadeusz Banek, Edward Kozłowski, Politechnika Lubelska, Lublin 2014, 31–40.

[5] Kowalik P., *Zastosowanie nieliniowego programowania binarnego do wyboru zmiennych objaśniających metodą wskaźników pojemności informacyjnej Hellwiga*, [in:] Natura i uwarunkowania ryzyka, Monografie Politechniki Łódzkiej, Łódź 2014, 284–295.
[6] Kowalik P., *Metoda wskaźników pojemności informacyjnej Hellwiga jako zadanie nieliniowego programowania binarnego,* [in:] Rola informatyki w naukach ekonomicznych i społecznych. Innowacje i implikacje interdyscyplinarne", Tom 2/2013 Wydawnictwo Wyższej Szkoły Handlowej, Kielce 2013, 215–224.
[7] Kowalik P., *Wykorzystanie arkuszy kalkulacyjnych do wyboru zmiennych objaśniających przy pomocy metody wskaźników pojemności informacyjnej (metody Hellwiga)*, [in:] Rola informatyki w naukach ekonomicznych i społecznych. Innowacje i implikacje interdyscyplinarne, Tom 2/2012, Wydawnictwo Wyższej Szkoły Handlowej, Kielce 2012, 168–178.
[8] Kowalik P, *Implementacja metody wskaźników pojemności informacyjnej (metody Hellwiga) w arkuszach kalkulacyjnych*, [in:] Rola informatyki w naukach ekonomicznych i społecznych. Innowacje i implikacje interdyscyplinarne, Tom 2, Wydawnictwo Wyższej Szkoły Handlowej, Kielce 2011, 186–194.
[9] Ludwiczak B, *Efektywność wydatków powiatów wschodniego regionu Polski w latach 2008–2012*, "Nierówności Społeczne a Wzrost Gospodarczy", nr 40 (4/2014), 125–136.
[10] Nowak E., *Zarys metod ekonometrii. Zbiór zadań*, Wydawnictwo Naukowe PWN, Warszawa 1994.
[11] Pawłowski Z., *Elementy ekonometrii*, Państwowe Wydawnictwo Naukowe, Warszawa, 1981.
[12] Rocki, M., *Metody doboru zmiennych w modelach ekonometrycznych jako zadania programowania matematycznego*, Ph.D. thesis, Institute of Econometrics, Main School of Planning and Statistics, Warszawa 1980.
[13] Rocki, M., *O pewnym zadaniu liniowego programowania calkowitoliczbowego jako metodzie doboru zmiennych w modelach ekonometrycznych*, "Przegląd Statystyczny", nr 1-2/1982.
[14] Rocki, M., *Własność koincydencji i zmienne katalityczne w doborze zmiennych objaśniających za pomocą zadania programowania zerojedynkowego*, "Przegląd Statystyczny", Tom 30, 1/2, 1983, 35–39.
[15] Rocki, M., *Ekonometria praktyczna*, Szkoła Główna Handlowa, Warszawa 2000.
[16] Sarnowski D., *Wartość diagnostyczna kryteriów selekcji 17-19-letnich koszykarzy*, Ph.D. thesis, Gdańsk University of Physical Education and Sport, Gdańsk 2016.

Andrii Kaminskyi[1], Ruslan Motoryn[2], Konstantyn Pysanets[3]

The effectiveness of the use of statistical data of credit histories bureaus in risk management systems

Keywords: credit bureau, credit risk management, reject analysis.

Abstract

The article is devoted to analysis of application the statistical potential of credit bureaus. Credit bureaus became an inherent part of modern credit market, especially at the segment of consumer lending. At the same time, possibilities of bureau`s statistics are not used effectively. The article suggests approaches to increasing effectiveness of risk management in consumer lending segment based on the study of data amassed by credit bureaus. In particular, the analysis of rejected applications, the analysis of statistical distribution of scoring inflows and bureau benchmarking are considered. The reject analysis gives the possibilities to improve rules for discrimination good and bad borrowers. The statistical analysis of market inflow is a good indicator for understanding risk environment. Bureau benchmarking which based on market statistics provides good comparison for understanding effectiveness separate creditors.

1 Introduction

The system of credit bureaus is one of the most important components of modern credit relations and an important infrastructural element of the credit market. Credit bureaus engage in collection and exchange of credit reports, thus reducing the information asymmetry between lenders and borrowers. Effectiveness of this system contributes to greater reliability of lending, strengthens and improves stability of the financial sector.

Effective performance of credit bureaus reduces credit risks and makes loans more affordable to responsible borrowers. In the corporate segment, credit bureaus help increase competitiveness of organizations.

Credit bureaus have existed and accumulated experience in the credit markets of some developed countries for more than 150 years. The first bureaus appeared

[1] Department of Statistics and Econometrics, Taras Shevchenko National University of Kyiv.
[2] Department of Quantitative Methods in Management, Faculty of Management. Lublin University of Technology.
[3] Department of Statistics and Econometrics, Taras Shevchenko National University of Kyiv.

in the second half of the 19th century in Austria (1860), Sweden (1890), Finland (1890) and, later on, in other countries of Western Europe. Today, credit bureaus operate in the credit markets at almost all countries. Credit information exchange systems have integrated into the institutional architecture of a developed market economy, both in Europe and in the world. In Europe, in the vast majority of countries (17 against 6), provision of credit information is optional. Genesis of credit bureaus and cross-country evidence are presented in [7,10,11,17].

The organization of the bureau system, its structure and functions differ from one market to another. In a number of countries, the bureau system is organized as a competitive environment of private institutions. A typical example is the system of credit bureaus in the United States of America, where there are three basic bureaus (TransUnion, Experian and Equifax) and a number of smaller bureaus serving individual regions or industries. Bureaus there share both positive and negative credit information. In Germany, the bureau of credit histories is a union of eight regional, legally and economically independent entities (SCHUFA). In France, this system is represented by a state register. In Denmark, Belgium, Spain, Australia, Mexico, Brazil and several other countries provide only negative credit information.

In 1997, a Credit Information Bureau (BIK) was established in Poland. The main task of BIK was to provide information on clients' creditworthiness. At present, more than 680 institutions participate in BIK information exchange systems. Information resources BIK cover over 137 million credit accounts owned by 23 million Poles ([21]). According to experts from the World Bank in 2013, Poland was one of the leaders in the region and globally in terms of the quality of the credit information.

Credit bureaus as inherent element of the consumer loans market were established in Ukraine at 2005. There are 7 credit bureaus in Ukraine in current period, though three largest bureaus cover 99% of individual borrowers ([22,23,24]). The market of personal loans uses bureaus very intensively. The market of corporative credits interacts with bureaus non-actively.

There are about 30 credit bureaus in Russia (the main five of them hold information about 95% of borrowers). Moreover, the Central Bank of Russia created a central catalogue of credit histories containing information on bureaus holding particular borrower's credit history.

It is necessary to note that some countries use approach based on establish state credit register. Bosnia and Herzegovina is an example.

Credit bureaus are focused on the collection and provision of credit information to participants in the credit market. At the same time, they accumulate enormous statistical information, and its use presents significant potential in improving the efficiency of risk management systems. This article suggests approaches for using the data of credit bureaus in order to raise the efficiency of credit risk management systems in consumer lending segment.

2 Functions of credit bureaus

Economic literature comprises a number theoretical and empirical studies devoted to the analysis of the functions of credit bureaus, credit reporting systems, and their role. Important contributions to the study of credit reporting were made by M. Miller, D. Rudmen, T. Jappelli, M. Pagano, S. Djankov, V.Simovic, M.Rothemund, M.Gerhardt and many others.

The system of credit information exchange through credit bureaus consists of economic, technical, and legal components. This article focuses on the economic one. Existing research (for example, [5]) has defined a number of credit bureaus' economic functions, including the following:

Reduction of information asymmetry risk. The risk of information asymmetry is inherent to credit relations, as lenders and borrowers often have different information at their disposal. When credit bureaus system is not existent, lenders can not properly assess the borrowers' creditworthiness. Borrowers may hide information about some negative aspects of their past and, vice versa, assign greater importance to the positive aspects. Taken together, this may lead to a potentially misguided decision made by lender and increases the credit risk. Moreover, borrowers may be over-credited and hide this information from the lender, applying for a new loan. Reduction in the information asymmetry with the help of credit bureaus decreases lenders' risk, helps reject adverse borrowers and ensures more favourable terms for responsible and trustworthy borrowers.

1) Reduction in costs on information collection and data analysis. In the absence of credit bureaus, creditors need to spend a lot of time searching for information about borrowers in different sources. As an alternative to bureaus, they may use the so-called 'blacklists', but their legitimacy is under big question. When credit bureaus are present on the market, they store information on all credit transactions, and when a credit institution makes a query to the bureau, the required information is promptly provided. Modern information technologies enable collection, systematization and provision of credit information in a highly efficient manner. The duration of the query to the bureau and receipt of the answer takes only a few seconds.
2) Reduction of moral hazard to borrowers. Failure to repay a loan may result from the economic to do so or from the borrower's reluctance to make the necessary payment. In the latter case, credit bureaus stimulate borrowers to develop more responsible attitude to fulfilling their obligations. Information about credit transactions is stored for a long period (e.g., in Ukraine it is stored for ten years), during which a borrower with a negative credit history will not be able to get a loan from banks working with the bureau.

The abovementioned functions have a significant impact on the credit market. In particular, reductions in information asymmetry and moral hazard allow lenders reduce interest rates on loans. Generally, lenders include risk premium

into the interest rate, which naturally affects the lending economy: with an increase of risk premium, the demand for loans falls.

There may be differences in the implementation of credit bureau functions in specific markets and they are usually reflected in the principles of bureau system organization and in market structure. Main differences in the organization of bureau systems are as follows:

1) obligation to provide information to the bureau by all lenders;
2) participation of the state in the credit bureau system;
3) information provided to bureau users (only negative, or both negative and positive);
4) need of agreement with the borrower on processing and transferring of his/her personal information to the bureau.

3 Implementation of credit bureaus into the risk management system

The information significance of credit bureaus is growing. The first reason of this is that the databases have currently accumulated enough statistics about borrowers to provide lenders with increasingly comprehensive information reports. The second reason is that the content of information available to the bureau is also growing. For example, currently, borrowers' photos are collected and stored in their credit history files helping to reduce potential risk of fraud. Yet, in such circumstances, the issue of effective implementation of the lender's interaction with the bureau into the general system of risk management is becoming increasingly important. Consequently, it calls for exploring the logic of interaction with other structural elements of risk management and defining its most effective option. General logic of risk management system is presented in [2,18]. The interaction credit bureaus with other components were considered for consumer lending in [12] and [15].

In our study, we analysed several conceptual approaches to the implementation of credit bureau services to the credit risk management system of consumer lending. As an assumption, we consider the situation when several credit bureaus operate in the market, as it represents the most common model in modern credit markets. Therefore, the model of work with bureaus involves the need to interact with several bureaus, as information provided by different bureaus may not overlap. However, an opposite argument in this case is that the higher cost of working with several bureaus will affect the economy of lending.

To illustrate the case, we assume that lenders interact with three major bureaus (for instance, this is a true case in Ukraine, where most lenders work with three credit bureaus). We distinguish two main models of the organization of interaction between the lender and the credit bureau:

- a model of sequential queries to different bureaus;
- a model of simultaneous queries to all bureaus.

The model of sequential queries is illustrated in Figure 1, and the model of parallel queries is presented in Figure 2.

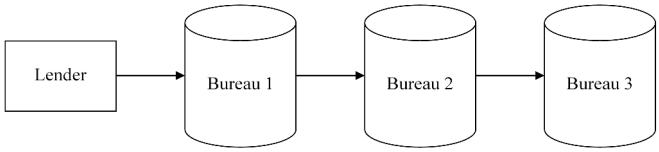


Figure 1. Model of sequential queries to credit bureaus.

Source: own elaboration.

The advantage of the first model is the reduction of risk management costs for the use of bureau services. If negative credit history is revealed in the first query to the Bureau 1, subsequent queries may be avoided. At the same time, it makes sense to make queries to the bureau on the basis of Hit Rate, i.e. the effectiveness of finding a credit history with the bureau. A disadvantage of this model is that fragmentary credit history is obtained, which does not allow to fully assess the risks. This refers to the situation when a query to Bureau 1 yields negative history, while Bureau 2 and Bureau 3 may hold positive credit history on other loans. For example, the negative credit history in the Bureau 1 may have been due to outstanding payments during the crisis period on the market.

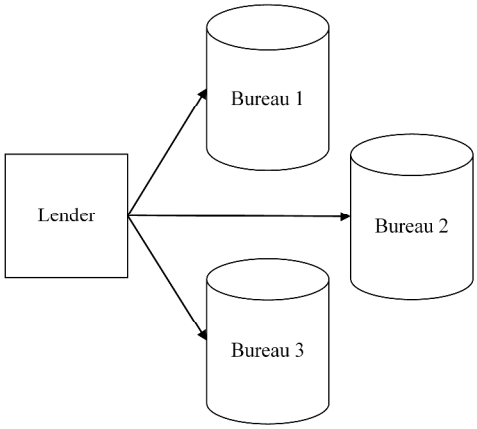


Figure 2. Model of simultaneous queries to credit bureaus.

Source: own elaboration.

The advantage of the second model is that it provides lenders with a comprehensive picture of the borrower's entire credit history. This allows making an

informed credit decision taking into account not only the negative aspects of credit history, but also the positive ones.

We studied the models of organization of the credit risk management system in the framework of interaction of queries to credit bureaus and other components of risk management, in particular, the 'black lists' and the internal application credit scoring. We distinguish between two main models of credit risk management in consumer lending. Their rationale is presented in Figure 3.

The first model involves initial check of the borrower against the internal 'black-lists' as well as other lender's procedures of identification and verification. A certain percentage of applications (e.g., according to statistical data in Ukraine, we estimate this level at 10%-15%) is rejected at this stage. Others potential borrowers are subjected to internal application scoring and creditworthiness check procedures. Here, the rejection rate is a little higher (we estimate it to reach 10%-20%). Following first two stages, 65%-80% of the potential borrowers from the incoming flow get checked through a query sent to the credit bureau. Based on information from the bureau, approximately 30% of borrowers with a negative credit history are rejected. Finally, the remaining 35%-50% of loan applications are approved.

In terms of scoring methodology, Figure 4 illustrates the borrower's assessment in the framework of the model. Area B represents the part of applicants' incoming flow rejected at the internal 'black lists' check. At the stage of scoring assessment, applications with a score below the cut-off point are rejected (darker area A on the left). Finally, at the third stage, rejection decisions are made on the basis of information received from credit bureaus, (area C in Fig. 4). Area D represents approved applications (potentially issued loans).

Using the second model of using credit bureaus as part of risk management system, credit bureau information, applications are initially checked with the bureau (and an area similar to C is rejected), then against the 'black lists' (area B), and eventually, another part of borrowers' applications are rejected following internal scoring (area A).

Both models described above have a number of advantages. The first model's advantage is that the costs it involves are lower. The queries to the credit bureaus are made only for those applicants who have successfully passed stages 1 and 2. Taking into account that 20% to 35% applications are rejected at these stages, it means the costs are reduced by similar values. For instance, a bank with an incoming inflow of 10,000 potential customers per month and paying a fee of $1 per query, may achieve cost savings in the amount of $2000-$3500 per month. One disadvantage of this model is the amount of time spent by banking personnel on the first two stages for clients with negative credit history. Thus, if processing of each application takes about one hour during the first two stages (filling in the application form, etc.), and then 30% of them are rejected following negative information from the credit bureau, then, additional expenses may reach up to 3600 hours per month.

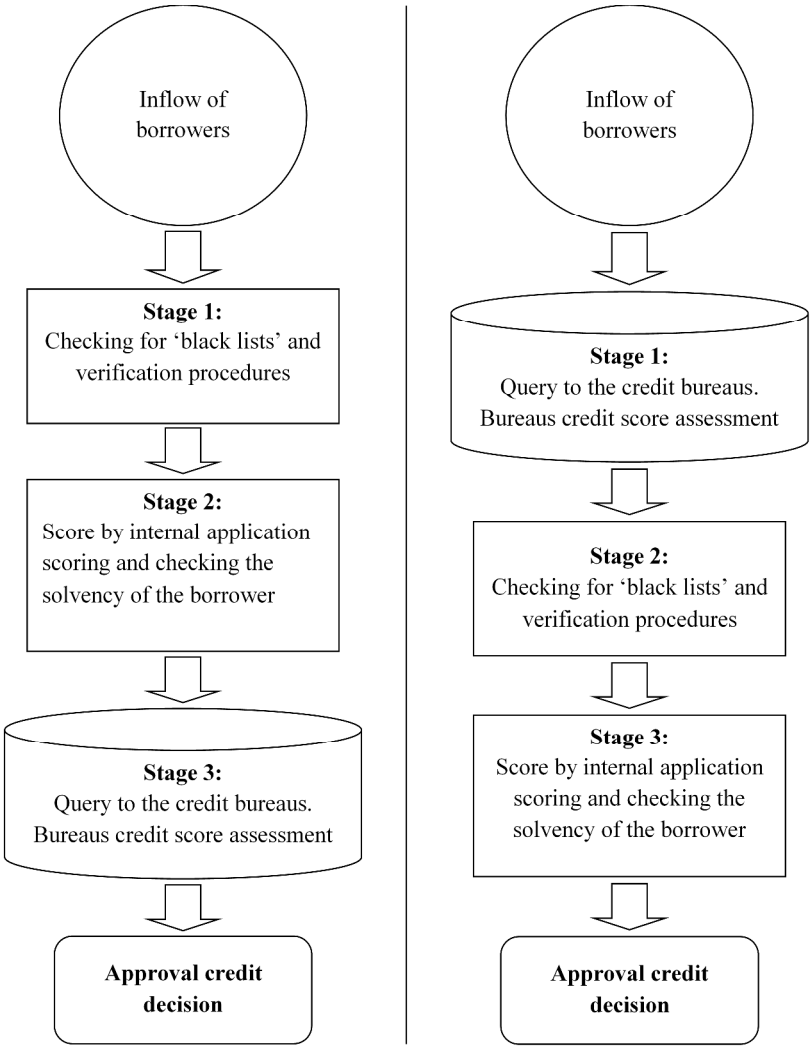


Fig. 3. Models of implementation of the credit bureau services in the system of credit risk management

Source: own elaboration.

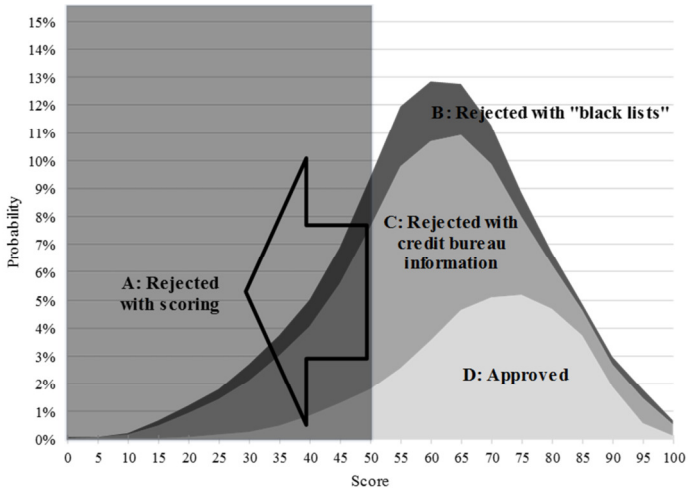


Figure. 4. Graphic illustration of processing borrower`s application

Source: own elaboration.

The second model, on the other hand, allows lender to save time spent on application processing. Having made query to the credit bureau and received negative information, lender can immediately stop application process and move to serving the next client. Operating efficiency in this case grows together with costs for credit bureau services.

The efficiency of these models may be defined by the state of the market. The first model would be preferable for lenders with a small application inflow. The second model can be effectively used in a dynamic market with large number of customers, in particular, with express loans and store loans. In order to evaluate the effectiveness of the models, we recommend comparing the cost of bureau services and the cost of applications processing for stages 1 and 2 of the first model.

Increasing effectiveness of risk management based on the use of credit bureau information

Nowadays, credit bureaus store huge amounts of market data. The idea of using bureaus' statistics for the study of economic problems was first proposed by David Burch in the 1970s. Burch used information from Dun's Market Identifiers (DMI), a private credit bureau, to find out the dependency of employment rates on firm relocations between US states ([3]).

There is a variety of ways how credit bureaus' data can be used to improve the effectiveness of risk management systems. Our study focuses on three aspects:

- reject service;
- statistical analysis of the incoming flow of applications; and
- benchmarking the lender's performance.

Reject service. In order to improve the efficiency of credit risk management system, we suggest using the statistical potential of credit bureaus for assessing loan applications that were rejected by lender at the verification stage (step 1) and at the application scoring stage (step 2). Analysis of rejection decisions by further monitoring the borrower's risks is called the analysis of rejected applications or the reject service. Feelders in [8] studied this phenomenon for commercial loans. Also, reject inference was analysed in [9].

To improve the effectiveness of verification process, it makes sense to first classify the rules making part of this process and being used to take the rejection decisions. Assume that rejections are based on k rules:P_1 ,…, P_k. Then, when lenders make queries to bureaus to receive additional information on borrowers whose applications were rejected in accordance with k rules, they can get information on whether other loans were granted after the rejection, and if they did, then whether they were paid back on time ("Good" loan status) or not ("Bad" loan status). A matrix of rejected applications can then be created shows at Table 1.

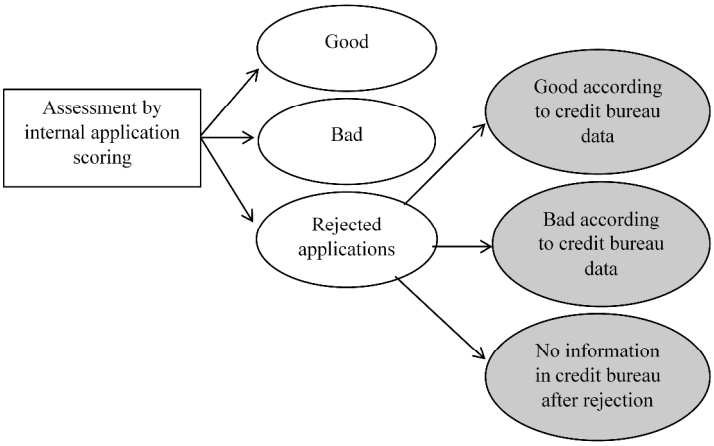
Table 1. Reject analysis for verification rules.

Verification rule	No loans received after rejection	Loan received after rejection and paid back on time	Loan received after rejection and not paid back on time
P_1	ND_1	G_1	B_1
…	…	…	…
P_k	ND_k	G_k	B_k

Source: own elaboration.

Economic analysis of rejected applications by certain rule helps compare and evaluate interest income from "good" and losses from "bad" cases. If the interest income from "good" exceeds the loss from "bad", verification should be changed by removing this rule.

Information provided by credit bureaus can be used for analysis of rejected applications based on application scoring values. Its logic is presented is Figure 5 and is based on consideration of "good" and "bad" results for rejected applications.



.

Figure 5. Rationale for analysis of reject service based on application scoring.

Source: own elaboration.

In the simplest application scoring model, applications go through one stage only where at a pre-defined cut-off point the decision is made on loan approval or rejection. Reject service analysis provides lenders with more detailed information on applicants whose loan applications had been rejected. Using this information, lenders can adjust decision-making rules in order to issue more loans in the future. It also allows improving the discriminatory power of application scoring by including additional information on good and bad statuses of borrowers' existing loans from credit bureau into classification of good and bad loans at the stage of scoring system development or upgrading.

Thus, information from the credit bureau allows additional statistical calculations to optimize the risk management system, in particular through the improvement of the verification and application scoring stages. Analysis of rejected applications (reject service) can improve the effectiveness of discriminating between good and bad applications at the first and second stages.

Let us consider example of application reject service procedure. Authors have applied described reject service logic for one Ukrainian financial company, which issued consumer loans in 2013-2017. During this period 33210 applications were rejected in credit granting. The rejected procedures were based on the model which present at the left part of Fig.3 ("black lists" and internal application scoring). The statistics from three basic Ukrainian bureaus which were analysed by authors indicate that 18421 (55,59%) rejected applicants have received loans in other financial institutions after rejections. The information in credit bureaus about other 14789 (=33210-18421) is absent after rejections. To all ap-

pearance, these rejected applicants were really "Bad" and any creditor did not want to grant loans for them.

18421 rejected applicants on the base of bureaus information were divided to 11347 ("Goods") and 7074 ("Bads"). "Goods" repaid loans successfully without delinquency more than 91 days. "Bads" had delinquency more 91 days and 93% of them did not go out from delinquency during one year after it begun.

We used statistical analysis for reject factors. Here we present factors which were included in internal application scoring. There were 13 basic application factors. First group of factors included socio-demographic factors: age, borrower gender, marital status, number of children, education level. Second group included professional factors: job position and time at employment. Third group of factors was devoted to welfare indicators: applicant`s monthly income, ownership indicators, time of car using. Fourth group was focused on loan characteristics: required amount of loan, duration of loan, recurring loan. Abovementioned characteristics formed core of internal application score, but did not exhaust all indicators for estimation.

We analysed Information Value (IV) statistic which is good screener for predictor variables of application scoring. Calculation of IV (see, for example [1], [20]) was done by formula (1):

$$IV = \sum_{i=1}^{n} (Good\_odds_i - Bad\_odds_i) \cdot \ln\left(\frac{Good\_odds_i}{Bad\_odds_i}\right) \cdot 100\% \qquad (1)$$

$Good\_odds_i$ is a share of "Good" applications for attribute i among all "Good" cases and Bad_odds_i is a share of "Bad" ones for attribute i amoung all "Bad" cases. According to IV methodology, its statistic values interpretation is:

- $IV < 0{,}02$ – factor has no predictive influence;
- $0{,}02 \leq IV < 0{,}1$ – low influence power;
- $0{,}1 \leq IV < 0{,}3$ – average influence power;
- $0{,}3 \leq IV < 0{,}5$ – statistically high predictive influence;
- $IV \geq 0{,}5$ – influence should be checked because of suspicious high influence power of factor.

Statistic IV is using in application scoring by signing higher weights to those predictors which have higher IV.

Firstly, we analysed statistics IV for Goods and Bads from creditor portfolio of loans. Then we expanded pool for analyses according to Fig.5. The results are presented at the Table 2 below.

Table 2. IV statistics for creditor`s data and for expanded data through reject service.

Factors, which included into application scoring	IV statistic for the data from creditor loans portfolio	IV statistic for the expanded data, which include rejective queries
Age	0.056120	0.054181
Borrower gender	0.001998	0.002449
Marital status	0.028445	0.027949
Number of children	0.020031	0.032251
Education level	0.017229	0.014506
Job position	0.019830	0.005123
Time at employment.	0.000212	0.002403
Applicant`s monthly income	0.123439	0.257549
Ownership indicators,	0.006214	0.007006
Time of car using	0.001763	0.001906
Required amount of loan	0.005119	0.088092
Duration of loan	0.000114	0.011906
Recurring loan	0.159000	0.297540

Source: authors` calculation.

IV statistics are similar for some predictors and differs for other. The differences in our case are concerned with “applicant`s monthly income”, “required amount of loan”, “duration of loan” and “recurring loan”. It means, that comparing this information values with ones, used for internal credit scoring development, lender can improve its scorecard, for example by correcting weights of risk factors or including additional factors, or change decision rules based on the results of the analysis.

Thus, information from the credit bureau allows for additional statistical calculations to optimize the risk management system, in particular through the improvement of the verification steps and application scoring decisions step.

Analysis of rejected applications (reject service) can improve the effectiveness of discrimination between goods and bads applications at the first and second stages.

4 Statistical analysis of applications inflow

The statistical potential of the credit bureau provides an opportunity to calculate the average market values of certain parameters of the borrower incoming flow. As example, at the Table 3 we illustrate market inflow at the form of risk distribution. Data reflect Ukrainian consumer credit market inflow for January-

September of 2017 year. R1-R15 are risk classes of International Bureau of Credit Histories (R1 indicate low risk, R15 indicate high risk).

Table 3. Inflow risk distribution in Ukrainian consumer credit market (%).

R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15
1.6	3.3	3.7	5.7	7.3	6.5	7.7	8.4	9.7	8.8	7.8	8.3	6.6	6.9	7.7

Source: authors` calculation on the base of [24].

They can be used to compare the level of inflow risk to market risk for each individual lender. Identifying this difference is directly related to the effectiveness of lender's risk management. Therefore, if the incoming flow of applications is riskier than market, risk management should be 'tougher'. Conversely, if the incoming flow is better compared to market average, then the risk management system may be 'softened'. Graphs in Figure 6 illustrate this approach. They show a comparison of the inflows of the entire market (in dark grey) against banks A and B (in light gray). Incoming flows are shown by the distribution functions of scores of potential borrowers applying separately to these banks, and in general on the market. R1-R15 are scoring classes (a total of 15 classes), each characterized by borrowers' default probability. Class R1 corresponds to almost zero probability of default, and R15 to 100%, other classes have interim default probability values between R1 and R15.

The upper graph shows that the inflow of borrowers to Bank A is characterized by a higher number of borrowers with lower credit risk than average on the market. Bank B's inflow, by contrast, is worse than the market average. This assessment is a unique tool provided by the bureau and raising the question: why is the flow of borrowers to the bank is worse than the market average? If lender's inflow can be improved by changes in marketing policies, then it will have impact on risk management.

5 Benchmarking effectiveness of the risk management system

Assessing effectiveness of credit risk management is a rather challenging task. Traditional indicators for such assessment include different types of overdue rates, rate of approved applications, rate of return on arrears, etc. However, these indicators, considered for an individual lender, do not reflect the impact of entire market conditions. A sound approach to evaluate lender performance is to use specific benchmarking that would reflect market average values. Then, comparison of individual lender's values to average market values gives an opportunity to assess their effectiveness properly. Among others this problems investigated in [14].

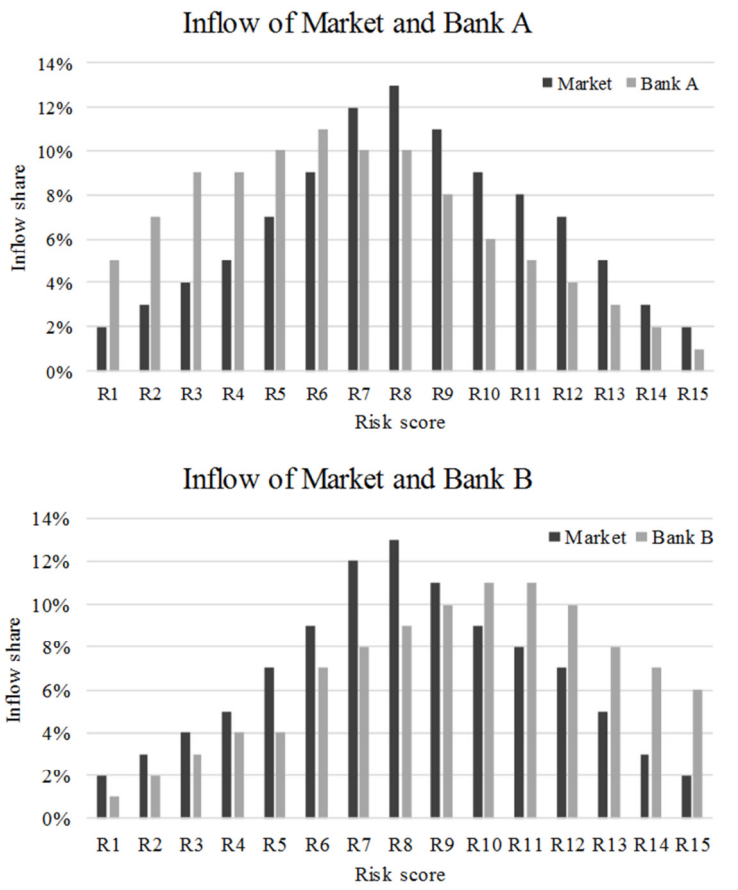


Figure 6. Comparison of market and bank inflows.

Source: authors` calculation on the base of [24].

Statistical data from credit bureaus provides lenders with meaningful information on the effectiveness of risk management indicators for the market in general. Based on them, lenders may develop benchmarking that will include dynamics of changes in the market.

We propose to consider the following five indices as indicators of the credit risk management effectiveness:

Bad Rate (BR). The percentage of borrowers with approved applications who have overdue payments of over 90 days. BR is one of basic indicators of the risk management system. Credit bureau data may be used to calculate the value of this indicator for the entire market or for a specific segment. Comparing BR of a particular lender to the market bad rate average helps assess the effectiveness of risk management in general.

First Payment Default (FPD). This is an indicator of the effectiveness of risk management in counteracting fraudulent actions. Failure to make the first payment on the loan, as a rule, is a sign of fraud.

Approval Rate (APR). This rate is a very important indicator that stands for the percentage of approved loans compared to application inflow. APR not only defines the effectiveness of differentiation between bad and good applications, but also illustrates the quality of the inflow. It is also an indirect indicator of the effectiveness of the risk management system.

Average Score Value (ASV). Bureau inflow ASV may be considered as an averaged indicator of inflow risk level. It is used to compare average score values for lender's inflow with similar market values. A more advanced approach may include comparing the distribution functions of the scoring values of the market inflow and values for the lender's inflow.

Rate of Collected Delinquency (RCD). This indicator compares the level of BR with overdue amounts repaid. RCD is an indicator of quality in managing overdue loans portfolio. Based on data from credit bureau, this rate can be calculated for the market and then compares to lender's RCD.

Benchmarking model based on abovementioned indicators is presented in Figure 7.

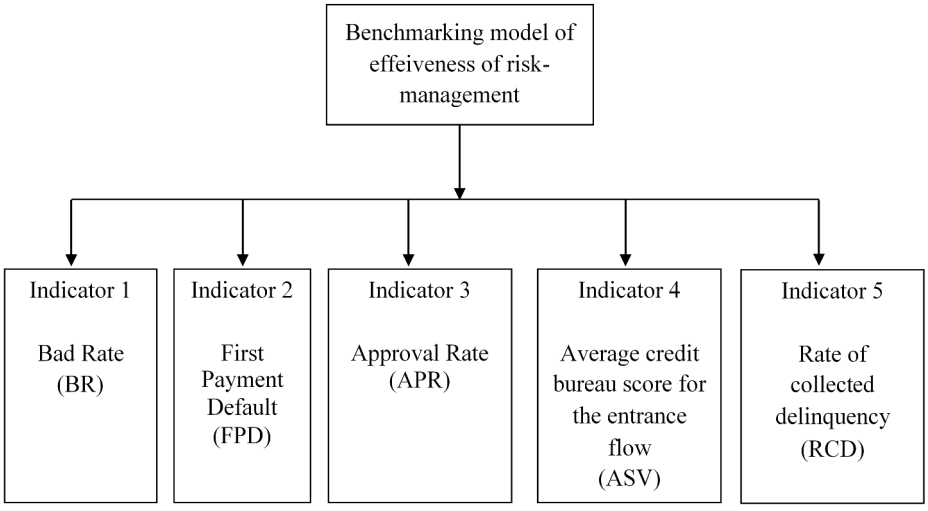


Figure 7. Benchmarking model based on credit bureau data.

Source: own elaboration.

To illustrate how the model works, let us consider it for two lenders A and B, whose data on risk management is presented in Table 4.

Table 4. Benchmarking of two lenders against market average

	Market (credit bureau statistics)	Lender A	Lender B
BR (%)	8%	6%	15%
FPD (%)	1,5%	1,5%	2%
APR (%)	40%	50%	60%
ASV (from 0 to 100)	54	62	49
RSD	8%	11%	4%

Source: own elaboration.

Taking market indicators as base value (value equals 1) for comparison, we have obtained the following benchmarking result, presented in Figure 8.

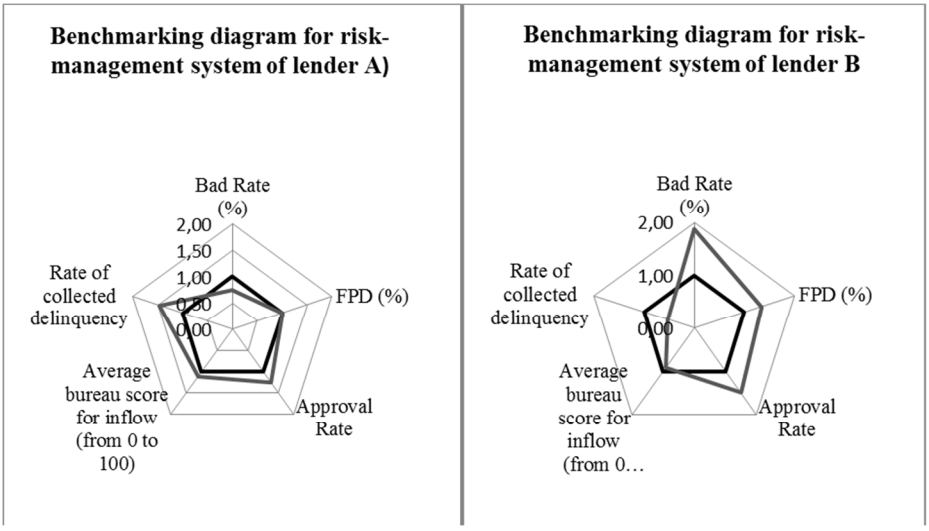


Figure 8. Benchmarking for lenders A and B: comparison of effectiveness of risk-management systems.

Source: own elaboration.

From comparison of the parameters of the risk management systems of lender A and market average, we conclude that the position of lender A is better than market average. Conversely, indicators of lender B are worse than market average. Therefore, the risk management system of lender A is considered more effective than that of lender B.

Naturally, benchmarking for assessing the effectiveness of risk-management may include more indicators that could be calculated using credit bureau data.

6 Conclusions

Credit bureaus today are inherent constituent of credit market. Especially, they are highly developing at the segment of consumer lending. Creditors include queries to bureau into the loan issuing procedure, especially when dealing with risk assessment risk procedures. Moreover, credit bureaus accumulate great volume of different statistics. These statistics using form good potential for analysis and solving different economic problems. This potential may be use more intensively. Our findings suggest that data stored by credit bureaus presents significant opportunities for improving effectiveness of credit risk management systems.

Indeed, implementation of effective credit risk-management system is a very important objective for lenders. Credit risk management systems are complex for large banks. All procedures have been realized automatically. Verification of effectiveness may be characterizes by non-clear, fuzzy results. Statistics of credit bureaus provide some good instruments for increasing effectiveness risk management systems. Comprehensive analysis of rejected applications (reject service), analysis of application inflow and benchmarking of the risk management system effectiveness based on data collected by credit bureaus has significant impact on the effectiveness of lenders' risk management systems

Bibliography

[1] Anderson R., *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*, Oxford University Press, 2007.
[2] Bessis J., *Risk Management in Banking*, John Wiley & Sons Ltd, 2002.
[3] Birch D. *The Job Generation Process*, –Cambridge, MA: MIT Program on Neighborhood and Regional Change, 1979.
[4] Bos J., Haas R., Millone M., *Show Me Yours and I'll Show You Mine: Sharing Borrower Information in a Competitive Credit Market*, "CentER Discussion Paper Series", No. 2015-027.
[5] *Credit Reporting Systems and the International Economy*, ed. M. Miller, Cambridge: MIT Press, 2003.
[6] Dattakumar R., Jagadeesh R., *A review of literature on benchmarking*, "Benchmarking: An International Journal", V.10, Iss: 3, 2003, 176–209.
[7] Djankov S., McLiesh C., Shleifer A. *Private credit in 129 countries*, Journal "Journal of Financial Economics", V. 84., Iss. 2., 2007, 299–329.
[8] Feelders A., *Credit scoring and reject inference with mixture models*, "International Journal of Intelligent System in Accounting, Finance and Management", V. 8., 1999, 271–279.
[9] Hand D. Henley W., *Can reject inference ever work?*, "IMA Journal of Mathematics Applied in Business & Industry", V.5, 1994, 45–55.

[10] Jappelli T., Pagano M., *Information sharing, lending and defaults: Cross-country evidence*, "Journal of Banking & Finance", V.26, 2002, 2017–2045.
[11] Kaminsky A., *Genesis and structure of credit bureau system in Ukraine, "*Bulletin of Taras Shevchenko National University of Kyiv. Economics", V. 10 – (151), 2013, 31–36.
[12] Kaminsky A., *Credit bureau in risk-management system of consumer crediting, "*Business-Inform", № 4, 2013, 372-376.
[13] Kaminsky A., *The assessment of bank market position on the base of bureau of credit histories benchmarking*, "Visnyk of the Lviv University. Series Economics", V. 50, 2013, 133–140.
[14] Kaminsky A., *Credit bureau benchmarking as a tool for estimation of bank's position at the market*, "Bulletin of Taras Shevchenko National University of Kyiv. Economics", V. 1(166), 2015, 60–64.
[15] Kaminsky A., *Methodological fundamentals of credit bureau potential using in credit activity*, "Scientific papers NaUKMA. Economics", V. 17, 2015, 38–43.
[16] Pysanets K., *Selection of instruments for credit scoring systems development*, "Nauka i Studia", V.23, Przemyśl, 2013, 23–34.
[17] Rothemund M., Gerhardt M., *The European Credit Information Landscape*, "Financial Markets ECRI Research Report", January 2011.
[18] Schroeck G., *Risk Management and Value Creation in Financial Institutions*, John Wiley & Sons Ltd., 2002.
[19] Simovic V., *The impact of the functional characteristics of a credit bureau on the level of indebtedness per capita: Evidence from East European countries*," Baltic Journal of Economics", V.11, Iss. 2, 2011, 101–130.
[20] Thomas L., Edelman D., Crook J., *Mathematical Modeling and Computation Credit Scoring and Its Applications. – Series: Mathematical Modeling and Computation*, 2002.
[21] https://www.bik.pl.
[22] http://ubki.ua.
[23] http://www.pvbki.com.
[24] https://credithistory.com.ua.
[25] https://bank.gov.ua.

Tomasz Warowny[1]

Estimating the probabilities of a simultaneous occurrence of random phenomena

Keywords: probability, random variables, uniform distribution, binomial distribution, normal distribution.

Abstract

In this paper random variables with uniform, binomial and Gauss distributions are considered. Those variables describe the times of the start and duration of random phenomena. It was shown how to estimate the probability of simultaneous occurrence of selected phenomena in the fixed period. Relevant definitions, formulas and examples are provided.

1 Introduction

In many areas of their activity, people make decisions in conditions of uncertainty because many phenomena that influence their behaviour and decisions are carried out in a random way. These can be economic, social and natural phenomena. It is difficult to predict them all, determine when they will occur and how long they will last. For example, when investing in stock market shares, the investor would like to see no bear market during his investment; the person planning a holiday would like to have nice weather in the chosen time; in the production process it is advisable that too many machines not to fail at the same time to avoid production downtime. Therefore it is necessary that phenomena for which there is no certainty as to the time of their occurrence are described by means of random variables. In case of many phenomena it is often important to know the value of the likelihood of simultaneous occurrence. The paper presents, with certain assumptions about the distributions of random variables, how to describe them and how to estimate the probabilities of their simultaneous occurrence. Relevant definitions, formulas and examples are also provided.

[1] Department of Quantitative Methods in Management, Faculty of Management, Lublin University of Technology, e-mail: t.warowny@pollub.pl

2 Description of phenomena by means of random variables with the uniform distribution

Definition. Random variable X has a uniform distribution on a measurable set A, if for each measurable set B is

$$P(X \in B) = \frac{m(A \cap B)}{m(A)},$$

where m denotes the measure of a set.

Let us consider two phenomena denoted by Z_1 and Z_2, respectively. Let us introduce the following notations and assumptions:

- X- random variable representing the moment of the beginning of the phenomenon Z_1,
- Δx - time of duration of the phenomenon Z_1, it is a fixed value,
- Y - random variable representing the moment of the beginning of the phenomenon Z_2,
- Δy - time of duration of the phenomenon Z_2, it is a fixed value,
- X, Y- independent random variables with the uniform distributions on the interval $[0, T]$.

Let us calculate the probability of an event „phenomena will last simultaneously at least $\Delta t \geq 0$". It is obvious that the condition $\Delta x, \Delta y \geq \Delta t$ must be satisfied.

Let us consider two cases.

Case 1. The phenomenon Z_1 will begin not later than the phenomenon Z_2. In order for them to occur simultaneously for at least Δt, the following conditions must be met

$$\boldsymbol{Y \geq X} \qquad \text{and} \qquad \boldsymbol{Y \leq X + \Delta x - \Delta t.}$$

The above statement is illustrated by Fig. 1.

In many areas of their activity, people make decisions in conditions of uncertainty because many phenomena that influence their behaviour and decisions are carried out.

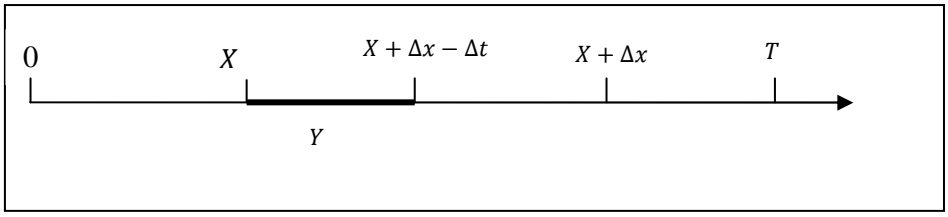


Fig. 1. A geometrical interpretation of Case 1.

Source: own elaboration.

Case 2. The phenomenon Z_2 will begin not later than the phenomenon Z_1. In order for them to occur simultaneously for at least Δt the following conditions must be met

$$\boldsymbol{Y \leq X} \qquad \text{and} \qquad \boldsymbol{X \leq Y + \Delta y - \Delta t},$$

what is illustrated by Fig. 2.

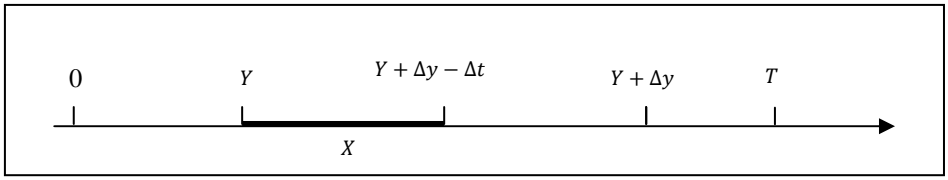


Fig. 2. A geometrical interpretation of Case 2.

Source: own elaboration.

Thus we have

$$\begin{cases} \boldsymbol{Y \geq X} \\ \boldsymbol{Y \leq X + \Delta x - \Delta t} \end{cases} \qquad \text{or} \qquad \begin{cases} \boldsymbol{Y \leq X} \\ \boldsymbol{Y \geq X - \Delta y + \Delta t}. \end{cases}$$

The set of points with coordinates (X, Y) satisfying the above conditions is a set of lines on Fig. 3.

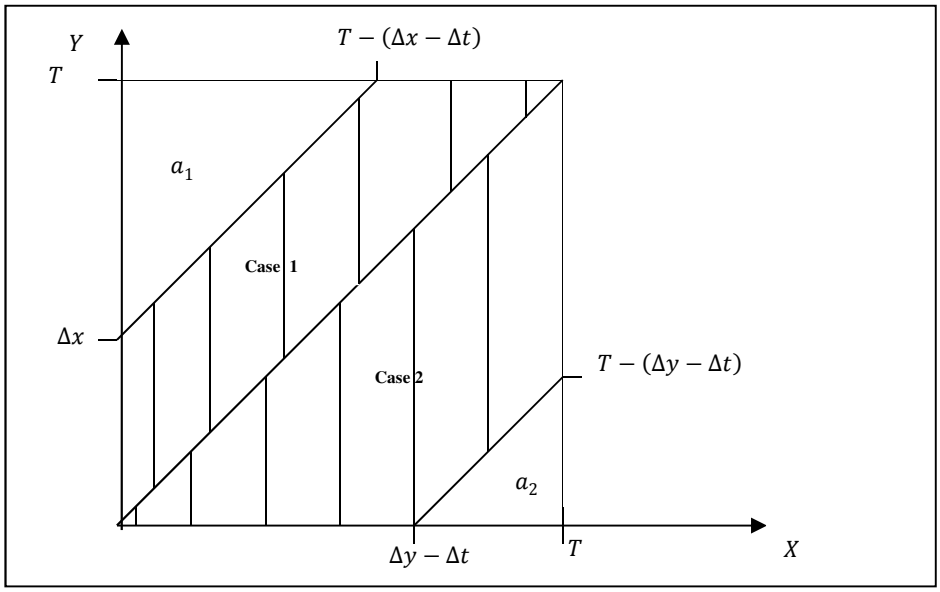


Fig. 3. A geometrical interpretation of an event „the phenomena will occur for at least Δt".

Source: own elaboration.

By the definition of the uniform distribution we conclude that t the probability of an event "the phenomena will occur simultaneously for at least Δt" is equal to

$$P = \frac{T^2 - a_1 - a_2}{T^2}, \tag{1}$$

where:

$$a_1 = \frac{[T - (\Delta x - \Delta t)]^2}{2}, a_2 = \frac{[T - (\Delta y - \Delta t)]^2}{2},$$

what results in

$$P = \frac{2T^2 - [T - (\Delta x - \Delta t)]^2 - [T - (\Delta y - \Delta t)]^2}{2T^2}. \tag{2}$$

Example 1. Two machines operate independently one of another. Statistically, in each of them a piece (bearing) breaks down once a month (720 hours). Replacement of the bearing stops the first machine for 20 hours, the second one for 15 hours. The production process is stopped if the machines are stopped simultaneously for more than 5 hours. What is the probability that within a month the production will be stopped?
We have: $T = 720, \Delta x = 20, \Delta y = 15, \Delta t = 5$.
By using (2) we obtain

$$P = \frac{2 \cdot 720^2 - [720 - (20 - 5)]^2 - [720 - (15 - 5)]^2}{2 \cdot 720^2} = 0.034.$$

The probability that within a month the production will be stopped is then equal to 0.034.

In the case of many decision problems, it may often be important to answer the question: how long must the selected phenomenon last (e.g. Δx) so that the probability in question had a fixed value?

To answer the above question it is necessary to solve a quadratic equation of the variable Δx in the following form:

$$P = \frac{2T^2 - [T - (\Delta x - \Delta t)]^2 - [T - (\Delta y - \Delta t)]^2}{2T^2} = \beta, \tag{3}$$

where β is the probability required by the decision maker.

Addressing Example 1 let us consider another example.

Example 2. How much should the time of replacement of the broken bearing in the first machine be decreased so that the probability of stopping the production process would be 0.02?
We have: $T = 720, \Delta y = 15, \Delta t = 5, \beta = 0{,}02, \Delta x =?$

The solution of the equation

$$\frac{2T^2 - [T - (\Delta x - \Delta t)]^2 - [T - (\Delta y - \Delta t)]^2}{2T^2} = \beta = 0.02$$

is $\Delta x = 9.48$. Replacement of the broken bearing in the first machine should last about 9.5, the probability of stopping the production process would be 0.02.

When we assume in (2) that $\Delta t \to 0$, we obtain

$$P = \frac{2T^2 - [T - (\Delta x - \Delta t)]^2 - [T - (\Delta y - \Delta t)]^2}{2T^2} \to \frac{2T^2 - [T - \Delta x]^2 - [T - \Delta y]^2}{2T^2}. \tag{4}$$

Formula (4) is the probability of an event that two phenomena will occur simultaneously at any given time. Hence we obtain that both phenomena will not occur simultaneously with probability equal

$$1 - P = P' = \frac{2T^2 - 2T(\Delta x + \Delta y) + (\Delta x)^2 + (\Delta y)^2}{2T^2} \tag{5}$$

Example 3. The environmental guard received information from a trusted source that the next night, the trash from the production facility would be dumped into the forest. The exact time is unknown, but it is known that this illegal procedure will start between 22:00 and 6:00 and will last for 2 hours. Due to the workload, the guard will be able to appear at the crime scene at a random, independent moment, and be able to stay there for no longer than an hour. In order that the dishonest owner of the production facility could be punished, he must be caught in the act. What is the probability that the guard will be able to witness the dumping of trash into the forest?

We have: $T = 8$, $\Delta x = 2$, $\Delta y = 1$. After using (5) we obtain

$$P = \frac{2T(\Delta x + \Delta y) - (\Delta x)^2 - (\Delta y)^2}{2T^2} = \frac{2 \cdot 8 \cdot (2 + 1) - 2^2 - 1^2}{2 \cdot 8^2} = 0{,}336.$$

A guard who can spend one hour on waiting has about 33% chance of catching the deceitful owner of the plant.

Suppose the inspector wants the probability to be equal at least β. How much time would he have to spend on waiting in the woods? In this case, the inequality of the variable Δy must be solved

$$P = \frac{2T(\Delta x + \Delta y) - (\Delta x)^2 - (\Delta y)^2}{2T^2} \geq \beta.$$

After transforming the above quadratic inequality to the form

$$-(\Delta y)^2 + 2T \cdot \Delta y + 2T \cdot \Delta x - (\Delta x)^2 - 2\beta T^2 \geq 0$$

it is easy to determine its solution which has the form of

$$\Delta y \in \left[T - \frac{\sqrt{\Lambda}}{2}, T + \frac{\sqrt{\Lambda}}{2}\right],$$

where $\Lambda = 4(T^2 + 2T \cdot \Delta x - (\Delta x)^2 - 2\beta T^2)$. Taking into account $\Delta y \leq T$ we obtain

$$\Delta y \in \left[T - \frac{\sqrt{\Lambda}}{2}, T\right]. \tag{6}$$

Suppose the inspector wants the probability of proving the crime to be at least 0.6. We have $T = 8$, $\Delta x = 2$, $\beta = 0{,}6$. By substituting these values into formula (6) we obtain $4{,}1 \leq \Delta y \leq 8$. The inspector would have to reserve at leaat 4 hours and 6 minutes for waiting.

3 Description of phenomena by means of random variables with the binomial distribution

Until now, we have assumed that each of the considered phenomena would certainly occur. The time of their beginning was random. Let's assume now that the phenomena at a given time may exist (but not necessarily) with certain probabilities. Due to the limited scope of this paper, only certain specific cases will be considered. Situations where many phenomena will occur in a given period will be considered, and the probability of simultaneous occurrence of a fixed number of those phenomena will be estimated. The following definitions and properties will be needed.

Definition. A random variable S_n has a binomial distribution with the parameters (n, p), $n \in \mathrm{N}$, $0 < p < 1$ if

$$P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, \ k = 0{,}1, \ldots, n. \tag{7}$$

If p is the probability of occurrence of some event in a single experiment, then $P(S_n = k)$ is the probability of occurring this event exactly k time in n experiments. It can be shown that the expected value of the above random variable is $E(S_n) = np$ and the variance is $s^2(S_n) = np(1-p)$. Variable S_n is often referred to as the number of successes in Bernoulli's scheme.

Example 4. The probability that within one working day the machine will crash is $p = 0{,}1$. Failure will stop the machine until the end of the day. The possibilities of subsequent failures are independent of each other. What is the probability that in the next 30 days the machine will be immobilized for no more than 5 days?

Let random variable have binomial distribution with parameters (n, p). Let us compute

$P(S_{30} \leq 5) = P(S_{30} = 0 \vee S_{30} = 1 \vee S_{30} = 2 \vee S_{30} = 3 \vee S_{30} = 4 \vee S_{30} = 5)$
$= P(S_{30} = 0) + P(S_{30} = 1) + P(S_{30} = 2) + P(S_{30} = 3) + P(S_{30} = 4) +$
$+P(S_{30} = 5).$

Using (7) we obtain

$$P(S_{30} \leq 5) = \sum_{k=0}^{6} \binom{30}{k} \cdot 0,1^k \cdot 0,9^{30-k} = 0,927.$$

To estimate the probabilities of the simultaneous occurrence of phenomena described by binomial distribution random variables, the de Moivre-Laplace theorem is often used as a special case of the central limit theorem [1, p.350 and later].

The de Moivre–Laplace theorem. If S_n is a sequence of binomial random variables with parameters (n, p), then for any real numbers α_1, α_1 such that $\alpha_1 < \alpha_2$ there is a formula:

$$\lim_{n\to\infty} P\left(\alpha_1 < \frac{S_n - np}{\sqrt{np(1-p)}} < \alpha_2\right) = F(\alpha_2) - F(\alpha_1),$$

where F is a distribution function of normal distribution.

The above statement can also be formulated as follows:

$$P\left(\frac{S_n - np}{\sqrt{np(1-p)}} < \alpha\right) \to F(\alpha), \qquad n \to \infty, \qquad \alpha \in R$$

what means that the random variable $\frac{S_n - np}{\sqrt{np(1-p)}}$ is asymptotically normal.

Remark. If Y is a random variable of the continuous type, then $P(Y \leq \alpha) = P(Y < \alpha)$. For discrete-type random variables, such an inequality does not need to be hold. If S_nt is a sequence of random variables with a binomial distribution with parameters (n, p), hence discrete-type ones, then for non-negative integers $k_1, k_2$the following formula is often used in the following way [3, p.235]:

$$P(k_1 \leq S_n \leq k_2) = P(k_1 - 0,5 < S_n < k_2 + 0,5) =$$
$$= P\left(\frac{k_1 - 0,5 - np}{\sqrt{np(1-p)}} < \frac{S_n - np}{\sqrt{np(1-p)}} < \frac{k_2 + 0,5 - np}{\sqrt{np(1-p)}}\right).$$

Then, by the de Moivre – Laplace theorem we obtain

$$\lim_{n\to\infty} P(k_1 \leq S_n \leq k_2) = F\left(\frac{k_2 + 0,5 - np}{\sqrt{np(1-p)}}\right) - F\left(\frac{k_1 - 0,5 - np}{\sqrt{np(1-p)}}\right).$$

It is possible then to apply the following approximation

$$P(k_1 \le S_n \le k_2) \approx F\left(\frac{k_2 + 0{,}5 - np}{\sqrt{np(1-p)}}\right) - F\left(\frac{k_1 - 0{,}5 - np}{\sqrt{np(1-p)}}\right). \quad (8)$$

If $k_1 = k_2 = k$, then

$$P(k \le S_n \le k) = P(S_n = k) \approx F\left(\frac{k + 0{,}5 - np}{\sqrt{np(1-p)}}\right) - F\left(\frac{k - 0{,}5 - np}{\sqrt{np(1-p)}}\right). \quad (9)$$

Simultaneously we know that $P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

Estimation of precision of approximation in the de Moivre – Laplace theorem is following [2, p.173 and later]:

Example 6. In an office facility 1000 light bulbs are lit. For each light bulb, the probability of blurring is 0.1 during the day.

a) What is the probability that no more than 100 bulbs will burn in the day?
b) What is the probability that over 200 bulbs burned during the day?
c) What is the probability that the number of burnt bulbs is between 100 and 120 during the day?

Solution. Let S_n be the number of successes in the Bernoulli scheme (by success we mean the burning of the bulb). The probability of success is equal

$p = 0.1; n = 1000.$

a) We have to calculate $P(S_n \le 100)$. Using (7) we have

$P(S_n \le 100)=$

$$= P(S_n = 100) + P(S_n = 99) + \ldots + P(S_n = 1) + P(S_n = 0) = 0.5266.$$

If we use de Moivre–Laplace theorem, we will obtain

$$P(S_n \le z) = P\left(\frac{S_n - np}{\sqrt{np(1-p)}} \le \frac{z - np}{\sqrt{np(1-p)}}\right) \approx F\left(\frac{z - np}{\sqrt{np(1-p)}}\right).$$

Hence

$$P(S_n \le 100) \approx F\left(\frac{100 - 1000 \cdot 0{,}1}{\sqrt{9.49}}\right) = F(0) = 0.$$

The probability that no more than 100 light bulbs will burn during the day is equal to 0.5266.

b) We have to calculate $P(S_n > 200) = 1 - P(S_n \le 200)$.

$$P(S_n \le 200) \approx F\left(\frac{200 - 1000 \cdot 0{,}1}{\sqrt{9.49}}\right) = F(32.46) \approx 1.$$

The probability that more than 200 light bulbs will burn during the day is close to zero.

c) We have to calculate $P(100 \le S_n \le 120)$. Using (8) we have

$$P(k_1 \le S_n \le k_2) \approx F\left(\frac{k_2 + 0{,}5 - np}{\sqrt{np(1-p)}}\right) - F\left(\frac{k_1 - 0{,}5 - np}{\sqrt{np(1-p)}}\right).$$

Hence

$$P(100 \le S_n \le 120) \approx F\left(\frac{120 + 0{,}5 - 100}{\sqrt{9.49}}\right) - F\left(\frac{100 - 0{,}5 - 100}{\sqrt{9.49}}\right)$$
$$\approx 1 - 0.435 = 0.565.$$

Finally, let us recall the commonly known property of normal distribution called the three sigma rule.

Property. If X is a random variable with a normal distribution with parameters m, s, then $P(m - 3s \le X \le m + 3s) \approx 0.997$.

We know that if the random variable has S_n, a binomial distribution with parameters (n, p), then the expected value of S_n is equal to $E(S_n) = np$, and the standard deviation is equal to $s(S_n) = \sqrt{np(1-p)}$. By the de Moivre-Laplace theorem, $\frac{S_n - np}{\sqrt{np(1-p)}}$ has an asymptotically normal distribution with parameters 0 and 1. Based on the above property, we obtain that

$$P\left(-3 \le \frac{S_n - np}{\sqrt{np(1-p)}} \le 3\right) \approx 0.997$$

or

$$P\left(-3\sqrt{np(1-p)} + np \le S_n \le 3\sqrt{np(1-p)} + np\right) \approx 0.997.$$

The above means that the number of successes in the Bernoulli scheme with the probability of near to 1 is within the range

$$\left(-3\sqrt{np(1-p)} + np, 3\sqrt{np(1-p)} + np\right).$$

In Example 6 we have $n = 1000$, $p = 0.1$, so the interval is

$$\left(-3\sqrt{1000 \cdot 0.1 \cdot 0.9} + 1000 \cdot 0.1; 3\sqrt{1000 \cdot 0.1 \cdot 0.9} + 1000 \cdot 0.1\right)$$
$$= (90.76; 109.24).$$

This means that with the probability close to 1, the number of light bulbs that will burn during the day is between 91 and 109,
then $P(m - 3s \le X \le m + 3s) \approx 0.997$.

Example 7. Two cark parks with separate entrances are planned to be built near the airport. It is estimated that in any time of the day there will be parked 500 cars by average in the both car parks together. We assume that cars will be arriving separately and their drivers will choose one of the two car parks with probability $p = 0.5$, independently one from another. What should be the smallest

number of places on each of the car parks so that there would be a free place for each of incoming cars with probability 99%?

Solution. Let S_n be a number of successes in the Bernoulli scheme. By "success" we mean a choice of one of the car parks and finding a free place on it. The probability of the success is $p = 0.5, n = 500$.

By using the de Moivre–Laplace theorem we must find such a number z, so that

$$P(S_{500} \leq z) = 0.99.$$

We know that

$$P(S_n \leq z) = P\left(\frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{z - np}{\sqrt{np(1-p)}}\right) \approx F\left(\frac{z - np}{\sqrt{np(1-p)}}\right).$$

Hence

$$F\left(\frac{z - np}{\sqrt{np(1-p)}}\right) = 0.99,$$

so

$$\frac{z - np}{\sqrt{np(1-p)}} = F^{-1}(0.99) = 2.3263.$$

We obtain

$$\frac{z - 500 \cdot 0.5}{\sqrt{500 \cdot 0.5 \cdot (1 - 0.5)}} = 2.3263,$$

so $z = 276$.

There should be 276 places on each of the car parks so that there would be a free place for each of incoming cars with probability 0.99.

4 Final conclusions

In many areas of human activity, the decision-making process may be influenced by a variety of phenomena that may occur in the future and have a significant impact on the outcome of the decisions. Simultaneous occurrences of many random phenomena can bring both positive and negative effects.

In this paper it is presented how phenomena can be described with random occurrences. For example, random variables with uniform and binomial distribution are used. It is presented how to calculate the probabilities of simultaneous occurrence of these phenomena. The knowledge of these probabilities, in many cases, may limit the possibility of making a wrong decision and thus reduce its negative effects.

Bibliography

[1] Billlingsley P., *Probability and Measure*, 3rd edition, John Wiley & Sons. Inc., New York 1995.

[2] Jakubowski J., Sztencel R., *Wstęp do teorii prawdopodobieństwa*, wyd. 2, SCRIPT, Warszawa 2001.

[3] Krysicki W., Bartos J., Dyczka W., Królikowska A., Wasilewski M., *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach, część I, Rachunek prawdopodobieństwa*, Wydawnictwo Naukowe PWN, Warszawa 2003.