On option pricing in illiquid markets with jumps

Youssef El-Khatib\textsuperscript{1} and Abdulnasser Hatemi-J\textsuperscript{2}

\textsuperscript{1} UAE University, Department of Mathematical Sciences, Al-Ain, P.O. Box 17551, United Arab Emirates (E-mail: Youssef.Elkhathib@uaeu.ac.ae)

\textsuperscript{2} UAE University, Department of Economics and Finance, Al-Ain, P.O. Box 17555, United Arab Emirates (E-mail: Ahatemi@uaeu.ac.ae)

Abstract. One of the shortcomings of the Black and Scholes model on option pricing is the assumption that trading of the underlying asset does not affect the price of that asset. This assumption can be fulfilled only in perfectly liquid markets. Since most markets are illiquid, this assumption might be too restrictive. Thus, taking into account the price impact in option pricing is an important issue. This issue has been dealt with, to some extent, for illiquid markets by assuming a continuous process, mainly based on the Brownian motion. However, the recent financial crisis and its effects on the global stock markets have propagated the urgent need for more realistic models where the stochastic process describing the price trajectories involves random jumps. Nonetheless, works related to markets with jumps are scant compared to the continuous ones. In addition, these previous studies do not deal with illiquid markets. The contribution of this paper is to tackle the pricing problem for options in illiquid markets with jumps as well as the hedging strategy within this context, which is the first of its kind to the best knowledge.

Keywords: Options pricing, Illiquid markets, Jump diffusion, Incomplete markets.

1 Introduction

Financial derivatives are important tools for dealing with financial risk. An option is an example of such derivatives, which gives the right but not the obligation, to engage in a future transaction on some underlying financial asset. For instance, a European call option on an asset with the price $S_t \in [0,T]$ is a contract between two agents (buyer and seller), which gives the holder the right to buy the asset at a pre-specified future time $T$ (the expiration date) for an amount $K$ (called the strike). The buyer of the option is not obliged to exercise the option. When the contract is issued they buyer of the option needs to pay a certain amount of money called the premium. The payoff for this option is defined as $h(S_T) = \max(S_T - K, 0) = (S_T - K)^+$. The writer of the option receives a premium that is invested in the combination of the risky and risk free assets. The pricing problem is then to determine the premium, i.e. the price that the seller should charge for this option.

The pricing problem has been solved in the pioneer work of Black and Scholes\textsuperscript{2}. One of the shortcomings of the Black and Scholes model is the assumption that an option trader cannot affect the underlying asset price. However, it
is well-known that in a market with imperfect liquidity, trading does affect the underlying asset price (see, for example, Chan and Lakonishok[1], Keim and Madhavan[6], and Sharpe et al.[10]).

In Liu and Yong[7], the authors study the effect of the replication of a European option on the underlying asset price. They obtain a generalization of the Black Scholes pricing P.D.E. as the following:

$$\frac{\partial v}{\partial t}(S,t) + \frac{\sigma^2 S^2}{2} \frac{\partial^2 v}{\partial S^2}(S,t) - r v(S,t) = 0, \quad \text{for } (S,t) \in [0, +\infty[ \times [0,T]$$

$$v(S,T) = f(S), \quad 0 < S < \infty,$$

where $\lambda(S,t)$ is the price impact function of the trader. The classical BlackScholes P.D.E. is a special case of (2) when $\lambda(S,t) = 0$.

There are also several other papers that have studied the financial markets with jumps (among others, Merton[8], Dritschel and Protter[3], El-Khatib and Privault[4]) and El-Khatib and Al-Mdallal[5]. However, none of the previous studies based on the jump-diffusion approach deals with illiquid markets, to the best knowledge. This paper is extends the model of Liu and Yong[7] by including a jump-diffusion structure in the underlying option pricing model. This appears to be an important issue because the model that is suggested in this paper allows for the possibility to account for sudden and random significant changes in the market that might not be captured by the existing models in the literature such as the continuous model suggested by Liu and Yong[7]. Hence, the approach that is developed in this paper is expected to be more useful in financial risk management, especially in the cases in which the financial markets are under stress.

The disposition of the rest of the paper is the following. Section 2 introduces the jump-diffusion model for an illiquid market. Section 3 deals with the pricing problem of an option within the context of a jump-diffusion model along with the proof for the suggested solution. Section 4 concludes the paper.

2 A jump-diffusion model for illiquid markets

We start with presenting some necessary denotations. Let $(N_t)_{t \in [0,T]}$ be a Poisson process with deterministic intensity $\rho$. Let also $M_t = N_t - pt$ be its associated compensated process. The process $(B_t)_{t \in [0,T]}$ denotes a Brownian motion. The probability space of interest is $(\Omega, \mathcal{F}, P)$ with $(M_t)_{t \in [0,T]}$ and $(B_t)_{t \in [0,T]}$ being independent. Let $(\mathcal{F}_t)_{t \in [0,T]}$ signify the filtration generated by $(N_t)_{t \in [0,T]}$ and $(B_t)_{t \in [0,T]}$. The market is assumed to have two assets: a risky asset $(S_t)_{t \in [0,T]}$ and a risk-free denoted by $(A_t)_{t \in [0,T]}$. The maturity is $T$, the strike is $K$ and the payoff is $h(S_T) = (S_T - K)^+ = \max\{S_T - K, 0\}$. As in Liu and Yong[7], the return on the risk free asset indirectly depends on $S_t$ and the option trader’s trading in the stock market has a direct impact on the
stock price. This price impact, which an investor can cause by trading on an asset, functions in such way that it increases the price when buying the asset and it decreases the price when selling the asset. The price of the risk-free asset is given by

\[ dA_t = r(t, S_t) A_t dt, \quad t \in [0, T], \]

where \( r > 0 \) denotes the interest rate. The price of the risky asset is generated by the following stochastic differential equation:

\[ dS_t = \mu(t, S_t) dt + \sigma(t, S_t) (dW_t + a dM_t) + \lambda(t, S_t) d\theta_t, \quad t \in [0, T], \quad S_0 = x > 0, \]

where \( \mu \) and \( \sigma \) represent the expected return and volatility, respectively, the term \( a \) is a real constant and \( \lambda(S, t) \) denotes the price impact factor created by the trader via selling or buying the underlying asset. \( \theta_t \) is the number of shares that the trader has in the stock at time \( t \). Hence, \( \lambda(S, t) d\theta_t \) captures the price impact of trading. Before dealing with the pricing of a European option in a jump-diffusion illiquid market, we need to observe the following remark.

**Remark 1.** The parameter \( a \) in (4) determines the direction of the jumps\(^1\). In fact the following can be stated:

- If \( a < 0 \) the jumps are pushing the stock price down, i.e. the stock price is decreasing at each jump.
- If \( a = 0 \) then there are no jumps and therefore model (4) is reduced to the model in Liu and Yong\(^7\).
- If \( a > 0 \) the jumps are pushing the stock, i.e. the stock price is increasing at each jump.

### 3 Pricing of a European option in jump-diffusion illiquid market

Let \((V_t)_{t \in [0, T]}\) be the wealth process for the trader. Let also \((\psi_t)_{t \in [0, T]}\) denote the number of shares invested in the risk-free asset. Then, the value of the portfolio is given by

\[ V_t = \psi_t A_t + \theta_t S_t, \quad t \in [0, T]. \]

Assume that the number of shares of the risky asset satisfies the following condition:

\[ d\theta_t = \eta_t dt + \zeta_t (dW_t + b dM_t), \quad t \in [0, T]. \]

Let us consider a European call option with the payoff defined as \( h(S_T) := (S_T - K)^+ \). In order to replicate the option for a perfect hedge, we search for a strategy \((\psi_t, \theta_t)_{t \in [0, T]}\) which, at the expiration date of the option, leads to having a value of the underlying wealth to be equal to the payoff, that is \( V_T = h(S_T) \). Then we can state the following corollary.

\(^1\) it affects also the jumps size.
Lemma 1. Let \( \text{we need Itô formula which is given by the following lemma (see Protter}[9]). \)

The above system is called FBSDE (forward-backward stochastic differential equations). In order to derive the P.D.E. for the European option price, we need to solve the following system of stochastic differential equations.

Corollary 1. The wealth process for the trader of the jump-diffusion model in section 2 satisfies the following stochastic differential equation:

\[
dV_t = \{r(t, S_t)V_t + [\mu(t, S_t) - r(t, S_t)]\theta_t]dt + \theta_tS_t[\lambda(t, S_t)\zeta_t + \sigma(t, S_t)]dW_t + \theta_tS_t[a\sigma(t, S_t) + b\lambda(t, S_t)\zeta_t]dM_t \tag{7}
\]

Proof. By using equations (3), (4), (5) and (6) we have the following:

\[
dV_t = \psi_t dA_t + \theta_t dS_t
= \frac{V_t - \theta_t S_t}{A_t} dA_t + \theta_t S_t \left[ \mu(t, S_t) dt + \sigma(t, S_t)(dW_t + adM_t) + \lambda(t, S_t) d\theta_t \right]
= \{r(t, S_t)V_t + [\mu(t, S_t) - r(t, S_t)]\theta_t]dt + \theta_tS_t \left[ \sigma(t, S_t)(dW_t + adM_t) + \lambda(t, S_t) [\eta dt + \zeta_t (dW_t + bdM_t)] \right]
= \{r(t, S_t)V_t + [\mu(t, S_t) - r(t, S_t)] + \lambda(t, S_t) \eta] \theta_t S_t] dt
+ \theta_t S_t \left[ \lambda(t, S_t) \zeta_t + \sigma(t, S_t) \right] dW_t + \theta_t S_t \left[ a\sigma(t, S_t) + b\lambda(t, S_t) \zeta_t \right] dM_t,
\]

which ends the proof.

Our aim in this paper is to price the European option with payoff \( h(S_T) \) where \( S_T \) is given by (4). We replicate the European option by searching a wealth \((V_t)_{t \in [0, T]} \) which leads to the terminal value \( V_T = h(S_T) \). Thus, as in Liu and Yong[7], we need to solve the following system of stochastic differential equations.

\[
d\theta_t = \eta_t dt + \zeta_t (dW_t + bdM_t), \\
dS_t = \left[ \mu(t, S_t) + \lambda(t, S_t) \eta_t \right] dt + \left[ \sigma(t, S_t) + \lambda(t, S_t) \zeta_t \right] dW_t
+ \left[ a\sigma(t, S_t) + b\lambda(t, S_t) \zeta_t \right] dM_t,
\]

\[
dV_t = \{r(t, S_t)V_t + [\mu(t, S_t) - r(t, S_t)] + \lambda(t, S_t) \eta] \theta_t S_t] dt
+ \theta_t S_t \left[ \lambda(t, S_t) \zeta_t + \sigma(t, S_t) \right] dW_t + \theta_t S_t \left[ a\sigma(t, S_t) + b\lambda(t, S_t) \zeta_t \right] dM_t,
\]

\[
\theta_0 > 0, \quad S_0 > 0, \quad V_T = h(S_T), \tag{8}
\]

The above system is called FBSDE (forward-backward stochastic differential equations) system. In order to derive the P.D.E. for the European option price, we need Itô formula which is given by the following lemma (see Protter[9]).

Lemma 1. Let \( g, l, \) and \( k \) be three adapted processes such that

\[
\int_0^t |g_s| ds < \infty, \quad \int_0^t |\theta_s|^2 ds < \infty, \quad \text{and} \quad \int_0^t |t_s| ds < \infty.
\]

Let \( X = (X_t)_{t \in [0, T]} \) be the process defined by

\[
dX_t = g_t dt + l_t dW_t + k_t dM_t.
\]

For any function \( G \in C^{1,2}([0, T] \times (\infty, \infty], \) we have

\[
G(t, X_t) = G(0, X_0) + \int_0^t (\partial_s G(s, X_s) + (g_s - k_s \rho) \partial_x G(s, X_s)) ds,
\]

\[
\text{subject to } G(T, X_T) = h(X_T).
\]

\[
332
\]

International Conference, Mataró (Barcelona), Spain 25 - 28 June 2013
Proposition 1. Let in the jump-diffusion illiquid market presented in section 2. The next proposition provides the P.D.E. for the price of the European option for the underlying option price is given by

\[
\frac{\partial G(t, X_t)}{\partial t} + \frac{1}{2}\sigma^2 \frac{\partial^2 G(t, X_t)}{\partial x^2} + \frac{1}{2}\lambda \frac{\partial G(t, X_t)}{\partial x} - a\theta G(t, X_t) = \rho G(t, X_t) + \theta \frac{\partial G(t, X_t)}{\partial x}.
\]

(9)

Equation (9) can be written in the following format:

\[
G(t, X_t) = G(0, X_0) + \int_0^t \left[ \partial_r G(s, X_s) + (\mu_s - \kappa_s \rho) \partial_x G(s, X_s) + \right.
\]

\[
\left. \frac{1}{2}\sigma^2 \partial^2 G(s, X_s) + \rho G(s, X_s + k_s) - G(s, X_s) \right) ds
\]

\[
+ \int_0^t \left[ G(s, X_s + k_s) - G(s, X_s) \right] dM_s
\]

\[
+ \int_0^t \lambda \partial_x G(s, X_s) dW_s.
\]

(10)

The next proposition provides the P.D.E. for the price of the European option in the jump-diffusion illiquid market presented in section 2.

Proposition 1. Let \( f(t, S_t) \) denote the price of the European option at time \( t \in [0, T] \) for the model presented in section 2. Then the corresponding P.D.E. for the underlying option price is given by

\[
r(t, S_t) V_t + [\mu(t, S_t) - r(t, S_t) + \lambda(t, S_t) \eta_t] \theta_t = S_t f(t, S_t)
\]

\[
+ \frac{1}{2} \sigma^2 (t, S_t) \partial_s^2 f(t, S_t) + \rho S(t, S_t) (\partial_t f(t, S_t) - (1 + a \sigma(t, S_t) + b \lambda(t, S_t) \eta_t)) - f(t, S_t - )
\]

\[
+ \lambda \partial_x f(t, S_t) - f(t, S_t - )
\]

with the terminal condition \( f(T, S_T) = h(S_T) \). Moreover, the market is incomplete and there is no strategy leading to the terminal wealth \( V_T = h(S_T) := f(T, S_T) \). However, the number of shares \( \theta \) that minimizes the variance is given by

\[
\theta_t = \frac{(\sigma + \lambda \eta_t)^2 S_T \partial_s f + \rho S(\partial_t f - (1 + a \sigma + b \lambda \eta_t)) - f)}{(\sigma + \lambda \eta_t)^2 S_T^2 + \rho S(\partial_t f - (1 + a \sigma + b \lambda \eta_t))^2}.
\]

Proof. Let \( (\theta, S, V) \) be an adapted solution of the FBSDE (8) and assume that there exists a smooth function \( f \in C^{3,1}(\mathbb{R}_+ \times \mathbb{R}_+, \mathbb{R}) \) such that \( f(t, S_t) \) represents the price of the European option at time \( t \in [0, T] \). Since the price of the option at maturity is equal to the payoff, then \( f(T, S_T) = h(S_T) \). Now, using Itô formula (10) we obtain

\[
df(t, S_t) = \{ \partial_t f(t, S_t) + \mu(t, S_t) \theta_t - (\sigma(t, S_t) \partial_s f(t, S_t) + \partial_s f(t, S_t) + \rho S(t, S_t) \partial_t f(t, S_t)) \}
\]

\[
\{ (\sigma(t, S_t) + \lambda(t, S_t) \eta_t) S_t \partial_s f(t, S_t) + \theta_t \} dt
data M_t.
\]

(11)
By comparing equations (7) and (11) one can deduce that it is impossible to find a strategy \((\eta_t, \zeta_t)_{t \in [0,T]}\) that results in the terminal wealth \(V_T = h(S_T) := f(T, S_T)\). Thus, we put the term belonging to \(dt\) equations (7) and (11) equal to each other, which gives the P.D.E. of the option price and then we minimize the distance between the wealth \(V_T\) and the price \(f(T, S_T) = h(S_T)\) over the number of shares of the underlying asset, i.e. \(\theta_t\). The P.D.E. of the option price in this case is

\[
\begin{align*}
    &r(t, S_t) V_t + [\mu(t, S_t) - r(t, S_t) + \lambda(t, S_t) \eta_t] \theta_t S_t = \\
    &\quad \partial_t f(t, S_t) + (\mu(t, S_t) + \lambda(t, S_t) \eta_t - \rho[a \sigma(t, S_t) + b \lambda(t, S_t) \zeta_t]) S_t \partial_S f(t, S_t) \\
    &\quad + \frac{1}{2} \sigma(t, S_t) \zeta_t^2 S_t^2 \partial^2_S f(t, S_t) + \rho (f (t, S_t) - 1 + a \sigma(t, S_t) \\
    &\quad + b \lambda(t, S_t) \zeta_t) - f(t, S_t-))
\end{align*}
\]

with the terminal condition

\[
f(T, S_T) = h(S_T).
\]

To find the number of shares \(\theta_t\) invested in \(S_t\) we need to solve the following problem:

\[
\text{Minimize}_\theta E[H^2(\theta)],
\]

where \(H(\theta) := (h(S_T) - V_T)\). By using (7), (11) and (12) we have

\[
E[H^2(\theta)] = E \left[ \left( \int_0^T (\sigma(t, S_t) + \lambda(t, S_t) \zeta_t) S_t (\partial_S f(t, S_t) - \theta_t) \right)^2 dt \right] \\
+ E \left[ \left( \int_0^T (f(t, S_t) - 1 + a \sigma(t, S_t) + b \lambda(t, S_t) \zeta_t) - f(t, S_t-) \right) S_t \theta_t \right] \\
- \theta_t S_t [a \sigma(t, S_t) + b \lambda(t, S_t) \zeta_t] dM_t \right)^2]
\]

\[
E \left[ \int_0^T (\sigma(t, S_t) + \lambda(t, S_t) \zeta_t) S_t (\partial_S f(t, S_t) - \theta_t) \right]^2 dt \right] \\
+ E \left[ \int_0^T \rho (f(t, S_t) - 1 + a \sigma(t, S_t) + b \lambda(t, S_t) \zeta_t) - f(t, S_t-) \right] \\
- \theta_t S_t [a \sigma(t, S_t) + b \lambda(t, S_t) \zeta_t] \right)^2 dt \right]
\]

where

\[
l(x) = (\sigma + \lambda \zeta)^2 S^2 (\partial_S f - x)^2 + \rho (f(t, S_t) - 1 + a \sigma + b \lambda \zeta) - f - x S (a \sigma + b \lambda \zeta)^2.
\]

The minimum is obtained at \(l'(x) = 0\), which yield the following result:

\[
2(\sigma + \lambda \zeta)^2 S^2 (\partial_S f - x) - 2S (a \sigma + b \lambda \zeta) \rho (f(t, S_t) - 1 + a \sigma + b \lambda \zeta) - f \\
- x S (a \sigma + b \lambda \zeta) = 0,
\]
and
\[
\theta_t = \frac{(\sigma + \lambda \zeta)^2 S^2 \partial_S f + \rho S (a\sigma + b\lambda \zeta) \left( f(t, S_t - (1 + a\sigma + b\lambda \zeta)) - f \right)}{(\sigma + \lambda \zeta)^2 S^2 + \rho S^2 (a\sigma + b\lambda \zeta)^2},
\]

which ends the proof.

It is worth mentioning that in the case where there are no jumps, i.e. when \(a = b = 0\), then \(\theta = \partial_S f\) and the P.D.E. in the previous proposition is reduced to the P.D.E. that is obtained in Liu and Yong\(^7\), assuming there are no dividends.

4 Conclusion

Option pricing is an integral part of modern risk management in increasingly globalized financial markets. The classical Black and Scholes model is regularly used for this purpose. However, one of the main pillars that makes this model operational is the underlying assumption that the markets are perfectly liquid. This assumption is, nonetheless, not fulfilled in reality since perfectly liquid markets do not exist. In our opinion the question should not be whether the markets are illiquid or not, the question should be about the degree of illiquidity. Thus, taking into account the fact that markets are illiquid can improve on the precision of the underlying option pricing.

This paper is the first attempt, to our best knowledge, that extends the existing literature on option pricing by introducing a jump-diffusion model for illiquid markets. This seems to be a more realistic approach to deal with a market that is incomplete. A solution for the option pricing within this context is provided along with the underlying proof. The suggested solution might be useful to investors in order to determine the optimal value of an option in a market that is characterized by illiquidity.

References

1. L. Chan and J. Lakonishok. The behavior of stock prices around institutional trades. *Journal of Finance* 50:1147–1174, 1995.
2. B. Fischer and M. Scholes. The Pricing of Options and Corporate Liabilities. *Journal of Political Economy* 81 (3):637-654, 1973.
3. M. Dritschel and Ph. Protter. Complete markets with discontinuous security prices. *Finance & Stochastics* 3(2):203-214, 1999.
4. Y. El-Khatib and N. Privault. Hedging in complete market driven by normal martingales. *Applicatiores Mathematicae* 30(2):147–172, 2003.
5. Y. El-Khatib and Q.M. Al-Mdallal. Numerical simulations for the pricing of options in jump diffusion markets, *Arab Journal of Mathematical Sciences*, 18(2): 199–208, 2012.
6. D. Keim and A. Madhavan. The upstairs market for large-block transactions: analysis and measurement of price effects. *Review of Financial Studies* 9: 1–36, 1996.
7. H. Liu and J. Yong. Option pricing with an illiquid underlying asset market, *Journal of Economic Dynamics and Control* 29: 2125–2156, 2005.
8. R. C. Merton. Option pricing when underlying stock returns are discontinuous, *Journal of Financial Economics*, 3:125–144, 1976.

9. Ph. Protter. *Stochastic integration and differential equations. A new approach*. Springer-Verlag, Berlin, 1990.

10. W. F. Sharpe, G. J. Alexander and J. V. Bailey. *Investments*. Prentice Hall, New Jersey, 1999.
Variations of PageRank with application to linguistic data

Christopher Engström¹, Sergei Silvestrov², and Thierry Hamon³

¹ Division of Applied mathematics, School of Education, culture and communication, Mälardalen University, Box 883, 72123 Västerås, Sweden (E-mail: Christopher.engstrom@mdh.se)
² Division of Applied mathematics, School of Education, culture and communication, Mälardalen University, Box 883, 72123 Västerås, Sweden (E-mail: Sergei.silvestrov@mdh.se)
³ LIM&BIO (EA3969), Université Paris 13, Sorbonne Paris Cité, 74, rue Marcel Cachin, 93017 Bobigny Cedex France (E-mail: thierry.hamon@univ-paris13.fr)

Abstract. In this paper we will give some alternative formulations of PageRank originally used by Google to rank home pages in a search query. By looking at PageRank from a probabilistic perspective using Markov chains we will see how to fix some of the limitations in the original definition of PageRank to make it usable for comparison not only within a system, but also between systems.

Last we will take a short look at an application of PageRank to linguistic data for identification of "important" terms in a corpus of biomedical texts. We show how the method can be applied for this kind of data as well as highlighting some of the limitations of the method and future work.

AMS Mathematical Subject Classification 2010: 68T50, 60J20, 68U35,15A51, 05C50, 05C90, 68M11

Keywords: PageRank, Markov chains, text mining.

1 Background

PageRank can be seen as the stationary distribution of a Markov chain described by a random walk on a directed graph with the following rules: In every step with probability $c$ we either follow any edge out of the current vertex with equal probability $1/n$ where $n$ is the number of edges going out of the current vertex. With probability $(1-c)$ we instead pick a new vertex at random with no consideration of current vertex [5]. We do not allow any vertex to link to itself and special care needs to be taken of vertices with no outgoing edges (dangling nodes), usually by letting them link to all vertices instead. PageRank in it’s original form can be seen below [1].

Definition 1. PageRank $R^{(1)}$ for the vertices $V$ in a directed graph $G$ is defined as the right eigenvector with eigenvalue one to the matrix

$$M = c(A + gu^T) + (1-c)ue^T$$

where $0 < c < 1$, $A$ is the adjacency matrix of $G$ with non-zero elements $a_{ij} = 1/n_i$ where $n_i$ is the number of non-zero elements on row $i$, $u$ is a non-negative column-vector of length $|V|$ with $\|u\|_1 = 1$, $e$ is a column-vector of
length $|V|$ with elements $e_i = 1$, $g$ is a column-vector of length $|V|$ with elements $g_i = 1$ if $v_i$ is a dangling node, and $g_i = 0$ otherwise. $R^{(1)}$ is normalized such that $||R^{(1)}||_1 = 1$ where $||x||_1 = \sum_i |x_i|$ is the $L^1$-norm.

If we consider the random walk on a graph then $u$ is a probability distribution of where to go whenever we start over in a new random vertex either with probability $(1 - c)$ in every step or whenever we end up in a dangling node. $u$ often have elements equal to $1/|V|$ but changes can be made to decrease or increase the influence of certain vertices. It’s easy to see that $M$ is a column stochastic irreducible matrix thus we have a guarantee that $R^{(1)}$ exist and is a positive vector from Perron-Frobenius for non-negative irreducible matrices [3].

While $M$ generally is not sparse, $A$ often is a very large and sparse matrix and as such the Power method is often used to calculate PageRank by iterating

$$R^{(1)}_{n+1} = MR^{(1)}_n = cA^T R^{(1)}_n + d_n u$$

$$d_n = 1 - \sum cA^T R^{(1)}_n$$

where we use that the part not depending on $A$ in the multiplication is proportional to $u$, $(c(gu^T)^T + (1 - c)ue^T) R^{(1)} \propto u$ and $||R^{(1)}||_1 = 1$.

Since $R^{(1)}$ is normalized to some constant (even if that constant is the number of vertices) we cannot reliably compare PageRank between different graphs, this since dangling nodes are assumed to link to all vertices. If the graphs are big with a similar structure this is unlikely to cause a problem, however for small graph’s it can make a large difference.

We would also like to allow vertices to have a weightsum of links out less than one, for example to differentiate between a clique and one with no edges.

## 2 PageRank from a probability perspective

We will look at PageRank as a Random walk on a graph and define a slightly different variation of PageRank.

**Definition 2.** Consider a random walk on a graph described by $cA$. We walk to a new vertex from our current with probability $0 < c < 1$ and stop with probability $1 - c$. Using a constant weight vector $u$ with all elements equal we define PageRank $R^{(2)}$ for a single vertex as

$$R^{(2)}_j = \left( \sum_{e_i \in S, e_i \neq e_j} P(e_i \to e_j) + 1 \right) \left( \sum_{k=0}^{\infty} (P(e_j \to e_j))^k \right)$$

(2)

where $P(e_i \to e_j)$ is the probability to hit node $e_j$ starting in node $e_i$. This can be seen as the expected number of visits to $e_j$ if we do multiple random walks, starting in every vertex once.
Note that we use $A$ and not $M$ which is modified for dangling nodes nor do we normalize PageRank to some constant. This means that we can now compare PageRank between different graphs without problem, neither do we need to consider dangling nodes as if linking to all vertices. The weight sum of edges out of a vertex can now be anything between zero and one as well.

**Theorem 1.** PageRank $R^{(2)}$ is proportional to $R^{(1)}$ ($R^{(2)} \propto R^{(1)}$).

**Proof.** $(cA^T)^k_{ij}$ is the probability to be in node $e_i$ starting in node $e_j$ after $k$ steps. Multiplying with the one vector $e$ therefor gives the sum of all the probabilities to be in node $e_i$ after $k$ steps starting in every node once (constant $u$). The expected total number of visits is the sum of all probabilities to be in node $e_i$ for every step starting in every node:

$$R^{(2)}_j = \left( \sum_{k=0}^{\infty} (cA^T)_{ij}^k \right) e_j$$  \hspace{1cm} (3)

\[
\sum_{k=0}^{\infty} (cA^T)^k
\]

is the Neumann series of $(1 - cA^T)^{-1}$ which is guaranteed to converge since $cA^T$ is non-negative and have column sum $< 1$. This gives:

$$R^{(2)} = \left( \sum_{k=0}^{\infty} (cA^T)^k \right) e = (1 - cA^T)^{-1}e = (1 - cA^T)^{-1}nu$$  \hspace{1cm} (4)

where we use that $u$ is constant and $n$ is the number of vertices. Then

$$R^{(1)} = MR^{(1)} \iff (cA^T - I)R^{(1)} = -(cu^T + (1 - c)ue^T)R^{(1)}$$  \hspace{1cm} (5)

Since every column of $u_1g^T$ is either equal to $u$ or zero and all columns equal to $u$ for $ue^T$ we can see that $-(cu^T + (1 - c)ue^T)R^{(1)}$ will be proportional to $u$. This can be written as:

$$(1 - cA^T)R^{(1)} = ku \iff R^{(1)} = (1 - cA^T)^{-1}ku \propto (1 - cA^T)^{-1}nu = R^{(2)}$$  \hspace{1cm} (6)

We note that although we made the assumption $u$ constant, it is quite easy to show a similar relation for other $u$ as well by defining $R^{(2)}$ slightly differently.

### 2.1 Relation between normalized and non normalized PageRank

Since we know that they are proportional it’s easy to get from the non-normalized PageRank $R^{(2)}$ to the normalized PageRank $R^{(1)}$ by simply dividing by $||R^{(2)}||_1$:

$$R^{(1)} = \frac{R^{(2)}}{||R^{(2)}||_1}$$  \hspace{1cm} (7)

But we do not yet have any way to go from $R^{(1)}$ to $R^{(2)}$. For this we formulate the following:
Theorem 2. The relation between normalized PageRank $R^{(1)}$ and non-normalized PageRank $R^{(2)}$ can be written:

$$R^{(2)} = \frac{nR^{(1)}}{d}$$

where $n$ is the number of nodes and $d = 1 - \sum cA^\top R^{(1)}$ is the value gained from the last step in the power method.

Proof. If we let $S_a$ be the dangling nodes in our system we can rewrite

$$d = 1 - \sum cA^\top R^{(1)} = (1 - c) + \sum_{a \in S_a} R_a^{(2)}$$

If there’s no dangling nodes then

$$\sum R^{(2)} = n \sum_{k=0}^\infty c^k = \frac{n}{1 - c}$$

The loss from a dangling node $a$ can be seen as the PageRank $R_a^{(2)}$ (expected visits to $a$) times the sum of probability going from there if there where no dangling nodes present, this gives

$$\sum R^{(2)} = \frac{n - \sum_{a \in S_a} R_a^{(2)} c}{1 - c}$$

where $S_a$ contains all the dangling nodes in our system. By (7) (9) and (11)

$$R^{(1)} = \frac{R^{(2)} (1 - c)}{n - \sum_{a \in S_a} R_a^{(2)} c}$$

$$d = (1 - c) + \frac{\sum_{a \in S_a} R_a^{(2)} c(1 - c)}{n - \sum_{a \in S_a} R_a^{(2)} c} \Leftrightarrow \sum_{a \in S_a} R_a^{(2)} c = n - \frac{(1 - c)n}{d}$$

$$R^{(2)} = \frac{(n - \sum_{a \in S_a} R_a^{(2)} c) R^{(1)}}{1 - c} = \frac{n(1 - c)R^{(1)}}{d(1 - c)} = \frac{nR^{(1)}}{d}$$

We now have the possibility of comparing the PageRank of different disjoint system immediately by storing the results in $P^{(2)}$ instead. It’s however important to note one restriction: while we can weight vertices differently within a system, we cannot weight systems differently since the weight vector of every subsystem is normalized (and later weighted depending on number of nodes $n$ and its $d$—value). This means that as the weight for one vertex in a subsystem is lowered, other vertices in the same subsystem effectively gets a higher weight compared to those in other systems.

We look at a large system $S$ composed of multiple systems $S_1, S_2, \ldots, S_N$ where we want to use a global weight vector $V$ rather than one for each subsystem. If $v_i$ is the part of $V$ corresponding to nodes in system $S_i$, then as
we normalize $v_i$ and calculate $P^{(2)}$ we would get the correct internal weighting between nodes in $S_i$. In order to get the correct weighting compared to other subsystems we therefore only need to find the correct weight for the subsystem itself. Summation over $v_i$ gives the relative weight of subsystem $S_i$. This gives a slightly new definition:

**Definition 3.** $R_{S_i}^{(3)}$ for system $S_i \in S$ is defined as:

$$R_{S_i}^{(3)} = \frac{R_{S_i}^{(1)} ||v_i||_1}{d_i}$$

where $v_i$ is the part of the global weight vector $V$ belonging to the nodes in system $S_i$. [9]

We note that $||V||_1 \neq 1$ but can now be any non-negative vector. If we also choose $V$ such that we define a "default" weight rather than a total sum of weights, we see that as we change one subsystem the rank of other subsystems never change. For example we could use the default weight 1 and change individual weight from there. If $V$ is the one-vector we easily find that $R^{(2)} = R^{(3)}$. Now that we have this version of PageRank we can compute and make changes to individual subsystems $S_i$ without the need to re-compute PageRank of any other subsystems, while still being able to compare PageRank between subsystems making the method very suitable for continuous updates and parallel computations.

### 3 Using PageRank to find annotated terms in a text corpus

Regarding the low coverage of the existing terminologies to identify terms in corpus [4,7,17], approaches have been proposed to automatically extract terms from texts, i.e. noun phrases referring to linguistic entities in a specialized domain (e.g. medicine, biology, law, electric power plant, etc.) for text corpora (see [6]). Usually, proposed approaches are based on linguistic rules or and collocations. However, even if the extracted terms are useful for terminology building, document indexing or text mining, the quality of the results, especially the precision, is not sufficient. The various strategies of term extraction have difficulties to identify relevant terms among the huge amount of term candidates: terms and irrelevant extracted noun phrases could have similar linguistic characteristics. To tackle this problem, several statistical measures based on term or word frequency, but also lexical or contextual clues have been proposed [8,15] without proposing convicting methods for filtering or ranking the extracted terms: a single measures is not sufficient to rank the terms correctly and we assume that relations shared by terms have to take into account to achieve this aim. In that respect, we focus here on the term ranking by exploiting graph structure and PageRank. PageRank also gives lots of customization options in personalization vector $V$, constant $c$, and how we choose to weight the edges in the graph making it suitable for a machine learning approach.
3.1 Description of data

Our working data is based on the Genia corpus\(^1\) [14]. This text corpus consists of 1999 abstracts issued from Medline portal\(^2\). The abstracts concern the transcription factors in human blood cells (abstracts indexed by the MeSH terms: *human, blood cell and transcription factor*). The corpus contains 18,545 sentences and 436,967 words. Each abstract is also annotated with terms referring to physical biological entities (organisms, proteins, cells, genes) as well as biologically meaningful terms (e.g. molecular functions). Thus, 96,582 terms are annotated in the abstracts (see Figure 1).

To address potential mechanisms by which [estr]ogens suppress eryth[ro]poiesis, we have examined their effects on GATA-1, an erythroid transcription factor that participates in the regulation of the majority of erythroid cell-specific genes and is necessary for full maturation of erythrocytes.

**Fig. 1.** Excerpt of the Genia corpus: sentence annotated with terms.

We consider this term annotation as our Gold Standard: the list of terms extracted and ranked by our term extraction approach is compared to this reference. In that respect, for the experiment, we randomly split the corpus in two parts: the training set includes 60% of the corpus (1,200 abstracts) while the remaining part of the corpus (799 abstracts) will be the test set. A preprocessing of the text corpora is performed with the Ogmios platform [11]: texts are segmented in words and sentences, and part-of-speech categories and lemma are associated to the words thanks to the TreeTagger [18].

In our experiment, the list of terms have been extracted with linguistic rules [2]. Parsing patterns taking into account the morpho-syntactic variation, are recursively applied to the text and provide noun phrases which seems to be relevant in the targeted domain (see Figure 2).

Additional information is also associated to the extracted terms:

- Syntactic dependencies between term components are computed according to assigned Part-of-Speech tags and shallow parsing rules. Each term is considered as a syntactic binary tree composed of two elements describing the syntactic role of the component in the term: the head component is the main word of the term, and the expansion component modifies the main word. Each component is also considered as a multi-word or a single-word term. For instance, *full maturation* is the head component of the term analyzed in Figure 3. As several terms can share components, the list of terms can be view as a syntactic network or a graph.

---

\(^{1}\) version 3.02, available at [http://www.nactem.ac.uk/genia/genia-corpus/event-corpus](http://www.nactem.ac.uk/genia/genia-corpus/event-corpus)

\(^{2}\) [http://www.ncbi.nlm.nih.gov/pubmed](http://www.ncbi.nlm.nih.gov/pubmed)
• Statistical measures are provided with each term: term frequency, number of documents where appears a term, \textit{C-Value} \cite{10,16}, \textit{TF-IDF}, measures based on syntactic and contextual clues \cite{8}, etc.

To address potential mechanisms by which \textit{estrogens} suppress \textit{erythropoiesis}, we have examined their effects on \textit{GATA-1}, an erythroid transcription factor that participates in the regulation of the majority of erythroid cell-specific genes and is necessary for \textit{full maturation of erythrocytes}.

\textbf{Fig. 2.} Excerpt of the Genia corpus: terms automatically extracted.

\textbf{Fig. 3.} Parsing tree of the term \textit{full maturation of erythrocytes}

\section*{3.2 Method}

The first thing to consider is how to create a sub-stochastic matrix given a tree such as in Figure 3 for use in our method. We have two different relations "head" and "modifier" which we could choose to weight differently, since they are also directed "A is a "head" of B" we could choose to give different weights depending on direction as well. This gives four different types of edges (to head, from head, to modifier, from modifier), we give every type of edge a non-negative weight which is then used to create our system matrix \(A\) by normalizing every row such that they have sum one.

The next point is choosing constant \(0 < c < 1;\) for a very low \(c\) PageRank is mainly decided by the personalization vector \(V\) and the immediate neighbors of a vertex in the graph. For a large \(c\) the personalization vector \(V\) have little effect on the ranking; instead some specific graph structures give a greatly inflated rank (such as two vertices linking only to each other). As \(c\) get close to 1 there are also problems with stability and convergence \cite{13}\cite{12}. As with the edge weights it’s hard to say what \(c\) constitutes an optimal value, initially we choose to start with \(c = 0.85\) (as used by Google initially).

Last we need to choose \(V\), here a lot of variation can be done such as weighting the terms depending on frequency or similar measurements or by weighting depending on type of vertex. Since we have a tree-structure for our graphs it might be worth it to weight depending on level in the tree or if it can
be divided in "head" and "modifier" or not. Initially we tried using $V$ with a uniform weight and weights equal to the frequency of corresponding terms.

### 3.3 Initial results

We compare the results to using only the frequency measurement with PageRank using a couple of different parameter values.

![Fig. 4. Precision using frequency or constant $V = e$ with different values of $c$ and all edge weights $W$ equal.](image1)

![Fig. 5. Precision using frequency $V = \text{freq.}$ with $c=0.85$ and different edge weights $W=\{th, tm, fh, fm\}$.](image2)

Where $th$ and $tm$ is the weight of edges pointing at a head or modifier respectively and $fh$ and $fm$ are the edges pointing from a head or modifier respectively. From Figure 4 we can see that we have similar results when using PageRank as when using frequency. Using frequency seem to have a slightly better precision initially, but then drops off faster than the PageRank variations. From changing $c$ we cannot see any major changes, lowering it does seem to even out the curve slightly. We see that both frequency and PageRank to have a very high precision for the first couple of terms but then rapidly fall. Using the frequency by itself or as weights $V$ seem to give the best results.

In Figure 5 we can see that the result differ significantly depending on how we choose to create our system matrix. We see that we get a better result if we have a higher chance to go down the tree ($th, tm$ high) rather than the opposite, and that the head ($th, fh$ high) seems to be more important than the modifier.

### 4 Conclusions and future work

We have seen from our initial studies on our training data that we can expect to get a slight improvement compared to using only the frequency. Changes in $V$ have some effect and can likely be tuned further by looking at other measures or weighting depending on type of vertex. The main thing influencing the result however seem to be how the system matrix is created. Not only the weights $\{th, tm, fh, fm\}$, it’s also possible let $c$ depend on vertex properties such as its depth in the tree. While both frequency and PageRank seem to give similar
results, it’s not clear if they find the same annotated terms or not. While we can expect to find some improvement using PageRank, how large and if the same parameter values can then be used on another text corpus remains to be seen. The next step is in using a machine learning approach to try and optimize parameters for the model and try it on our test data, if this is successful we can then see if it proves useful for another text corpus. One limitation in creating the system matrix is the normalization of rows, while we could potentially have a row sum less than one, we cannot have one larger than one and still guarantee convergence. This is something we would like to look at in the future.

Acknowledgements

This research was supported in part by the Swedish Research Council (621-2007-6338), Swedish Foundation for International Cooperation in Research and Higher Education (STINT), Swedish Royal Academy of Sciences, Crafoord Foundation, the EU exchange programme Erasmus, Medical informatics and Bioinformatics laboratory LIM&BIO at the Bobigny School of Medicine (SMBH) of Université Paris 13 and The School of Education, Culture and Communication at Mälardalen University.

References

1. F. Andersson and S. Silvestrov. The mathematics of internet search engines. *Acta Appl. Math.*, 104(2):211–242, 2008.
2. S. Aubin and T. Hamon. Improving term extraction with terminological resources. In T. Salakoski, F. Ginter, S. Pyysalo, and T. Pahikkala, editors, *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, number 4139 in LNAI, pages 380–387. Springer, August 2006.
3. A. Berman and R. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Number del 11 in Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1994.
4. O. Bodenreider, T. C. Rindflesch, and A. Burgun. Unsupervised, corpus-based method for extending a biomedical terminology. In *Workshop on Natural Language Processing in the Biomedical Domain (ACL2002)*, pages 53–60, 2002.
5. K. Bryan and T. Leise. The $25,000,000,000$ eigenvector: The linear algebra behind google. *SIAM Review*, 48(3):569–581, 2006.
6. M. T. Cabr, R. Estop, and J. Vivaldi. Automatic term detection: a review of current systems. In *Recent Advances in Computational Terminology*, John Benjamins, Amsterdam, Philadelphia, 2001.
7. C. G. Chute, S. P. Cohn, K. E. Campbell, D. E. Olivier, and J. R. Campbell. The content coverage of clinical classifications. *Journal of American Medical Informatics Association*, 3:224–233, 1996.
8. P. Drouin. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–117, 2003.
9. C. Engström. Pagerank as a solution to a linear system, pagerank in changing systems and non-normalized versions of pagerank. Master’s thesis, Mathematics, Centre for Mathematical sciences, Lund Institute of Technology, Lund University, May 2011:E31. LUTFMA-3220-2011.
10. K. T. Frantzi, S. Ananiadou, and J. Tsujii. Automatic term recognition using contextual clues. In Proceedings of the Second Workshop on Multilinguality in software Industry: The AI Contribution (MULSAIC’97) - Workshop WL1, IJ-CAI’97, Nagoya, Japan, August 1997.

11. T. Hamon, A. Nazarenko, T. Poibeau, S. Aubin, and J. Derivire. A robust linguistic platform for efficient and domain specific web content analysis. In Proceedings of RIAO 2007, Pittsburgh, USA, 2007. 15 pages.

12. T. Haveliwala and S. Kamvar. The second eigenvalue of the google matrix. Technical Report 2003-20, Stanford InfoLab, 2003.

13. S. Kamvar and T. Haveliwala. The condition number of the pagerank problem. Technical Report 2003-36, Stanford InfoLab, June 2003.

14. J.-D. Kim, T. Ohta, Y. Teteisi, and J. Tsujii. Genia corpus - a semantically annotated corpus for bio-textmining. Bioinformatics, 19(suppl. 1):i180–i182, 2003.

15. I. Korkontzelos, I. P. Klapaftis, and S. Manandhar. Reviewing and evaluating automatic term recognition techniques. In B. Nordstrm and A. Ranta, editors, Advances in Natural Language Processing (6th International Conference on NLP, GoTAL 2008), number 5221 in LNAI, pages 248–259. Springer, August 2008.

16. D. Maynard and S. Ananiadou. Identifying terms by their family and friends. In Proceedings of COLING 2000, pages 530–536, Saarbrucken, Germany, 2000.

17. A. T. McCray, A. C. Browne, and O. Bodenreider. The lexical properties of the gene ontology (GO). In Proceedings of the AMIA 2002 Annual Symposium, pages 504–508, 2002.

18. H. Schmid. Probabilistic part-of-speech tagging using decision trees. In D. Jones and H. Somers, editors, New Methods in Language Processing Studies in Computational Linguistics, 1997.
DECREASE IN RISK ERRONEOUS CLASSIFICATION
THE MULTIVARIATE STATISTICAL DATA
Farhadzadeh E.M., Farzaliyev Y.Z., Muradaliyev A.Z.
Azerbaijan Scientific-Research and Design-Prospecting Institute of Energetic
AZ1012, Ave. H.Zardabi-94, e-mail:fem1939@rambler.ru

ABSTRACT
Objective estimation of parameters of individual reliability is an indispensable condition of an opportunity of decrease in operational expenses for maintenance service and repair of the equipment and devices of electro power systems. The method of decrease in risk of erroneous classification of multivariate statistical data offered. The method based on imitating modeling and the theory of check of statistical hypotheses.

I. INSTRUCTION
Estimation parameters of individual reliability of the equipment of power supply systems provides classification of final population of multivariate statistical data of operation, tests and restoration of deterioration on the set versions of attributes (VA) [1].

VA reflects features of a design, a condition of operation, feature of occurrence of refusals and carrying out of repairs of the equipment. Expediency of classification on each of population VA is established by comparison of statistical functions of distribution (s.f.d.) final population of statistical data \( F^n_\Sigma (X) \) and s.f.d. samples \( n \) random variables from this population on \( i \) versions of \( V \) attribute \( F^i_\nu(X) \), where \( \nu=1, k; k \)-number of attributes of random variable \( X \) (for example, durations of emergency repair); \( i=1, rk; r_k \)-number of versions \( k \) an attribute. If s.f.d. \( F^n_\Sigma (X) \) and \( F^i_\nu(X) \) differ not casually, in other words, sample \( \{X\}_n \) where \( n \)-number of random variables of sample, it is not representative classification of data at an estimation of parameters of individual reliability is expedient and on the contrary. It is necessary to note, that unlike sample of a general data population (analogue: infinite set of random variables with uniform distribution in an interval \([0,1]\)), which imposing appearance is set by some significance value \( \alpha \), sample of final population of multivariate data on set VA is not casual, as a matter of fact, and it can appear only representative. In particular, sample can appear representative, if for considered data set VA not significant.

II. RECOMMEND METHOD
In a basis of comparison \( F^n_\Sigma (X) \) and \( F^i_\nu(X) \) there is a statistical modeling (by means of computer program RAND) \( n \) pseudo-random numbers \( \xi \), random variables of sample equal to number, with uniform distribution in an interval \([0,1]\).

Indispensable condition thus is consistency s.f.d. \( F^i_\nu(\xi) \) to the uniform law of distribution \( F^n_\Sigma (\xi) \), in other words, casual character of distinction \( F^n_\xi(\xi) \) and \( F^i_\nu(\xi) \). It is obvious, that from the uniform law of change of random numbers \( \xi \) at all consistency does not follow the uniform law s.f.d. \( F^i_\nu(\xi) \) with the set significance value \( \alpha \). Use at modeling statistical analogue \( F^i_\nu(X) \) s.f.d. \( F^i_\nu(\xi) \), essentially differing from \( F^n_\xi(\xi) \), leads to erroneous increase in value of the greatest divergence of distribution of this analogue \( F^n^*_\nu(X) \) from \( F^i_\nu(X) \) and by that to growth of probability of the erroneous decision at classification of data.
Representative character of sample \( \{ \xi_n \} \) at the decision of a problem of an estimation of expediency of classification of multivariate data it was supervised Kolmogorov's by criterion [2].

According to this criterion sample \( \{ \xi_n \} \) it is unpresentable, if

\[
D_n > d_{n, (1-\alpha)}
\]

where:

\[
D_n = \max(D_n^+, D_n^-)
\]

\[
D_n^+ = \max \{ D_i^+ \}; \quad 1 \leq i \leq n
\]

\[
D_i^+ = \left( \frac{i - \xi_i}{n} \right)
\]

\[
D_n^- = \max \{ D_i^- \}; \quad 1 \leq i \leq n
\]

\[
D_i^- = \left( \frac{\xi_i - i - 1}{n} \right)
\]

\( d_{n, (1-\alpha)} \) – critical value of statistics \( D_n \) provided that \( F_\Sigma(\xi) \) and \( F_\Sigma^*(\xi) \) differ casually

In [3] it is marked, that estimation \( D_n \) under the formula

\[
D_n = \max \left\{ \left| D_i^+ \right| ; \xi_i \right\} \quad 1 \leq i \leq n
\]

leads to incorrect decisions on a parity \( F_\Sigma(\xi) \) and \( F_\Sigma^*(\xi) \).

The similar remark can be found and in [4]. The reason of such discrepancy does not stipulate. At uncertain in advance \( n \), decrease in time of calculation, according to [3], is reached by application of exact approach Stephens, which tabulated critical values \( d_{n, (1-\alpha)} \), depending from \( n \) and \( \alpha \), reduces to dependence only from \( \alpha \). Sample \( \{ \xi_n \} \) it is unpresentable, if

\[
A \cdot D_n > C_{1-\alpha}
\]

where:

\[
A = \left( \sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}} \right)
\]

For example, at \( n=4 \) size \( A=2.175 \) and for \( \alpha=0.1 \) critical value \( C_{1-\alpha}=1.224 \), and at \( \alpha=0.05 \) size \( C_{1-\alpha}=1.358 \).

Application of a method of the decision of «a return problem» when it is in advance known, that sample \( \{ \xi_n \} \) it is unpresentable, has shown, that criteria (1) and (8) for values most often used in practice \( \alpha=0.05 \) and \( \alpha=0.1 \) not casual character of divergence \( F_\Sigma(\xi) \) and \( F_\Sigma^*(\xi) \) at small \( n \) establish only for those cases when it does not raise the doubts. For acknowledgement of this statement, we shall consider a following example. Let random numbers \( \psi \) have uniform distribution \( F_\Sigma(\psi) \) in an interval \([0.5; 1]\). Casual sample is set \( \{ \psi_n \} \) with \( n=4: \{0.86346; 0.50672; 0.91424 \text{ and } 0.67210 \} \). Check up the assumption of imposing appearance of this sample for the uniform law of distribution of a random variable \( \xi \) in an interval \([0,1]\).

Results of calculations are resulted in table 1.

| i | \( F_\Sigma(\psi_i) \) | i/n | \( D_i^+ \) | \( D_i^- \) | The note |
|---|------------------|-----|-----------|-----------|---------|
| 1 | 0.507            | 0.25| -0.257    | +0.506    | \( D_i^+ = 0.086; D_i^- = 0.506 \) |
| 2 | 0.672            | 0.5 | -0.172    | +0.422    | \( D_n=0.506; D_n<d_{4, 0.0}=0.565 \) |
| 3 | 0.863            | 0.75| -0.113    | +0.363    | \( AD_n=1.101; \) |
| 4 | 0.914            | 1.00| +0.086    | +0.164    | \( AD_n<C_{0.9}=1.224 \) |

As sample follows from table 1 \( \{ \psi \}_{4} \) does not contradict the assumption of imposing appearance rather \( F_\Sigma(\xi) \) at \( \alpha=0.1 \).

These features and some assumptions of the reasons of their occurrence [5] have demanded to pass from the analysis of absolute values of the greatest divergence of distributions \( F_\Sigma(\xi) \) and \( F_\Sigma^*(\xi) \), to the analysis of the valid values of the greatest divergence (St). Thus under «the greatest
divergence $F_{\Sigma}(\xi)$ and $F_{\nu}^{*}(\xi)$» we shall understand the greatest on the module vertical distance between $F_{\Sigma}(\xi)$ and $F_{\nu}^{*}(\xi)$ with $i=1, n$.

Calculations $S_{t_n}$ were spent according to the algorithm, integrated which block diagram is resulted in figure 1.

![Fig.1. Block diagram of algorithm of calculation of the greatest divergence of distributions $F_{\Sigma}(\xi)$ and $F_{\nu}^{*}(\xi)$](image)

Application of formulas of type

$$S_{t_n} = \max \left( \xi_i - \frac{i}{n} \right) \quad 1 \leq i \leq n \quad (10)$$

calculation on the computer leads to erroneous results. For example, according to table 1 the maximal value among four realizations of size $D_t^+$ will, $D_t^+ = 0.086$, and the greatest vertical divergence between $F_{\Sigma}(\xi)$ and $F_{\nu}^{*}(\xi)$ it is equal $D_t^+ = -0.256$

Results of ordering of given realizations $S_{t_n}$ presented in table 2 and allow concluding:

1. Quintile distributions $F^{*}(S_{t_n}) = \alpha$ and $n \geq 2$ are equal on size and are opposite on a sign (distinction in a sign is caused by distinction of formulas 4 and 10) quintiles distributions $F(D_{t_n}) = 2\alpha$ (see tabl.16 [2]);
2. Distribution $F^{*}(S_{t_n})$ is asymmetrical. In the illustrative purposes on fig. 2 are resulted s.f.d. $F^{*}(S_{t_n})$ for of some $n$. The assumption of symmetry of distribution $F(S_{t_n})$ it is possible to explain discrepancy of probability practically equal quintile distributions $F^{*}(S_{t_n})$ and $F(D_{t_n})$;
3. Than $\xi_n$ it is less, that negative value on sign $S_{t_n}$ on size will be more, since $S_{t_n} = (\xi_{n-1})$. On experimental data the least value $S_{t_n}$ for $n=2$ has appeared equal $S_{t_n} = 0.992$, and the greatest $S_{t_n} = +0.489$ at sup equal, accordingly, 1 and 0.5.
### Table 2

Some results of an estimation s.f.d. $F^*(S_n)$

| $F(S_n)$ | 0.025 | 0.05 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 0.95 | 0.975 |
|----------|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| n        |       |     |     |     |     |     |     |     |     |     |     |     |     |
| 2        | -0.842 | -0.775 | -0.684 | -0.551 | -0.473 | -0.149 | -0.363 | -0.304 | -0.239 | -0.060 | 0.184 | 0.285 | 0.343 |
| 3        | -0.709 | -0.635 | -0.566 | -0.471 | -0.400 | -0.335 | -0.296 | -0.252 | -0.200 | -0.145 | 0.231 | 0.299 | 0.372 |
| 4        | -0.623 | -0.567 | -0.494 | -0.414 | -0.355 | -0.302 | -0.253 | -0.217 | -0.173 | 0.155 | 0.240 | 0.319 | 0.377 |
| 5        | -0.567 | -0.511 | -0.449 | -0.370 | -0.318 | -0.274 | -0.232 | -0.190 | -0.147 | 0.164 | 0.246 | 0.309 | 0.360 |
| 6        | -0.523 | -0.469 | -0.411 | -0.338 | -0.292 | -0.252 | -0.215 | -0.173 | -0.127 | 0.171 | 0.244 | 0.303 | 0.358 |
| 7        | -0.481 | -0.438 | -0.384 | -0.318 | -0.274 | -0.235 | -0.201 | -0.162 | -0.113 | 0.165 | 0.235 | 0.290 | 0.342 |
| 11       | -0.389 | -0.353 | -0.309 | -0.255 | -0.219 | -0.189 | -0.110 | -0.129 | -0.097 | 0.160 | 0.216 | 0.260 | 0.302 |
| 16       | -0.333 | -0.295 | -0.258 | -0.215 | -0.184 | -0.158 | -0.134 | -0.103 | 0.207 | 0.105 | 0.150 | 0.194 | 0.232 |
| 22       | -0.280 | -0.253 | -0.221 | -0.183 | -0.157 | -0.135 | -0.113 | -0.083 | 0.105 | 0.137 | 0.176 | 0.210 | 0.235 |
| 29       | -0.246 | -0.219 | -0.193 | -0.160 | -0.138 | -0.119 | -0.099 | -0.068 | 0.098 | 0.126 | 0.158 | 0.186 | 0.212 |
| 40       | -0.208 | -0.187 | -0.164 | -0.136 | -0.119 | -0.102 | -0.084 | -0.050 | 0.089 | 0.112 | 0.140 | 0.164 | 0.185 |
| 60       | -0.173 | -0.156 | -0.137 | -0.114 | -0.097 | -0.083 | -0.069 | 0.054 | 0.077 | 0.096 | 0.118 | 0.138 | 0.155 |
| 90       | -0.142 | -0.127 | -0.111 | -0.092 | -0.079 | -0.068 | -0.055 | 0.051 | 0.067 | 0.081 | 0.100 | 0.116 | 0.130 |
| 120      | -0.122 | -0.110 | -0.096 | -0.080 | -0.068 | -0.059 | -0.047 | 0.047 | 0.060 | 0.072 | 0.089 | 0.102 | 0.114 |
| 150      | -0.110 | -0.099 | -0.086 | -0.071 | -0.062 | -0.053 | -0.042 | 0.041 | 0.053 | 0.065 | 0.079 | 0.092 | 0.104 |
4. In distribution $F^*(S_{tn})$ distinguish the bottom $\bar{S}_{tn}$ and top $\bar{S}_{tn}$ boundary values with a significance value $\alpha$, i.e.

$$F^\bar{S}_{tn}(\frac{\alpha}{2},(1-\frac{\alpha}{2})$$  \hspace{1cm} (11)$$

5. It is established, that if $0.25 \leq F^*(S_{tn}) \leq 0.75$, i.e. if $\alpha \leq 0.5$

$$\bar{S}_{tn} = \left(\frac{1}{n} + \frac{St}{n}\right)$$  \hspace{1cm} (12)$$

For example, for $n=4$ and $\alpha=0.10$ according to distribution $F^*(S_{tn})$ (see tabl.2) $\bar{S}_{tn} = -0.567$, and $\bar{S}_{4} = +0.319$. At the same time under the formula (12)

$$-(0.25-0.567=0.317=\bar{S}_{4})$$

If $n=29$ and $\alpha=0.2$, that $\bar{S}_{tn} = -0.193$ and $\bar{S}_{n} = 0.158$. The size $\bar{S}_{n}$ under the formula (12) is equal - (0.034-0.193)=0.159

On fig. 3 histograms of distribution of negative and positive values $S_{tn}$ for $n=4$ and $n=29$ are resulted.

As follows from fig. 3, negative values $S_{tn}$ essentially exceed positive values $S_{tn}$ on relative number and an interval of change. Proceeding from i. 3 it is clear, that it not casually and does not testify about unpresentable samples. With growth $n$ the parity of negative and positive values $S_{tn}$ decreases and aspires to unit. For $n=2$ negative values $S_{tn}$ make 87.5%, and for $n=29$ - 61%, and for
n=150 – 55%. Thus, even at n=150 quintile distributions $F^*(S_{tn})$ at $\alpha=0.05$ and $\alpha==0.95$ are not equal $[-0.099; +0.092]$. Histograms also explain laws of distribution $F^*(S_{tn})$ resulted on fig.2.

On fig. 4 curve changes of boundary values of statistics $S_{tn}$ for of some values s.f.d. are resulted, $F^*(S_{tn})$. Criterion of the control of imposing appearance of sample $\{\xi\}_n$ with a significance value $\alpha$ thus looks like:

$$S_{tn} < S_{tn} < \overline{S_{tn}}$$

(13)

![Fig.4.](image)

Fig.4. Laws of change of boundary values of the greatest divergence of distributions $F\Sigma(\xi)$ and $F^*_V(\xi)$

Let's designate positive values $S_{tn}$ through $S_{tn}^+$, and negative values- $S_{tn}^-$

In view of i.1. and the equations (12), sample $\{\xi\}_n$ with a significance value $\alpha\leq0.5$ can be accepted representative, if

$$S_{tn}^+ < \left\lfloor d_{n,(1-2\alpha)} - \frac{1}{n} \right\rfloor < \left\lfloor S_{tn}^- \right\rfloor < d_{n,(1-2\alpha)}$$

(14)

As

$$\left( S_{tn}^+ + \frac{1}{n} \right) = \left| S_{tn}^- \right|$$

criterion (13) for a significance value $\alpha$ can be presented, as

$$\left( S_{tn}^+ + \frac{1}{n} \right) = \left| S_{tn}^- \right| = d_{n,(1-2\alpha)}$$

(15)

Here it is necessary to pay attention to discrepancy of the equations of importance $S_{tn}$ and $d_{n,(1-2\alpha)}$.

If again to address to data of table 1 it is easy to notice, that the interval criterion (13), allowing to consider a sign on the greatest divergence $S_{tn}$, also is unable to establish unpresentable character of sample $\{\psi\}_n$.

It is known, that decrease in risk of the erroneous decision at classification of data can be reached by the account not only errors I type, but also the II types [4].

The most simple decision of this problem would be comparison $S_{tn}$ between $F\Sigma(\xi)$ and $F^*_V(\xi)$ with boundary values of the interval $[S_{tn};S_{tn}]$ corresponding a significance value $\alpha=0.5$. It is that limiting case of values $\alpha$ when $S_{tn}=0$. Thus a errors II type $\beta=(1-\alpha)$, i.e. also it is equal 0.5. If $\alpha$ to accept it is less, than 0.5 the errors II type increases $\beta$.

In real conditions:
- configurations $F\Sigma(\xi)$ also $F^*_V(\xi)$ are various, i.e. $S_{tn}\neq0$;
- for the same value $S_{tn}$ size $(\alpha+\beta)$ less or it is equal to unit;
- in process of increase $S_{tn}$ size $(\alpha+\beta)$ decreases, reaches the minimum ($S_{tn,opt}$) and then increases;
- if $S_{tn} < S_{tn, opt}$, then $\alpha > \beta$, if $S_{tn} > S_{tn, opt}$, then $\alpha < \beta$;

- distinction between $\alpha$ and $\beta$ increases in process of increase in a divergence between $S_{tn}$ and $S_{tn, opt}$.

Comparison of realizations $S_{tn}$ to boundary values $S_{tn}$ and $S_{tn, opt}$, calculated accordingly, for $F^*(S_{tn}) = 0.25$ and $F^*(S_{tn}) = 0.75$, allows to not calculate s.f.d., which defines a errors II type $\beta$, that it is possible to carry to advantages of this way. Its lacks are necessity of increase twice numbers of modeled realizations of distribution $F^*_V(\xi)$, unjustified decrease in disorder $S_{tn}$, the heuristic approach.

Algorithm of calculation s.f.d., describing the greatest deviation $F_\Sigma(\xi)$ and $F^*_V(\xi)$, provided that $F^*_V(\xi)$ it is unrepresentable, consists of following sequence of calculations:

1. It is modeled next (from necessary N realizations) their sample $n$ random numbers;
2. It is formed s.f.d. $F^*_V(\xi)$;
3. The greatest divergence between $F$ is defined $F_\Sigma(\xi)$ and $F^*_V(\xi)$. Designate this size as $S_{tn,e}$ where the index «e» corresponds to empirical character of sample. Having defined statistical characteristics of this sample $\{F^*_V(\xi)\}$ and $S_{tn,e}$, start formation s.f.d. $F^*_V(\psi)$ on realizations of the greatest divergence between functions of distribution $F_\Sigma(\xi)$ and set (N) s.f.d. $F^*_V(\psi)$, modeled on s.f.d. $F^*_V(\xi)$. For what:
4. On s.f.d. $F^*_V(\xi)$ distribution is formed

$$F^*_V(\psi) = \begin{cases} 
0 & \text{if } \psi \leq \psi_1 \\
\frac{i-1}{n+1} + \frac{(\psi - \psi_i)}{(\psi_{i+1} - \psi_i)(n+1)} & \text{if } \psi_1 < \psi < \psi_{n+1} \\
1 & \text{if } \psi \geq \psi_{n+1} 
\end{cases} \quad (16)$$

5. Under standard program RAND the random number is modeled $\xi$ with uniform distribution in an interval $[0,1]$;
6. On distribution (16) calculated corresponding probability $\xi$ random number $\psi$. Calculations are spent under the formula

$$\psi = \psi_i + (\psi_{i+1} - \psi_i)([\xi \cdot (n+1) - (i-1)]$$

with $i=1, (n+1)$

7. Items 5 and 6 repeat n time;
8. On sample $\{\psi\}_n$ is under construction s.f.d. $F^*_V(\psi)$;
9. The greatest divergence between $F_\Sigma(\xi)$ and $F^*_V(\psi)$ is defined. Designate it through $S^*_n$;
10. Items (5+9) will repeat N time;
11. Average value of a random variable $S^*_n$ defined. Designate it through $M^*(S^*_n)$;
12. On N to values, $S^*_n$ it is formed s.f.d. $F^*(S^*_n)$.

If to assume, that distribution $F^*(S^*_n)$ corresponds to the normal law of distribution, average value $M^*(S^*_n)$ is equal $S_{tn,e}$ and corresponds $F^*(S^*_n) = \beta = 0.5$, for all realizations $S_{tn,e}$, which probability $0.1 < \alpha < 0.5$, the preference should be given to assumption $H_2$. However, the assumption of the normal law of distribution of function $F^*(S^*_n)$ mismatches the validity. As an example on fig.5 the histogram of distribution of realizations $S^*_n$ for s.f.d. is resulted. $F^*_V(\psi)$, resulted in table 1.
Let’s enter into consideration two assumptions:

- \( H_1 \): sample \( \{ \psi \}_n \) reflects laws of distribution \( F_2(\xi) \);
- \( H_2 \): sample \( \{ \psi \}_n \) does not reflect law of distribution \( F_2(\xi) \).

The recommended algorithm of decision-making depends on a parity of average values of realizations \( St_n \) and \( St_n^* \). In this connection the distribution describing risk of the erroneous decision in function \( St_n \) designate \( Sh1(St_n) \), and in function \( St_n^* - Sh2(St_n) \).

\[
\text{At } M'(St_n) < M'(St_n^*) \\
\left\{
\begin{array}{l}
Sh1(St_n) = 1 - F'(St_n) \\
Sh2(St_n) = F(St_n^*)
\end{array}
\right.
\]  

Algorithm of decision-making looks like:

\[
\left\{
\begin{array}{l}
\text{If } St_n \geq St_n^*, \text{ then } H_2, \text{ else } \\
\text{If } St_n \leq St_n^*, \text{ then } H_1, \text{ else } \\
\text{If } Sh1(St_n) \ll Sh2(St_n), \text{ then } H_2, \text{ else } H_1
\end{array}
\right.
\]  

(18)

At \( M'(St_n) \geq M'(St_n^*) \)

\[
\left\{
\begin{array}{l}
Sh1(St_n) = 1 - F'(St_n^*) \\
Sh2(St_n) = F'(St_n^*)
\end{array}
\right.
\]  

Algorithm of decision-making looks like:

\[
\left\{
\begin{array}{l}
\text{If } St_n \geq St_n, \text{ then } H_1, \text{ else } \\
\text{If } St_n \leq St_n^*, \text{ then } H_2, \text{ else } \\
\text{If } Sh1(St_n^*) \gg Sh2(St_n), \text{ then } H_2, \text{ else } H_1
\end{array}
\right.
\]  

(19)

(20)

(21)

In the illustrative purposes on fig. 6 functions of distribution \( Sh1(St_n) \) and \( Sh2(St_n) \) are resulted, calculated according to table 1.
Fig. 6. Laws of change s.f.d. $F^*(St_n)$ and $F'(St^*_n)$ for $n=4$: a – s.f.d. $F^*(St_n)$; b - $F'(St^*_n)$

As $M^*(St_n)$ it has appeared less than $M^*(St^*_n)$ functions of distribution $Sh1(St_n)$. and $Sh2(St_n)$, were calculated accordingly under the formula (18).

In table 3 numerical values of the parameters defining result of the decision are systematized. As follows from tab. 3 as $Sh1(St_{n,e})<Sh2(St_{n,e})$, the preference, according to (19) is given assumption $H_2$. In other words, attraction to the statistical analysis of size of a errors I type and errors II types, allows distinguish unrepresentable samples.

| The basic parameters of calculation |
|------------------------------------|
| Parameter | Conditional designation | Estimation |
|-----------|-------------------------|------------|
| 1. Number casual sample | $n$ | 4 |
| 2. Average value of the greatest divergence of distributions $F_\xi(\xi)$ and $F'_\psi(\psi)$ | $M^*(St_n)$ | -0.207 |
| 3. Average value of the greatest divergence of distributions $F_\psi(\psi)$ and $F'_\psi(\psi)$ | $M^*(St^*_n)$ | 0.292 |
| 4. Empirical value of the greatest divergence of distributions $F_\xi(\xi)$ and $F'_\psi(\psi)$ | $St_{n,e}$ | 0.257 |
| 5. Boundary values of an interval of change $St_n$ c $\alpha=0.1$ | $St_n$ | 0.319 |
| top | $St_n$ | -0.567 |
| bottom | $St_n$ | 0.544 |
| 6. Boundary values of an interval of change $St^*_n$ with $\alpha=0.01$ | $St^*_n$ | 0.292 |
| top | $St^*_n$ | 0.09 |
| bottom | $St^*_n$ | 0.42 |
| 7. Probability $St_{n,e}$ on s.f.d. $[1-F'(St_n)]$ | $Sh1(St_{n,e})$ | |
| on s.f.d. $F'(St^*_n)$ | $Sh2(St_{n,e})$ | |
| 8. The assumption is accepted | $H$ | $H_2$ |

It is necessary to note, that attraction to an estimation of character of a divergence of distributions $F_\xi(\xi)$ and $F'_\psi(\psi)$ distributions $F'(St^*_n)$ for all realizations samples it is unjustified, as for of some from them, for example at $Sh1(St_n)\geq0.5$ sample $\{\psi\}_n$ it is most truly representative, and at $Sh1(St_n)\leq0.1$ – it is unrepresentable.

There for calculations s.f.d. $F'(St^*_n)$ offered to spend for following conditions:
1. \(M^*(St_n) < M^*(St_n^*)\)
   \[
   \begin{align*}
   St_{n,0.05} &< St_{n,0.95} < St_{n,0.95} \\
   St_{n,0.25} &> St_{n,0.95} > St_{n,0.75}
   \end{align*}
   \]
   \(22\)

2. \(M^*(St_n) > M^*(St_n^*)\)
   \[
   \begin{align*}
   St_{n,0.05} &< St_{n,0.95} < St_{n,0.95} \\
   St_{n,0.25} &> St_{n,0.95} > St_{n,0.75}
   \end{align*}
   \]
   \(23\)

Critical values of statistics \(St_n\) for \(F*(St_n)=0.25\) and average values \(M^*(St_n)\) for \(N=25000\) realizations \(St_n\) and of some \(n\) are resulted in table 4.

**Table 4**

| \(N\) | \(n\) | \(St_n\) \((F^*(St_n)=0.25)\) | \(M^*(St_n)\) | \(N\) | \(n\) | \(St_n\) \((F^*(St_n)=0.25)\) | \(M^*(St_n)\) |
|---|---|---|---|---|---|---|---|
| 1 | 2 | -0.498 | -0.33 | 9 | 22 | -0.17 | -0.047 |
| 2 | 3 | -0.435 | -0.254 | 10 | 29 | -0.149 | -0.037 |
| 3 | 4 | -0.385 | -0.207 | 11 | 40 | -0.127 | -0.027 |
| 4 | 5 | -0.343 | -0.173 | 12 | 60 | -0.105 | -0.019 |
| 5 | 6 | -0.312 | -0.146 | 13 | 90 | -0.086 | -0.012 |
| 6 | 7 | --0.294 | -0.133 | 14 | 120 | -0.074 | -0.00- |
| 7 | 11 | -0.235 | -0.087 | 15 | 150 | -0.067 | -0.008 |
| 8 | 16 | -0.198 | -0.063 |   |   |   |   |

The computer technology of an estimation of parameters of individual reliability assumes automation of process of classification of multivariate data. For what, as initial data boundary values of statistics \(St_n\) should entered. In this connection, by analogy to formulas (8) and (9), the opportunity of an estimation of dependence of boundary values \(St_n\) from \(n\) was of interest.

The equations of regress received under the standard program of sedate transformation, are characterized by factor of determination \(R^2\): (\(R^2> 0.999\)) and for of some \(Sh1(St_n)=\alpha/2\) look like:

- for \(Sh1(St_n)=0.025\) \(St_n = (1.23n^{0.52} - 1)/n = (B_1n^{0.52} - 1)/n\)
  \(24\)
- for \(Sh1(St_n)=0.075\) \(St_n = -1.23n^{-0.48} = -B_1/n^{0.48}\)
  \(25\)
- for \(Sh1(St_n)=0.05\) \(St_n = (1.12n^{0.52} - 1)/n = (B_2n^{0.52} - 1)/n\)
  \(26\)
- for \(Sh1(St_n)=0.095\) \(St_n = -1.12n^{-0.48} = -B_2/n^{0.48}\)
  \(27\)
- for \(Sh1(St_n)=0.1\) \(St_n = (0.98n^{0.52} - 1)/n = (B_3n^{0.52} - 1)/n\)
  \(28\)
- for \(Sh1(St_n)=0.9\) \(St_n = -0.98n^{-0.48} = -B_3/n^{0.48}\)
  \(29\)
- for \(Sh1(St_n)=0.25\) \(St_n = (0.75n^{0.52} - 1)/n = (B_4n^{0.52} - 1)/n\)
  \(30\)
- for \(Sh1(St_n)=0.75\) \(St_n = -0.75n^{-0.48} = -B_4/n^{0.48}\)
  \(31\)

The equation of dependence of constant factors \(B\) from \(\alpha\) with factor of determination \(R^2\): (\(R^2> 0.993\)) looks like:

\[B = 0.652[Sh1(St_n)]^{-0.175}\]

Thus, the bottom and top boundary values of statistics \(St_n\) in view of the equation (12) calculated under following formulas:

\[
\begin{align*}
St_n &= -0.652[Sh1(St_n)]^{0.175} \cdot n^{-0.48} \\
\overline{St_n} &= -\left[\frac{St_n - 1}{n}\right]
\end{align*}
\]
For practical calculations $\bar{S}_n$ and $\bar{S}_T$ more often formulas (27) and (12) used.

**CONCLUSIONS**

1. The interval nonparametric criterion of the control of conformity samples from $n$ pseudo-random numbers is offered to the uniform law in an interval $[0,1]$;
2. In a basis of criterion there is a distinction of distributions of positive and negative values of the greatest divergence of distributions $F_{\xi}(\xi)$ and $F_{\psi}(\xi)$;
3. Transition from statistics $D_n$ to statistics $S_T$ more often formulas allows not only to simplify algorithm of calculation greatest divergences $F_{\xi}(\xi)$ and $F_{\psi}(\xi)$, but also to estimate an opportunity of use of statistics $S_T$ at an estimation of the greatest divergence s.f.d. $F_{\xi}(X)$ and $F_{\psi}(X)$, to estimate risk of the erroneous decision $Sh_1(S_T)$;
4. Increase of accuracy of the control of conformity of distribution $S_T^n$ to the uniform law reached by practical realization of recommended algorithm of the decision-making considering not only a errors I type, but also the errors II type.

**LITERATURE**

1. Farhadzadeh E.M., Muradaliyev A.Z., Farzaliyev Y.Z. Quantitative estimation of individual reliability of the equipment and devices of the power supply system. Journal: «Reliability: Theory&applications. R&RATA (Vol.7 No.4 (27)) 2012, December., USA, p.53-62
2. Gnedenko B.V., Beljaev J.K., Solovyov A.D. Mathematical methods in the theory of reliability. "Science", 1965, 524 p.
3. Kelton B, Law A. Imitational modeling. Classics CS. 3 CPб.: Peter, Kiev: Publishing group BHV, 2004, 847 p.
4. Ryabinin I.A. The heart of the theory and calculation of reliability of ship electro power systems. Shipbuilding. 1971, 454 p.
5. Farhadzadeh E.M. Technique of a statistical estimation of critical values of empirical distribution from theoretical. «Methodical questions of research of reliability of greater systems of power» SEI SO SA USSR, 16, Grozny, 1978, p.39-49
Kernel Type Estimator of a Bivariate Average Growth Function

István Fazekas¹, Zsolt Karácsony², and Renáta Vas³

¹ Department of Applied Mathematics and Probability Theory, Faculty of Informatics, University of Debrecen, Kassai út 26, H-4028 Debrecen, Hungary (E-mail: fazekasi@inf.unideb.hu)
² Department of Applied Mathematics, Faculty of Mechanical Engineering and Informatics, University of Miskolc, H-3515 Miskolc-Egyetemváros, Hungary (E-mail: matkzs@uni-miskolc.hu)
³ Department of Applied Mathematics and Probability Theory, Faculty of Informatics, University of Debrecen, Kassai út 26, H-4028 Debrecen, Hungary (E-mail: vas.renata@inf.unideb.hu)

Abstract. The usual regression problem is considered and a bivariate function with general error is observed. The Gasser-Müller method is used to obtain the estimator of the unknown function. Under general and realistic conditions on the covariance structure of the error random field an upper bound is obtained for the mean squared error.

Keywords: Average growth function, kernel, Gasser-Müller estimator, mean squared error, bandwidth, Taylor series expansion.

1 Introduction

In this paper, we are dealing with the nonparametric estimation of the average growth function. The theory and the methods connecting to this subject have been developed intensively since the beginning of the sixties. Among nonparametric estimations one can speak about estimating the density function and the regression function. Several papers and monographs are published about them e.g. [14], [5] and [6]. Here, it is essential to consider the asymptotic behaviour of the estimations (e.g. the asymptotic behaviour of their error). Both the estimation of the density function and that of the regression function were extensively studied in the dependent case (i.e. for time series) and in the spatial case (i.e. for random fields). It is worth mentioning the case of infill-increasing sampling scheme (it means that the sequence of domains is increasing and at the same time the locations of observations become dense). Using this setup, [13] studied the asymptotic behaviour of the empirical distribution function. The importance of the infill-increasing sampling method is underlined in [7] and [12].

The growth curve model is used mostly in applied sciences such as biostatistics, medical research, epidemiology (see [15]) and pharmacokinetic research. In the case of time series, the points of observations are usually taken equally spaced during the process of sampling (it is called fixed-design model) but other types of sampling method can also be considered e.g. random design. In practical situations, typically repeated measurements are used and those can make
the estimators of the function $f$ asymptotically consistent. This was examined for the kernel regression estimators e.g. in the papers [11] and [10]. In [8], the authors dealt with the case when some monotone transformation of the time scale was allowed. The nonparametric regression model with correlated errors was considered in several papers, too. For example, in [1], the kernel construction of the regression estimator was modified, in [2], there are examinations for models with weakly stationary second order error processes and [3] contains the study of estimating the regression function from quantized observations.

In [4], the nonparametric estimation of the growth curve is considered in the case when there is a nonstationary error process and the autocovariance function does not have any specific form. More precisely, the following model was considered there:

$$Y_j(x_i) = f(x_i) + \varepsilon_j(x_i), \quad (j = 1, \ldots, m, \ i = 1, \ldots, n)$$

where $f$ is the unknown average growth curve and $(\varepsilon_j)$ is the error process. The model means that $m$ experimental units are taken and each of them produces $n$ measurements. The estimator of $f$ is based on the observations

$$\{Y_j(x_i) = f(x_i) + \varepsilon_j(x_i), \ j = 1, \ldots, m, \ i = 1, \ldots, n\}$$

where the locations of the observations are chosen as follows:

$$0 = x_1 < x_2 < \ldots < x_n = 1, \quad \max_i |x_i - x_{i-1}| = O\left(\frac{1}{n}\right).$$

The estimator of $f$ given by Gasser and Müller has the form

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} W_{h,i}(x)\overline{Y}(x_i)$$

where

$$\overline{Y}(x) = \frac{1}{m} \sum_{j=1}^{m} Y_j(x)$$

is the average of the observations in the point $x$,

$$W_{h,i}(x) = n \int_{m_{i-1}}^{m_i} \frac{1}{h} K\left(\frac{x-u}{h}\right) \, du$$

and the midpoints $\{m_i, \ i = 0, \ldots, n\}$ are defined by

$$m_0 = 0, \ m_i = \frac{x_i + x_{i+1}}{2} \quad \text{for} \ i = 1, \ldots, n-1, \ m_n = 1.$$ 

In [4], under realistic conditions on the covariance structure of the error process $(\varepsilon_j)$, an asymptotic upper bound was derived for the mean squared error

$$\mathbb{E}(\hat{f}_h(x) - f(x))^2$$

and the optimal bandwidth was also determined.

For claiming precisely the theorem, they had some assumptions:
(i) The autocovariance function $\rho$ of $(\epsilon_j)$ exists and is continuous on the square $[0,1]^2$.
(ii) $\rho(x,y)$ has left and right first-order derivatives at the diagonal $x = y$, that is
$$\rho^{(0,1)}(x,x^-) = \lim_{y \nearrow x} \frac{\partial \rho}{\partial y}(x,y) \quad \text{and} \quad \rho^{(0,1)}(x,x^+) = \lim_{y \searrow x} \frac{\partial \rho}{\partial y}(x,y)$$
exist and are continuous.
The jump function along the diagonal $\alpha(x) = \rho^{(0,1)}(x,x^-) - \rho^{(0,1)}(x,x^+)$ is assumed to be continuous and not identically equal to zero.
(iii) $\rho(x,y)$ is assumed to have continuous mixed partial derivatives up to order two off the diagonal $x \neq y$ in the unit square and satisfies:
$$A(i,j) = \sup_{0 \leq x \neq y \leq 1} |\rho(i,j)(x,y)| < \infty \quad \text{for all integers } i, j \text{ such that } 0 \leq i+j \leq 2.$$ And the theorem of [4] was the following:
If the autocovariance function $\rho$ satisfies Assumptions (i)-(ii)-(iii) and $f$ is a twice differentiable continuous function on $[0,1]$ with $f''(x) \neq 0$ for $0 < x < 1$, then as $n, m \to +\infty$,
$$E(f_h(x) - f(x))^2 = \frac{1}{m} \left( \rho(x,x) - \frac{1}{2} \alpha(x)C_K h \right)$$
$$+ \frac{h^4}{4} d_K^2(f''(x))^2 + O \left( \frac{1}{mn} + \frac{h^2}{n} \right) + o \left( \frac{h^4}{m} \right)$$
where
$$C_K = \int_{-1}^{1} \int_{-1}^{1} |u - v| K(u)K(v)du dv \quad \text{and} \quad d_K = \int u^2 K(u)du.$$ Moreover, if even $m/n = O(1)$ as $n, m \to +\infty$ is satisfied, then the asymptotic MSE is minimized by taking the bandwidth
$$h^*_x = \left( \frac{\alpha(x)C_K}{2d_K^2(f''(x))^2} \right)^{1/3} m^{-1/3}.$$ In this paper, we consider the spatial version of the above model. We get the analogue of the result of [4] for the spatial case.
More precisely, we shall study the problem of estimating the average growth function for the following fixed design model. We have $m$ experimental units, each of them producing $n \times n$ measurements of the response:
$$Y_k(u,v) = f(u,v) + \epsilon_k(u,v) \quad (k = 1, \ldots, m). \quad (1)$$ Here, $f : [0,1]^2 \to \mathbb{R}$ is the unknown average growth function and $(\epsilon_k)$ is the error random field. We assume that the random fields $\{\epsilon_k(x), x \in [0,1]^2, k =$
1, . . . , m\} are identically distributed and independent. Therefore they have the same covariance function

\[ \varrho(x, y) = \text{cov}(\varepsilon_k(x), \varepsilon_k(y)), \quad x, y \in [0, 1]^2. \]

We assume that \( \mathbb{E}\varepsilon_k(x) = 0, x \in [0, 1]^2 \). The estimator of \( f \) is based on the observations

\[ \{Y_k(u_i, v_j), \quad i, j = 1, \ldots, n, \quad k = 1, \ldots, m\} \tag{2} \]

where the locations of the observations are chosen as follows:

\[
0 = u_1 < u_2 < \ldots < u_n = 1, \quad \max_i |u_i - u_{i-1}| = O \left( \frac{1}{n} \right),
\]

\[
0 = v_1 < v_2 < \ldots < v_n = 1, \quad \max_j |v_j - v_{j-1}| = O \left( \frac{1}{n} \right). \]

The Gasser-Müller estimator of the bivariate function \( f \) has the form

\[
\hat{f}_h(u, v) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_{h,i,j}(u, v) \overline{Y}(u_i, v_j), \quad u, v \in [0, 1] \tag{3}
\]

where

\[ \overline{Y}(u, v) = \frac{1}{m} \sum_{k=1}^{m} Y_k(u, v) \]

is the average of the observations in the point \((u, v)\),

\[
W_{h,i,j}(u, v) = n^2 \int_{a_{i-1}}^{a_i} \int_{b_{j-1}}^{b_j} \frac{1}{h^2} \tilde{K} \left( \frac{(u, v) - (x, y)}{h} \right) dx dy.
\]

For simplicity, we use product type kernel functions i.e. \( \tilde{K}(x, y) = K(x)K(y) \) where \( K(\cdot) \) is a univariate kernel function. Therefore

\[
W_{h,i,j}(u, v) = n \int_{a_{i-1}}^{a_i} \frac{1}{h} K \left( \frac{u - x}{h} \right) dx \cdot n \int_{b_{j-1}}^{b_j} \frac{1}{h} K \left( \frac{v - y}{h} \right) dy
\]

and the midpoints \( \{a_i, i = 0, \ldots, n; \quad b_j, j = 0, \ldots, n\} \) are defined by

\[
a_0 = 0, \quad a_i = \frac{u_i + u_{i+1}}{2} \quad \text{for} \quad i = 1, \ldots, n - 1, \quad a_n = 1,
\]

\[
b_0 = 0, \quad b_j = \frac{v_j + v_{j+1}}{2} \quad \text{for} \quad j = 1, \ldots, n - 1, \quad b_n = 1.
\]

Section 2 contains the main result of the paper. It is an asymptotic upper bound for the mean squared error of the estimator \( \hat{f}_h \) when both the bandwidth \( h \) goes to zero and the number \( n \) of observations goes to infinity. The upper bound is expressed in terms of the derivatives of the covariance function. Its proof is based on the appropriate versions of the Taylor expansion. We have to divide the domains carefully and in each subdomain we have to find an adequate version of the Taylor expansion to avoid the singularities of the covariance function. Therefore, our result can be used for several particular covariance structures.
2 The main result

Our aim is to find the asymptotic behaviour of the mean squared error (MSE). We want to find realistic assumptions for the covariance function of $\varepsilon_k$.

(i) Let $K$ be a kernel function, that is, $K : (-\infty, \infty) \to [0, \infty)$, $K$ is even, $K$ is zero outside $[-1, 1]$, $\int_{-1}^{1} K(u)du = 1$.

Assume that the two variable kernel function $\hat{K}$ is of product form, i.e.

$$\hat{K}(u, v) = K(u)K(v), \quad u, v \in \mathbb{R}.$$ 

Let $d_K = \int_{-1}^{1} u^2 K(u)du$.

(ii) Assume that the average growth function $f : [0, 1]^2 \to \mathbb{R}$ has partial derivatives up to order three and the third order partial derivatives are bounded. Assume that the autocovariance of the random field $\varepsilon_k(u, v)$ exists. Let

$$\varrho((u_1, v_1), (u_2, v_2)) = Var(\varepsilon_k(u_1, v_1), \varepsilon_k(u_2, v_2)), \quad u_1, u_2, v_1, v_2 \in [0, 1]$$

hold ($k = 1, \ldots, m$).

(iii) The first order partial derivatives of $\varrho$ exist outside the diagonal, i.e.

$$\frac{\partial}{\partial x_1} \varrho((x, y), (s, t)) \text{ exist if } x \neq s, \quad \frac{\partial}{\partial x_2} \varrho((x, y), (s, t)) \text{ exist if } y \neq t,$$

moreover,

$$\sup_{x \neq s} \left| \frac{\partial}{\partial x_1} \varrho((x, y), (s, t)) \right| \leq A_1 < \infty, \quad \sup_{y \neq t} \left| \frac{\partial}{\partial x_2} \varrho((x, y), (s, t)) \right| \leq A_1 < \infty$$

where $\frac{\partial}{\partial x_1} \varrho((x, y), (s, t))$ and $\frac{\partial}{\partial x_2} \varrho((x, y), (s, t))$ denote the first order partial derivatives of $\varrho$ with respect to the first and the second variables, respectively.

(iv) The one-sided limits

$$\lim_{\xi \to u, \xi < u} \frac{\partial}{\partial x_1} \varrho((\xi, v_1), (u, v_2)) = \frac{\partial}{\partial x_1} \varrho((u^-, v_1), (u, v_2)),$$

$$\lim_{\xi \to u, \xi > u} \frac{\partial}{\partial x_1} \varrho((\xi, v_1), (u, v_2)) = \frac{\partial}{\partial x_1} \varrho((u^+, v_1), (u, v_2)),$$

$$\lim_{\eta \to v, \eta < v} \frac{\partial}{\partial x_2} \varrho((u_1, \eta), (u_2, v)) = \frac{\partial}{\partial x_2} \varrho((u_1, v^-), (u_2, v)),$$

$$\lim_{\eta \to v, \eta > v} \frac{\partial}{\partial x_2} \varrho((u_1, \eta), (u_2, v)) = \frac{\partial}{\partial x_2} \varrho((u_1, v^+), (u_2, v))$$

exist.

We say that $\frac{\partial}{\partial x_1} \varrho((x, y), (s, t))$ is continuous on the domain $x \leq s$ if it is continuous as a four variable function if $\frac{\partial}{\partial x_1} \varrho((s, y), (s, t))$ is interpreted as $\frac{\partial}{\partial x_1} \varrho((s^-, y), (s, t))$. The continuity on the domain $x \geq s$ is defined analogously. We say that $\frac{\partial}{\partial x_1} \varrho((x, y), (s, t))$ is continuous outside the diagonal if it is continuous both on $x \leq s$ and $x \geq s$. 

363
(v) Let $\frac{\partial}{\partial x_1} \varrho((x, y), (s, t))$ and $\frac{\partial}{\partial x_2} \varrho((x, y), (s, t))$ be continuous outside the diagonal.

(vi) The second-order partial derivatives $\frac{\partial^2}{\partial x_1 \partial x_3} \varrho((x, y), (s, t))$, $\frac{\partial^2}{\partial x_4 \partial x_1} \varrho((x, y), (s, t))$, $\frac{\partial^2}{\partial x_2 \partial x_2} \varrho((x, y), (s, t))$ and $\frac{\partial^2}{\partial x_2 \partial x_4} \varrho((x, y), (s, t))$ exist outside the diagonal (that is, for all $x, y, s, t \in [0, 1]$ except when $x = s$ or $y = t$) and they are bounded.

**Theorem 1** Let the model $Y_k(x, y)$ be defined by (1), the sample $Y_k(u_i, v_j)$ by (2) and the estimator $\hat{f}_h(u, v)$ by (3). Assume that the function $f$ satisfies assumption (ii), the kernel function $K$ satisfies (i), while the covariance function $\varrho((x, y), (s, t))$ satisfies (iii), (iv), (v) and (vi). Then, as $n \to \infty$ and $h \to 0$, we have that

$$E(\hat{f}_h(u, v) - f(u, v))^2 = \frac{1}{m} \varrho((u, v), (u, v))^2$$

$$- \frac{h}{m} 2B \left\{ \frac{\partial}{\partial x_1} \varrho((u^+, v), (u, v)) - \frac{\partial}{\partial x_1} \varrho((u^-, v), (u, v)) + \frac{\partial}{\partial x_2} \varrho((u, v^+), (u, v)) - \frac{\partial}{\partial x_2} \varrho((u, v^-), (u, v)) \right\}$$

$$+ \frac{1}{4} h^4 dK (\frac{\partial^2}{\partial x_1^2} f(u, v) + \frac{\partial^2}{\partial x_2^2} f(u, v))^2 + \left( \frac{h^2}{m} + \frac{1}{mn} + \frac{1}{n^2} \right) O(1)$$

where $dK = \int_{-1}^{1} u^2 K(u) du$ and $B = \int_{-1}^{1} \int_{-1}^{1} xK(x)K(s) ds dx$.

**Acknowledgement**

The publication was supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project has been supported by the European Union, co-financed by the European Social Fund.

The work is supported by the TÁMOP-4.2.2/B-10/1-2010-0024 project. The project is co-financed by the European Union and the European Social Fund.

This research was supported by the European Union and the State of Hungary, co-financed by the European Social Fund in the Framework of TÁMOP 4.2.4. A/2-11-1-2012-0001 ‘National Excellence Program’.

**References**

1. N. Altman: Kernel regression estimation of data with correlated errors. (1990) *Journal of the American Statistical Association* **85** 749–759.

2. K. Benhenni, S. Hedli-Griche, M. Rachdi: Estimation of the regression operator from functional fixed-design with correlated errors. (2010) *Journal of Multivariate Analysis* **101** 476–490.

3. K. Benhenni, M. Rachdi: Nonparametric estimation of the regression function from quantized observations. (2006) *Computational Statistics and Data Analysis* **50** 3067–3085.

4. K. Benhenni, M. Rachdi: Nonparametric estimation of average growth curve with general nonstationary error process. (2007) *Communications in Statistics - Theory and Methods* **36** 1173–1186.

5. D. Bosq: Nonparametric Statistics for Stochastic Processes. (1998) *Springer, New York - Berlin - Heidelberg.*
6. L. Devroye, L. Györfi: Nonparametric density estimation. The $L_1$ view. (1985) Wiley, New York.

7. I. Fazekas, A. Chuprunov: Asymptotic normality of kernel type density estimators for random fields. (2006) *Statistical Inference for Stochastic Processes* 9 161–178.

8. E. Ferreira, V. Núñez-Antón, J. Rodríguez-Poo: Kernel regression estimates of growth curves using nonstationary correlated errors. (1997) *Statistics and Probability Letters* 34 413–423.

9. T. Gasser, M.G. Müller: Estimating regression functions and their derivatives by the kernel method. (1984) *Scandinavian Journal of Statistics, Theory and Applications* 11 171–185.

10. W. Härdle: Applied Nonparametric Regression. (1989) *Cambridge University Press, Cambridge*. 19.

11. J. Hart, T. Wehrly: Kernel regression estimation using repeated measurements data. (1986) *Journal of the American Statistical Association* 81 1080–1088.

12. Zs. Karácsony, P. Filzmoser: Asymptotic normality of kernel type regression estimators for random fields. (2010) *Journal of Statistical Planning and Inference*, 140 872–886.

13. S.N. Lahiri: Asymptotic distribution of the empirical spatial cumulative distribution function predictor and prediction bands based on a subsampling method. (1999) *Probability Theory and Related Fields*, 114(1) 55–84.

14. B.L.S. Prakasa Rao: Nonparametric Functional Estimation. (1983) *Academic Press, INC. London*.

15. D. Von Rosen: The growth curve model: a review. (1991) *Communications in Statistics - Theory and Methods* 20(9) 2791–2822.
null
Modeling of Mortality in Older Ages by a Modified Gompertz-Makeham Function

Tomas Fiala\textsuperscript{1}, Petra Dotlacilova\textsuperscript{2}, and Jitka Langhamrova\textsuperscript{3}

\textsuperscript{1} Department of Demography, University of Economics, Prague, nam. W. Churchilla 4, 130 67 Praha 3, Czech Republic (E-mail: fiala@vse.cz)
\textsuperscript{2} Department of Demography, University of Economics, Prague, nam. W. Churchilla 4, 130 67 Praha 3, Czech Republic (E-mail: Petra.Dotlacilova@vse.cz)
\textsuperscript{3} Department of Demography, University of Economics, Prague, nam. W. Churchilla 4, 130 67 Praha 3, Czech Republic (E-mail: langhamj@vse.cz)

Abstract. One of the functions very often used for modeling of mortality at higher age is the Gompertz-Makeham function. Computations for many countries however indicate that for higher old ages (80 years and more) the values of Gompertz-Makeham function are systematically higher than the empirical values of age-specific death rates. Koschin et al.\cite{4} proposed to modify the Gompertz-Makeham function for higher old age. The paper brings computations of modeling mortality of persons of 60 years of age and older both by the standard Gompertz-Makeham function as well as by its modified form. The estimate of unknown parameters of the function is based on the weighted least square method. Computations were carried out in Excel. Results for various European countries are presented.

Keywords: mortality, force of mortality, age-specific mortality rates, Gompertz-Makeham function.

1 Introduction

One of the functions very often used for modeling of mortality at higher ages (usually 60 years and more) is the Gompertz-Makeham function

\[ \mu(x) := a + b \cdot c^x, \]  

where \( x \) denotes the age, \( \mu(x) \) is the force of mortality at the exact age \( x \), and \( a, b, c \) are parameters of the function. According to this model the rate of increase of the force of mortality at higher ages is supposed to be almost constant. If \( a = 0 \) the rate of increase of the force of mortality per one year of age would be equal exactly to the value of \( c \), for \( a \neq 0 \) the rate tends to \( c \) with increasing age.

The estimates of parameters \( a, b \) and \( c \) of the Gompertz-Makeham function were usually based on the data of mortality from the age of 60 years to approximately 80-85 years. For many countries the Gompertz-Makeham function corresponded almost perfectly with the real data of mortality. Data for higher ages were very often unavailable or non-reliable.
The length of life has been permanently increasing in last decades. More and more people survive beyond the age of 85 or more. And computations for many countries show that for the age over 85 years the values of Gompertz-Makeham function are systematically higher than the corresponding empirical values of age-specific death rates. It seems that the original Gompertz hypothesis of constant rate of increase of the force of mortality at higher ages is not valid for the age beyond approximately 85 years.

In Koschin et al.[4] the following modification of the Gompertz-Makeham function for higher ages is proposed

\[ \mu'(x) := a + b \cdot c^x, \quad x \leq x_0, \quad (2) \]

\[ \mu'(x) := a + b \cdot c^{x_0 + \frac{1}{2} \cdot \ln[d(x-x_0)+1]}, \quad x > x_0, \quad (3) \]

where \( d \neq 0 \) is the fourth parameter of the function. The recommended value of the age \( x_0 \) is about 83. This function is also mentioned eg. in Burcin et al.[2].

If \( d > 0 \) the values of the modified function (3) will be lower than the corresponding values of the standard function (1). For \( d < 0 \) but of course \( d > -\frac{1}{2(x-x_0)} \) the modified function would have higher values than the standard one. If \( d \) equals almost to zero the values of the modified function are almost equal to those of the standard function.

In this paper we shall compare the differences in modeling of mortality of 60 years and older males and females in selected European countries by the standard Gompertz-Makeham function (1) and by its modified form (2), (3) respectively.

2 Methodology

The basic data for our analysis were the numbers of deaths by sex and age at last birthday from the latest available year and the numbers of inhabitants by sex and age on 1\textsuperscript{st} January of the year the mortality has been analyzed and on 1\textsuperscript{st} January of the following year. The data from the Eurostat Statistics Database [6] have been used.

All the computations mentioned have been carried out separately for males and females.

The mid-year population in the year \( t \) at the age \( x \) can be estimated by the average of the population size at the given age at the beginning and at the end of the year

\[ S_{t,x} := \frac{S_{t,x} + S_{t+1,x}}{2}, \quad (4) \]

where \( S_{t,x} \) is the number of people at the age \( x \) at the beginning of the year \( t \), \( \tilde{S}_{t,x} \) is the mid-year population size in this age.

The age specific mortality rates have been computed by the well-known formula

\[ m_{t,x} := \frac{M_{t,x}}{S_{t,x}}, \quad (5) \]
where $M_{t,x}$ is the number of deaths of people in the age $x$ in the year $t$.

Under the assumption of uniform distribution of mortality in the age interval $<x;x+1)$ the force of mortality in the middle of this interval, i.e. at the exact age $x + 1/2$ can be estimated by the corresponding age specific mortality rate at the age $x$

$$\mu(x + 1/2) := m_x. \quad (6)$$

Under the usual assumption of independence and identical distribution of the probabilities of deaths the number $M_x$ of persons dying at the age $x$ has binomial distribution

$$M_x := B[\bar{S}_x \cdot \mu(x + 1/2); \bar{S}_x \cdot \mu(x + 1/2) \cdot (1 - \mu(x + 1/2))]. \quad (7)$$

The distribution of age specific mortality rates (using approximation of the binomial distribution by the normal distribution) is then

$$m_x := N[\mu(x + 1/2); \frac{\mu(x + 1/2) \cdot (1 - \mu(x + 1/2))}{\bar{S}_x}] \quad (8)$$

and finally the distribution of the residuals $m_x - \mu(x + 1/2)$ is

$$m_x - \mu(x + 1/2) := N[0; \frac{\mu(x + 1/2) \cdot (1 - \mu(x + 1/2))}{\bar{S}_x}]. \quad (9)$$

The residuals have normal distribution with zero expectation but their variances are different. The least squares method for estimation of the parameters of the modified Gompertz-Makeham function has to be based on minimization of sum of weighted squares.

$$SS'_{60-\beta} := \sum_{x=60}^{z} \frac{\bar{S}_x \cdot \mu'(x + 1/2) \cdot (1 - \mu'(x + 1/2))}{\bar{S}_x} \cdot (m_x - \mu'(x + 1/2))^2. \quad (10)$$

where $z$ is the highest age with "reasonable" empirical value of age specific mortality rate.

Because the Gompertz-Makeham function is non-linear in parameters, the minimization of the sum of weighted squares (10) has been carried out in Excel by the Solver procedure.

3 Main results

Modeling of the mortality beyond the age of 60 by the modified Gompertz-Makeham function has been carried out for selected European countries. We have preferred large countries having more inhabitants which diminishes the random errors in age specific mortality rates at higher age.

For each country the estimates of the parameters of the modified Gompertz-Makeham function (2), (3) (separately for males and females) have been estimated by minimizing the sum of squares (10). The empirical values of age
specific mortality rates are compared with both the values of the Gompertz-Makeham modified function and the values of standard Gompertz-Makeham function (1) with the same estimates of parameters $a$, $b$ and $c$ as in the modified version.

The main results are available in Table 1 and in the graphs below. The systematic drop in the values of age specific mortality rates at the highest ages in Italy, Romania and Russia seems to be unrealistic and indicates probable errors in the data. Data for other countries appear to be reliable.

The growth rate of force of mortality for Russian and Romanian males seems to be low but let us mention that the mortality is higher than in other countries. The force of mortality of Russian males crosses the level of 0.1 already at the age of 78 while in most countries it is as late as at the age of about 80-85.

| Country         | Gender | Maximal age of reliable data $z$ | Sum of squares standard function $SS_{60-z}$ | Sum of squares modified function $SS'_{60-z}$ | Force of mortality at the age of 100 standard function $\mu(100)$ | Force of mortality at the age of 100 modified function $\mu'(100)$ |
|-----------------|--------|----------------------------------|---------------------------------------------|---------------------------------------------|------------------------------------------------|------------------------------------------------|
| Austria         | males  | 106                              | 0.91                                        | 0.44                                        | 2512.58                                      | 146.08                                      |
| Austria         | females | 109                             | 0.88                                        | 0.42                                        | 6727.09                                      | 149.82                                      |
| Czech Republic  | males  | 99                               | 0.71                                        | 0.54                                        | 191.52                                       | 108.65                                      |
| Czech Republic  | females | 99                               | 0.83                                        | 0.47                                        | 1228.20                                      | 104.56                                      |
| France          | males  | 98                               | 0.62                                        | 0.45                                        | 1044.58                                      | 140.02                                      |
| France          | females | 98                               | 0.56                                        | 0.39                                        | 3682.16                                      | 327.85                                      |
| Germany         | males  | 94                               | 0.76                                        | 0.40                                        | 2298.28                                      | 465.37                                      |
| Germany         | females | 94                               | 0.88                                        | 0.49                                        | 4367.63                                      | 626.40                                      |
| Italy           | males  | 109                             | 0.74                                        | 0.42                                        | 16049.80                                     | 611.66                                      |
| Italy           | females | 109                             | 0.69                                        | 0.37                                        | 59406.36                                     | 697.00                                      |
| Netherlands     | males  | 106                             | 0.76                                        | 0.47                                        | 750.11                                       | 127.31                                      |
| Netherlands     | females | 106                             | 0.63                                        | 0.41                                        | 3094.05                                      | 151.85                                      |
| Poland          | males  | 99                               | 0.54                                        | 0.45                                        | 154.05                                       | 84.17                                       |
| Poland          | females | 99                               | 0.79                                        | 0.41                                        | 4976.28                                      | 154.82                                      |
| Romania         | males  | 99                               | 0.77                                        | 0.37                                        | 2601.87                                      | 616.63                                      |
| Romania         | females | 99                               | 1.04                                        | 0.46                                        | 18794.43                                     | 1019.03                                     |
| Russia          | males  | 99                               | 0.52                                        | 0.32                                        | 2736.13                                      | 1382.88                                     |
| Russia          | females | 99                               | 0.88                                        | 0.47                                        | 21064.80                                     | 1559.36                                     |
| Spain           | males  | 99                               | 0.73                                        | 0.37                                        | 4622.98                                      | 292.58                                      |
| Spain           | females | 99                               | 0.76                                        | 0.40                                        | 10565.51                                     | 185.99                                      |
| Sweden          | males  | 105                             | 0.75                                        | 0.51                                        | 996.03                                       | 82.28                                       |
| Sweden          | females | 109                             | 0.60                                        | 0.43                                        | 2431.03                                      | 260.00                                      |
| Switzerland     | males  | 99                               | 0.66                                        | 0.50                                        | 164.14                                       | 50.62                                       |
| Switzerland     | females | 99                               | 0.57                                        | 0.47                                        | 213.98                                       | 100.51                                      |

Table 1. Modeling of mortality in the age of 60 and higher for selected European countries
The differences among countries in the values of the force of mortality at the age of 100 estimated by the standard Gompertz-Makeham function are relatively high. The estimates based on the modified function are much lower and do not mutually differ so much.

4 Conclusions

The trends of mortality at the age 60 and more are almost in all European countries investigated very similar. The force of mortality is rising but the rate
Fig. 7. Germany - males
Fig. 8. Germany - females
Fig. 9. Italy - males
Fig. 10. Italy - females
Fig. 11. Netherlands - males
Fig. 12. Netherlands - females
Fig. 13. Poland - males
Fig. 14. Poland - females
Fig. 15. Romania - males

Fig. 16. Romania - females

Fig. 17. Russia - males

Fig. 18. Russia - females

Fig. 19. Spain - males

Fig. 20. Spain - females

Fig. 21. Sweden - males

Fig. 22. Sweden - females
of growth slows down beyond the age of 80. The modified Gompertz-Makeham function provides a more precise model of mortality in higher ages than the standard function. But for good estimation of the parameters of modified function we need reliable data of mortality by sex and age minimally until 90 years of age or better until 95 or more.

Other functions used for modeling mortality in higher ages can be found eg. in Burcin et al.[2]. The authors also developed special software for this purpose, see Burcin et al.[3]. More particular description of the Thatcher and Kannisto model is in [5] and [1].

References

1. L. Boleslawski, E. Tabeau. Comparing Theoretical Age Patterns of Mortality Beyond the Age of 80, 2001. In: E. Tabeau, A. Van den Berg Jeths, Ch. Heatcote (eds.). Forecasting Mortality in Developed Countries: Insights from a Statistical, Demographic and Epidemiological Perspective. s. 127-155. ISBN 978-0-7923-6833-5, 2001.

2. B. Burcin, K. Tesarkova, and L. Sidlo. Nejpouzívanejší metody vyrovnavání a extrapolační krivky umrtnosti a jejich aplikace na českou populaci. Demografie, 52, 2, 77-89, 2010.

3. B. Burcin, K. Hulikova Tesarkova, D. Komanek (2012). DeRaS: software tool for modelling mortality intensities and life table construction. Charles University in Prague, Prague. http://deras.natur.cuni.cz.

4. F. Koschin, T. Fiala, J. Langhamrova, V. Roubicek. Mortality in the Czech Republic in the Nineties, 1. vyd. Praha : VSE, 1998. 70 pp. ISBN 80-7079-586-7.

5. R., A. Thatcher, V. Kanisto and J., W. Vaupel. The Force of Mortality at Ages 80 to 120. ISBN 87-7838-381-1, 1998.

6. http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/searchdatabase.

The paper was written with the support of Internal Grant Agency of the University of Economics in Prague F4/24/2013 Mortality and Ageing of the Population of the Czech Republic.
Nonparametric Algorithms of Identification, Forecasting and Control for Economic Systems ⋆

Irina L. Fook,1 Irina Yu. Glukhova2 and Gennady M. Koshkin3

1 Department of Computer Science, Tomsk State University, 36 Lenin ave., 634050 Tomsk, Russia
(E-mail: foxil@sibmail.com)
2 Department of Applied Mathematics and Cybernetics, Tomsk State University, 36 Lenin ave., 634050 Tomsk, Russia
(E-mail: win32.86@mail.ru)
3 Department of Applied Mathematics and Cybernetics, Tomsk State University, 36 Lenin ave., 634050 Tomsk, Russia
(E-mail: kgm@mail.tsu.ru)

Abstract. The principal parts of mean square errors for the kernel plug-in estimators of the functions defining the ARX-process are found. To compare parametric and nonparametric identification algorithms and to study nonparametric control algorithms, we use the simulation. To investigate the dependence of Russian Federation’s Industrial Production Index on the dollar exchange rate, direct investments, and export for the period from September 1994 to January 2013, the proposed algorithms of identification and forecasting are applied.

Keywords: Kernel plug-in estimator, conditional mean, mean square error (MSE), ARX-process, nonparametric identification, forecasting algorithm, control.

1 Introduction

Suppose that a sequence \((Y_t)_{t=-\ldots,-1,0,1,\ldots}\) is generated by the ARX\((m,p,d)\)-process

\[Y_t = \Psi(Y_{t,m}, X_{t,s}) + \xi_t,\]

(1)

where \(Y_{t,m} = (Y_{t-i_1}, \ldots, Y_{t-i_m}), X_{t,s} = (X_{t-j_1}^1, \ldots, X_{t-j_1}^p, \ldots, X_{t-j_k}^1, \ldots, X_{t-j_k}^p)\), \(s = r + \ldots + k\), \(d = \max(r, \ldots, k)\), \(1 \leq i_1 < \ldots < i_m \ll n, 0 \leq j_1 < \ldots < j_k \ll n\), \(0 \leq j_1 < \ldots < j_k \ll n\) are known subsequences of natural numbers, \((\xi_t)\) is a sequence of i.i.d. random variables with zero mean, finite variance, zero third, and finite fourth moments, \(\Psi(Y_{t,m}, X_{t,s})\) is an unknown non-periodic function bounded on compact.

Models (1) are widely used for identification of economic systems and financial time series analysis. By identifying the model (1), we mean the problem of parametric or nonparametric estimation of a function \(\Psi\). In this paper, we assume that the process \((Y_t)_{t=-\ldots,-1,0,1,\ldots}\) is a strictly stationary process and satisfies the strong mixing (s.m.) condition with the s.m. coefficient (see Masry[10], Masry and Tjostheim[11], Kitaeva and Koshkin[3], [4])

\[\alpha(\tau) \approx e^{-\delta \tau}, \quad \delta > 0, \quad \tau \to \infty.\]

(2)

⋆ Supported by Russian Foundation for Basic Research, project no. 13-08-00744
Let \( Y_1, Y_2, \ldots, Y_n \) be observations generated by the process (1). As a model of the function of \( \Psi \) in (1), we take the conditional expectation

\[
b(y, x) = E(Y|Y_{t,m} = y, X_{t,s} = x) = E(Y|y, x), \quad (y, x) \in \mathbb{R}^{m+s}.
\]

According to Dobrovidov et al.[1], the integrals

\[
a_g(y, x) = \int q^g f(q, y, x) dq, \quad g = 0, 1,
\]

are basic functionals, where an \( f(q, y, x) \) is an unknown probability density function (p.d.f.) of a random vector \((Y_t, Y_{t,m}, Y_{t,s})\) in stationary conditions. Since

\[
a_0(y, x) = \int f(q, y, x) dq = p(y, x),
\]

where \( p(y, x) \) is the p.d.f. of a vector \((Y_{t,m}, Y_{t,s})\), then the conditional expectation can be written as

\[
b(y, x) = \frac{a_1(y, x)}{a_0(y, x)} = \frac{a_1(y, x)}{p(y, x)} = \int Y_t f(Y_t|y, x) dY_t.
\]

We take the kernel estimators of basic functionals \( a_g(y, x) \) at a point \((y, x)\) in the following form:

\[
a_{gn}(y, x) = \frac{1}{n-Q} \sum_{i=Q+1}^{n} \frac{Y_i}{m} \prod_{j=1}^{r} h_j \left( \frac{y - Y_{i,m}}{h^y} \right) \prod_{j=1}^{m} K_m \left( \frac{x - X_{i,s}}{h^x} \right) \prod_{j=1}^{k} h_{pj},
\]

where \( Q = \max(i_m, \max(j_r, \ldots, j_k)) \), \( h^y = (h_1, \ldots, h_m) \), \( h^x = (h_1^x, \ldots, h_p^x) \), \( h_{1j} = (h_{1j}, \ldots, h_{1r}) \), \( \ldots \), \( h_{pj} = (h_{pj}, \ldots, h_{pk}) \) are suitable bandwidths (positive numbers), \( K_m \) and \( K_s \) are \( m \)- and \( s \)-dimensional kernels. Thus, the kernel plug-in estimator of the conditional functional \( b(y, x) \) at a point \((y, x)\) and, hence, the function \( \Psi(y, x) \) in (1) is the ratio

\[
b_n(y, x) = \Psi_n(y, x) = \frac{\sum_{i=Q+1}^{n} \frac{Y_i}{m} \prod_{j=1}^{r} h_j \left( \frac{y - Y_{i,m}}{h^y} \right) \prod_{j=1}^{m} K_m \left( \frac{x - X_{i,s}}{h^x} \right) \prod_{j=1}^{k} h_{pj}}{\sum_{i=Q+1}^{n} \frac{1}{m} \prod_{j=1}^{r} h_j \left( \frac{y - Y_{i,m}}{h^y} \right) \prod_{j=1}^{m} K_m \left( \frac{x - X_{i,s}}{h^x} \right) \prod_{j=1}^{k} h_{pj}}.
\]

The identification problem of model (1) is an estimation problem of the function (cf. Kitaeva and Koshkin[3], Dobrovidov et al.[1])

\[
H(A) = H(a_0, a_1) = \frac{a_1}{a_0}, \quad a_0 > 0,
\]

(4)
where \( A = (a_0, a_1), \) \( a_g = a_g(u) = \int q^g f(q, u) dq, \) \( g = 0, 1, f(q, u) = f(z), z \in R^{m+s+1}, \) is the p.d.f. of a random vector \((Y_t, U_t) = (Y_t, Y_{t,m}, X_{t,s}) = Z_t\) in stationary conditions.

In this paper, we study the mean square (MS) convergence of the estimator (3) to the function \( \Psi \) determining the ARX-process (1).

2 The principal part of the MS error for a plug-in estimator of the function \( \Psi \)

We introduce the following notation:

\[
f_1(i+1)(i+j+1)(i+j+k+1)(z, v, u, w) \text{ is the } 4(m+s+1) \text{-dimensional p.d.f. of sample vectors } Z_1, Z_{i+1}, Z_{i+j+1}, Z_{i+j+k+1}, a_{1(i+1)(i+j+k+1)}^+(z, y, z', y') = \int_{R^4} |\nu \nu' \nu'' \nu'''| f_1(i+1)(i+j+k+1)(\nu, z, v, y', z', y') d\nu d\nu' d\nu'' d\nu''',
\]

\( Q + 1 \leq i, j, k < n, \) \( i + j + k \leq n - 1; \)

\[
a_{1(i+1)(i+j+k+1)}^+(z, y, z') = \int_{R^3} |\nu \nu' \nu''| f_1(i+1)(i+j+k)(\nu, z, v, y', z', y'') d\nu d\nu' d\nu'',
\]

\[
a_{1(i+1)}(z, z') = \int_{R^2} |\nu \nu'| f_1(i+j)(\nu, v, z') d\nu d\nu', \quad a^+_p(z) = \int |\nu|^p f(\nu, z) d\nu, \quad g = 0, 1, \quad L = m + s.
\]

**Definition 1.** A function \( K(u) \) belongs to the class of one-dimensional kernels \( K(\cdot) \in A_\nu \) if

\[
\int |K(u)| du < \infty, \quad \int K(u) du = 1, \quad \int |u^p K(u)| du < \infty,
\]

\( T_j = \int u^j K(u) du = 0, \quad j = 1, \ldots, \nu - 1, \quad T_\nu \neq 0, \) and \( K(u) = K(-u). \)

Below, to study the MS convergence of estimators, we use the same bandwidth \( h_n \) for each of the \( m + s \) variables in Theorem and the product of one-dimensional kernels as multidimensional kernels of the proper dimensions.

Denote \( \omega_{\nu g}(u) = \frac{T_\nu}{\nu!} \sum_{j=1}^L \frac{\partial^\nu a_g(u)}{\partial u_j^\nu}, \quad H_g = \frac{\partial H(A)}{\partial (a_g)}, \quad \sup_{u \in R^L} H_g = \sup_{u} H_g. \)

**Definition 2.** A function \( H(\cdot) : R^L \rightarrow R^1 \) belongs to the class \( \mathcal{N}_\nu(z) \) \((H(\cdot) \in \mathcal{N}_\nu(z)) if it is continuously differentiable up to the order \( \nu \) at a point \( z \in R^L \). A function \( H(\cdot) \in \mathcal{N}_\nu(R) \) if it is continuously differentiable up to the order \( \nu \) for each \( z \in R^L \).

**Theorem.** Assume that for the function \( H(A) \) in (4) and integers \( g, p = 0, 1, \beta = 0, 1, \ldots, q = L, T_L \), the following conditions hold:

1) \((Z_t)\) satisfies the s.m. condition, \( \int_0^\infty \tau^2 |\alpha(\tau)|^\beta d\tau < \infty, \) \( 0 < \delta < \infty; \)
2) \( a_{g+p}(\cdot) \in \mathcal{N}_0(R), \) \( a_{g(p+2+\delta)}^+(\cdot) \in \mathcal{N}_0(z); \) \( \sup_{u} a_{g+p}^+(u) < \infty, \sup_{u} a_{g+p}^- u(u) < \infty; \)
3) \( K(\cdot) \in A_{\nu}, \) \( \sup_{u \in R^L} |K(u)| < \infty; \)
4) \( a_0(u) > 0, a_g(\cdot) \in \mathcal{N}_\nu(R), \sup_{u} |a_g(u)| < \infty, \sup_{u} \left| \frac{\partial^\nu a_g(u)}{\partial u_1 \partial u_2 \cdots \partial u_q} \right| < \infty; \)
5) a non-increasing sequence \((h_n)\) is such that \( (d_n) = \left( h_n + \frac{1}{nh_n^L} \right) \downarrow 0, \) as
6) sup \( u \rightarrow \infty \) \( U_{ij}^{+}(i+j+1)(i+j+k+1)g(u, u, u) < \infty \),
7) sup \( i, j, k \geq 1 \) \( U_{ij}^{+}g(2+\delta)(u, u, u) < \infty \), sup \( u \rightarrow \infty \) \( U_{ij}^{+}g(2+\delta)(u, u) < \infty \),

Then \( E[H(A_n) - H(A)]^2 = \sum_{g,p=0}^{1} H_gH_p \left[ \frac{a_{t+p}(x)}{nh_n} \left( \int K^2(u)du \right) + \omega_{\nu\nu}(z)\omega_{\nu\nu}(z)h_n^{2\nu} \right] + o \left( h_n^{2\nu} + \frac{1}{nh_n} \right) \).

Note, in this formula according to (4), \( H_0 = -\frac{a_1}{a_0} \), and \( H_1 = \frac{1}{a_0} \). The proof of Theorem is based on the results, presented in Masry\[10\], Dobrovidov et al.\[1\], Kitaeva and Koshkin\[4\], Koshkin\[5\], Koshkin and Glukhova\[6\].

3 Comparison of parametric and nonparametric algorithms

Computer modelling is started by generation sequences of dependent observations, using the following processes:
\[
M(1): Y_n = 0.2Y_{n-1} + 0.11X_n^1 + 0.15X_n^1 + 0.3X_n^2 + 0.2X_n^3 + \xi_n,
M(2): Y_n = 0.01Y_{n-1}Y_{n-2} + 0.2X_n^1 + 0.03X_n^2 - X_n^3 - 0.7X_n^3 + \xi_n,
M(3): Y_n = e^{0.1Y_{n-1} + 0.2X_n^1 + 0.1X_n^1 + 0.01X_n^2 - 0.03X_n^3} + \xi_n.
\]

Here, variables \( X^1, X^2, \) and \( X^3 \) take values from the uniform distribution laws on the corresponding intervals \([2, 2.5]\), \([5, 6]\), \([8, 10]\), and random variables \( \xi_n \) are distributed by the normal distributions with zero mean and variances that are calculated for the models \( M(1) - M(3) \) by the formula
\[
\sigma^2 = \frac{Y_{max} - Y_{min}}{6} \cdot a,
\]
where the multiplier \( a \) is a level of noise \( \xi_n \), and the parameter \( a \) takes the values 0.01, 0.05, 0.1, 0.15, 0.2, and 0.5. We use in our simulation the following sample sizes: 50, 100, 200, and 500.

Note, the condition (2) holds for the models \( M(1) \) and \( M(2) \) (see Subsection 3.5.9 in Dobrovidov et al.\[1\]).

The identification algorithms for functions \( \Psi \) in (1) were obtained by the least squares method (LSM), iterative weighted least squares method (WLSM), and nonparametric approach. The LSM and WLSM estimators are obtained by making use of the MATLAB built-in functions.

The simulation of nonparametric algorithms is also based on the MATLAB. As a kernel \( K(u) \), we use the standard Gaussian density. The bandwidths are
found in two ways. In accordance with Stone[13], Hall[2], Li and Racine[9], Leung[8], the bandwidths are calculated by the cross-validation (CV) method. The second method of finding the bandwidth uses the estimate \( \hat{Y}_n \), based on the following empirical criteria:

\[
h_{j, \text{Empiric}} = C_0 \sigma_j n^{-1/4}, \quad C_0 = \arg\min_{0 < C < \infty} \left| \frac{1}{n-1} \sum_{i=Q}^{n-1} Y_i \sum_{i=Q}^{n-1} K_L \left( \frac{U_i - U_n}{h} \right) \right|
\]

\( j = 1, L \), where \( L \) is the dimension of a function \( \Psi \), \( \sigma^2_j \) is the sample variance of observations for the \( j \)-th variable, \( \hat{h} = (\hat{h}_y, \hat{h}_x) \).

Fig. 1. The dependence of the averaged identification errors on the noise level \( a \) for the model M(1), \( n = 50 \)

Fig. 2. The dependence of the averaged identification errors on the noise level \( a \) for the model M(2), \( n = 50 \)

Fig. 3. The dependence of the averaged identification errors on the noise level \( a \) for the model M(3), \( n = 50 \)

Fig. 4. The dependence of the averaged identification errors on the sample size with the noise level \( a = 0.15 \) for the model M(1)
Fig. 5. The dependence of the averaged identification errors on the sample size with the noise level $a = 0.15$ for the model $M(2)$.

Fig. 6. The dependence of the averaged identification errors on the sample size with the noise level $a = 0.15$ for the model $M(3)$.

For the models $M(1)$–$M(3)$, the values of the identification errors

$$A_n = \frac{1}{n - Q} \sum_{i=Q+1}^{n} \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right|,$$

$n = 50$, are presented in Figures 1–3. Here $\hat{Y}_i$ are the estimates of $Y_i$, obtained by the corresponding identification algorithms for our models. The results for other sample sizes are shown in Figures 4–6. Note, all the simulation results are averaged over 20 different samples of the same size.

According to Figures 1–3, the identification quality for the models $M(1)$–$M(3)$ and the different identification algorithms decrease with increasing the noise level. There is also a tendency of reduction of identification errors for all the models by increasing sample sizes. Further, note that for the non-linear model $M(3)$ nonparametric algorithms have advantages over parametric algorithms because of their adaptability.

4 Real data processing

We will examine the dependence of Russian Federation’s Industrial Production Index (IPI) $Y$ (see Figure 7) on the dollar exchange rate $X^1$, import $X^2$, and direct investment $X^3$ for the period from September 1994 to January 2013. The data are available from the following cites: http://www.gks.ru and http://sophist.hse.ru/. Due to the changing classification principles of economic activities in 2002, we consider two series of the data from September 1994 to December 2002 and from January 2003 to January 2013. Taking into account the corresponding trends of the IPI (see Figure 7), we apply the estimator (3) under

$$U_i = (Y_{i,1}, X_{i,4}) = (Y_{i-1}, X^1_i, X^2_i, X^3_i, X^3_{i-1}).$$
Fig. 7. The IPI for the period from January 1995 to January 2013

As a kernel $K(u)$, we take the Gaussian density. The bandwidths are such that

$$h_j = 1.1\sigma_j n^{-1/9}$$

for the data from September 1994 to December 2002, and

$$h_j = 1.23\sigma_j n^{-1/9}$$

for the data from January 2003 to January 2013, $j = 1, 2, 3, 4, 5$. The constants 1.1 and 1.23 are obtained by the above empirical criteria. To compare the non-parametric algorithms (3) and LSM-estimators, we have calculated the relative errors $A_n$ and relative average annual errors $A(t)$, $t = 1994, \ldots, 2013$, for both the approaches:

$$A_n = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right|, \quad A(t) = \frac{1}{12} \sum_{i=1}^{12} \left| \frac{Y_i(t) - \hat{Y}_i(t)}{Y_i(t)} \right|.$$

Here, $Y_i$ is the true value of the IPI and $\hat{Y}_i$ is either parametric or nonparametric estimate. The results of such a comparison are given in Figures 8 and 9.

The increase in the errors for 1998 and 2009 can be explained by Russian financial crisis ("Ruble crisis") in August 1998 and Global financial crisis in 2009.

To predict the IPI $Y$ for the data from 2002 to 2013 (cf. Simakhin[12]), we apply (3) under

$$U_i = (Y_{i,1}, X_{i,4}) = (Y_{i-1}, X_{i-1}^1, X_{i-1}^2, X_{i-1}^3, X_{i-2}^3).$$

Here, the bandwidths are equal to

$$h_{jt} = 0.94\sigma_j t^{-1/9},$$

where $\sigma_j^2$ are the sample variances of observations for the corresponding variables $X_j^j$, $j = 1, 2, 3, 4, 5$. 

381
The similarity of identification algorithms and forecasting algorithms leads one to expect the both should behave similarly. For the relative average annual errors $A(t)$, seen in Figures 9 and 10, one indeed observed that.

5 Control

Let the outputs $Y$ of models M(1) and M(3) should be levelled to the given values $Y^*(k), k = 1, \ldots, 5$ (see Figures 11 and 12) by making use of the controlled inputs $X^2$ and $X^3$ for the corresponding models.

For instance, outputs can be some qualities of produced goods and inputs can be certain parameters of the production technology. Note, we can take the direct investments as the controlled input in Section 4.

There is a problem of finding $X^2_*, X^3_*$ for M(1), M(3), and $X^3_*$ for the case of the real data in Section 4. The variables $X^1$, $X^2$, and $X^3$ take values from the uniform distributions on the corresponding intervals $[2,2.1]$, $[0,2]$, and $[0,0.2]$ for M(1), and $[2,2.1]$, $[0,0.2]$, and $[0,2]$ for M(3). Random variables $\xi_n$ were generated from the normal distributions with zero mean and variances, calculated by the formula (5).
The wanted outputs \( Y_*(k) \) were taken as 
\[
Y_*(1) = \min\{Y_3, \ldots, Y_{50}\}, \quad Y_*(2) = \frac{1}{48} \sum_{i=3}^{50} Y_i, \quad Y_*(3) = \max\{Y_3, \ldots, Y_{50}\}, \quad Y_*(4) = (Y_*(1) + Y_*(2))/2, \\
Y_*(5) = (Y_*(2) + Y_*(3))/2.
\]

Using the given value \( Y_*(1n) = Y_*(1) \), \( n = 51, \ldots, 70 \), on the base of the algorithm (3) as in Koshkin and Tarasenko\cite{7}, for the model M(1) one can construct the estimates of \( \hat{X}_2^*(1n) \) from a sequence of \( (Y_i, Y_{i-1}, X_{i}^1, X_{i-1}^1, X_{i}^2, X_{i-2}^3) \), 
\[
\hat{X}_2^*(1n) = \sum_{i=3}^{n} X_i^2 K_2 \left( \frac{Y_{i,2}(1n) - Y_{i,2}}{h_y} \right) K_2 \left( \frac{X_{i,2}^1 - X_{i,2}^1}{h_{1x}} \right) K \left( \frac{X_{i,2}^3 - X_{i,2}^3}{h_{33}} \right) \\
\sum_{i=3}^{n} K_2 \left( \frac{Y_{i,2}(1n) - Y_{i,2}}{h_y} \right) K_2 \left( \frac{X_{i,2}^1 - X_{i,2}^1}{h_{1x}} \right) K \left( \frac{X_{i,2}^3 - X_{i,2}^3}{h_{33}} \right), 
\]

where \( Y_{i,2}(1n) = (Y_*(1n), Y_{n-1}), \quad Y_{i,2} = (Y_i, Y_{i-1}), \quad X_{i,2}^1 = (X_i^1, X_{i-1}^1), \quad h_y = (h_1, h_2), \quad h_{1x} = (h_{11}, h_{12}).
\]

Fig. 11. The results of control for the model M(1)  
Fig. 12. The results of control for the model M(3)

| Models | M(1) | M(3) |
|--------|------|------|
| A(1)   | 0.024| 0.021|
| A(2)   | 0.014| 0.012|
| A(3)   | 0.024| 0.026|
| A(4)   | 0.011| 0.009|
| A(5)   | 0.020| 0.015|

Table 1. The relative errors
Here, we use the Gaussian kernel and bandwidths founded on the base of the empirical criteria (see Section 3). The relative errors are defined as

$$A(k) = \frac{1}{20} \sum_{i=51+(k-1)20}^{70+(k-1)20} \left| \frac{\hat{Y}_i - Y_\ast(k)}{Y_\ast(k)} \right|, \quad k = 1, \ldots, 5,$$

where, for example, \( \hat{Y}_i = 0.2Y_{i-1} + 0.11X_1^i + 0.15X_{i-1}^1 + 0.3X_2^2(1i) + 0.2X_3^3, \)

\( i = 51, \ldots, 150. \) The results are also obtained for the model M(3) (see Table 1).

References

1. A.V. Dobrovidov, G.M. Koshkin and V.A. Vasiliev. Non-Parametric State Space Models. Kendrick Press, Inc., Heber, UT 84032, USA, 2012.
2. P. Hall. Asymptotic properties of integrated square error and cross-validation for kernel estimation of a regression function. Z. Wahrscheinlichkeitstheorie Verw. Geb. 67, 175–196, 1984.
3. A.V. Kitaeva and G.M. Koshkin. Semi-recursive nonparametric identification in the general sense of a nonlinear heteroscedastic autoregression. Automation and Remote Control, 2, 92–111, 2010.
4. A.V. Kitaeva and G.M. Koshkin. Nonparametric semirecursive identification in a wide sense of strong mixing processes. Problems of Information Transmission, 46, 1, 22–37, 2010.
5. G.M. Koshkin. Deviation moments estimates of substitution and its piecewise-smooth approximations. Siberian Mathematical Journal, 40, 3, 605–618, 1999.
6. G.M. Koshkin and I.Yu. Glukhova. Nonparametric identification of nonlinear ARX-processes. Journal of Control and Computer Science. Tomsk State University, 3, 20, 55–61, 2012 (in Russian).
7. G.M. Koshkin and F.P. Tarasenko. Nonparametric algorithms for identifying and control of continuous-discrete stochastic objects. 8-th IFAC-IFORS Symposium on Identification and System Parameter Estimation. Beijing: Pergamon Press, 2, 882–887, 1988.
8. D. Leung. Cross-validation in nonparametric regression with outliers. Annals of Statistics, 33, 2291–2310, 2005.
9. Q. Li and J.S. Racine. Cross-validated local linear nonparametric regression. Statistica Sinica, 14, 485–512, 2004.
10. E. Masry. Probability density estimation from sampled data. IEEE Trans. Inf. Theory, IT–29, 5, 696–709, 1983.
11. E. Masry and D. Tjostheim. Nonparametric estimation and identification of nonlinear ARCH time series Econometric Theory, 11, 2, 258–289, 1995.
12. V.A. Simakhin. Adaptive robust nonparametric prediction algorithms. Journal of Control and Computer Science. Tomsk State University, 1, 14, 45–54, 2011 (in Russian).
13. C.J. Stone. Cross-validatory choice and assessment of statistical predictions (with discussion). Journal of Royal Statistical Society, 36, 111–147, 1974.
Optimal turbines layout in an offshore wind farm using evolutionary computation

D. Gallo-Marazuela, S. Salcedo-Sanz¹, A. Pastor-Sánchez¹, E. Alexandre¹, and A. Portilla-Figueras¹

Department of Signal Processing and Communications, Universidad de Alcalá, 28805 Alcalá de Henares, Madrid, Spain. (E-mail: sancho.salcedo@uah.es)

Abstract. The placement of the whole set of turbines of an offshore wind farm in northern Europe is tackled in this paper by means of an evolutionary algorithm. We consider a new model where the shape of the wind farm is taken into account, and a new model of wake is applied using an existing simulator called OpenWind. The proposed evolutionary algorithm works by choosing the places where the turbines must be installed (wind farm layout), in such a way that the energy produced by the wind farm is maximized. The results obtained have shown that the evolutionary approach is able to obtain excellent layouts, that maximize the power generation offered by the wind farm.

Keywords: Offshore wind farm design, optimal layouts, Evolutionary Computation.

1 Introduction

Wind energy is mainly produced in large production facilities known as wind farms. The number of new wind energy production facilities in the world has grown over 25% each year, consistently in the last five, and this growth is even larger in countries such as China [1] or the USA [2]. There are two types of wind farms, i.e., facilities located in land (onshore), and in the sea (offshore). Nowadays, onshore wind farms are more extended, though in the last few years the offshore counterpart is gaining terrain. Nonetheless, offshore wind farms are known to be more productive, and many companies are betting for this kind of facility when geographical conditions allow its installation. Moreover, recent studies have reported a significant increment in the installation of offshore wind farms over 30% with respect of previous years [2]. In Europe, these facilities have been installed for more than 20 years ago, and nowadays represent a significant part of wind energy production in countries such as Denmark, Sweden or the Netherlands [3]. In addition, other studies have shown that the potential of offshore wind energy in important economies such as China [1] or Europe [4] is much larger than its onshore counterpart. The main advantages of offshore wind farms are the availability of huge continuous areas for developing major projects, the higher wind speeds at the sea, less effects of turbulence or the elimination of visual impact and noise issues, among others. On the other hand, their main disadvantages are more expensive installation and connection to the electrical network, and limited access for maintenance operations, etc.

Usually, works on automatic layout optimization [5]-[14] propose a given new algorithm to tackle the problem, or some modification to existing ap-
approaches, tested on benchmark cases, with some assumptions to make the problem as real as possible. Such as approach is interesting, since it allows controlled simulations and experiments that provide a clear image of the proposed technique’s performance. On the other hand, experiments in real cases are complicated to be obtained, since they involve a complicated process that includes previous measures in-situ, and the test of different options and possibilities for the layout (some of them finally discarded), budget evaluation, etc. This paper tries to show part of the real process of an offshore wind farm design in northern Europe, that was carried out by using evolutionary computation algorithms as support automatic tool for establishing the initial layout. Given the wind farm shape, turbine model installed and after collecting wind data in-situ, a public available tool to simulate the wind farm production was used (Openwind [15]). This tool is able to obtain the average Energy Annual Production of a turbines layout, so it could be used as a black box tool to evaluate layouts. Then, studies on different possible structures for the wind farm design are carried out: first the best regular distribution of wind turbines embedded in the wind farm shape is studied, using an evolutionary algorithm. In this case, the algorithm tries to locate the best parameters of the grid, and it is assumed that all grid points within the wind farm shape will be used to install a wind turbine. Therefore, in this case there is not a control on the exact number of wind turbines finally installed. A second problem is also discussed and solved using a different evolutionary algorithm, i.e. the free design of the wind farm with fixed number of turbines N. In this case, the grid points indicate possible installation points for wind turbines, but only N of the points will be finally selected by the algorithm to install wind turbines. In the paper we discuss the results obtained in the real offshore wind farm considered, and the advantages and disadvantages of both strategies when applied to the real case.

The rest of the paper is organized as follows: Section 2 presents the two wind farm design strategies and associated problems that will be tackled with evolutionary algorithms. Section 3 discuses details of the evolutionary algorithms used to solve the problems previously stated. Discussion and analysis of the obtained results in a real offshore wind farm in northern Europe is done in Section 4. Finally, several conclusions of this research are outlined in Section 5.

2 Problem’s statement

In this paper we consider an offshore wind farm layout consisting on a regular embedded turbines design. The associated optimization problem can be defined as follows: Let us consider an offshore wind farm, with a given shape Ω. In this first problem definition we consider wind turbines dispositations in the wind farm in the form of a regular grid Ξ, embedded into Ω. In order to define the final layout in the wind farm, we first set the wind turbines into a square form S, with a given distance in rows (dr) and columns (dc), integer numbers, in metres. We also consider the possibility of a given offset in rows, given by an offset parameter do ∈ [0, 1], where 0 means that no offset is applied, and 1 stands for the maximum offset allowed. Finally, we define an orientation angle
for the turbines’ disposition \((\alpha \in \{0, 360^\circ\})\). Figure 1 shows an example of a regular disposition in a square, with the definition of the distances and angle involved.

Mathematically, the problem consists of, given a shape \(\Omega\), obtaining the vector of variables \(d_r, d_c, d_o, \alpha\) that defines \(S\), in such a way that the regular embedded layout obtained \(\Omega(S)\) maximizes the following objective function:

\[
f(\Omega(S)) = P, \tag{1}
\]

where \(P\) stands for the average Annual Energy Production (AEP) obtained from the wind farm.

3 Evolutionary computation algorithms as an automatic layout optimization tool

Evolutionary Algorithms (EAs) are robust problems’ solving techniques based on natural evolution processes. They are population-based techniques that codify a set of possible solutions to the problem, and evolve it through the application of the so called evolution operators [16]. Several points must be taken into consideration in the description of an EA. First, the problem encoding and objective function of the problem are basic pieces of an EA. The encoding of the problem is an important point, since it defines the search space in which the EA operates. The evolutionary operators, that make the population to evolve towards good quality solution, are other key points in an EA.

Figure 2 shows the structure of the EA proposed to solve this problem of offshore wind farm optimization. Basically, after the initialization of a population of \(\mu\) individuals at random, the different operators are applied in a loop fashion: first, an offspring population of the same size that the initial one is generated by means of the crossover operator, and two mutation mechanisms are then applied to generate diversity. A tournament selection operator [16] selects the \(\mu\) parents of the next generation among the complete population of parents and offspring of the current one, and the process continues until a
maximum number of generations is completed. In this case, a classical one-point crossover operator is used in this paper for both evolutionary algorithms considered. First, we form couples of parents at random, and then the operator is applied, to form two new offspring individuals. A Gaussian-based mutation is then applied with low probability. After the application of the mutation operator, the distances are rounded to be integer numbers (measured in meters), the offset parameter \( d_o \) is adjusted to be in \([0, 1] \) and the angle \( \alpha \in [0, 360^\circ] \).

### 4 Experimental part

In this section we tackle the optimization of a real offshore wind farm, located at northern Europe, considering the regular embedded strategy considered, by using the evolutionary algorithm introduced before. Figure 3 shows the wind farm location and its shape \( \Omega \). The Bonus 1 MW wind turbine, with a diameter \( D = 54\) m is considered as the model to be installed in the wind farm. The AEP value for the optimization algorithm is provided by the well-known Open Wind software ([15], freely available). Open Wind provides efficient wakes calculation and wind farm production estimation given the wind and terrain characteristics of the wind farm, and the turbine model.

Figure 4 shows the solution obtained just optimizing the AEP of the wind farm. Note how the number of wind turbines is large, and in this case the cost of the installation would increase exponentially and the wake effects among turbines are huge, so the energy production is affected. Note that, considering this optimization strategy, the designer is not able to control the number of turbines in each layout obtained. We therefore tackle a different strategy consisting in fixing the number of turbines within a given grid, embedded in the wind farm shape. A slightly different evolutionary algorithm must be considered, in such a way that a minimum distance of 3D is maintained among turbines. For this problem, Figure 5 shows the initial grid where turbines can be located. Figure 6 shows the best solution found by the algorithm with 20 turbines, and a total AEP of 84.29 GWh.
Fig. 3. Situation (Northern Europe) and form of the real offshore wind farm designed in this work.

Fig. 4. Best solution (regular embedded design) obtained with the evolutionary algorithm.

5 Conclusions

In this paper we have proposed a study on offshore wind farm design, considering layouts embedded in the shape of a real wind farm. We have applied an evolutionary algorithm to this end, optimizing a real offshore wind farm situated in northern Europe. A first design considering regular layouts has been described, and a small modification to control the number of turbines has also
Fig. 5. Possible location points in the free design problem.

Fig. 6. Best solution found by the evolutionary algorithm with $N = 20$ turbines.

been discussed. Obtained results have shown the goodness of the proposed approach.

Acknowledgements

This work has been partially supported by Spanish Ministry of Economy, under project number ECO2010-22065-C03-02.

References

1. X. Sun, D. Huang and G. Wu, “The current state of offshore wind energy technology development,” Energy, vol. 41, pp. 298-312, 2012.
2. A. Madariaga, I. Martinez de la Alegria, J. L. Martín, P. Eguía and S. Ceballos, “Current facts about offshore wind farms,” Renewable and Sustainable Energy Reviews, vol. 16, no. 5, pp. 3105-3116, 2012.
3. M. Bilgili, A. Yasar and E. Simsek, “Offshore wind power in Europe and its comparison with onshore counterpart,” Renewable and Sustainable Energy Reviews, vol. 15, pp. 905-915, 2011.
4. M. D. Esteban, J. J. Diez, J. S. López and V. Negro, “Why offshore wind energy?” *Renewable Energy*, vol. 36, pp. 444-450, 2011.
5. G. Mosetti, C. Poloni and B. Diviacco, “Optimization of wind turbine positioning in large wind farms by means of a genetic algorithm,” *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 51, no. 1, pp. 105-116, 1994.
6. G. Marmidis, S. Lazarou and E. Pyrgioti, “Optimal placement of wind turbines in a wind park using Monte Carlo simulation,” *Renewable Energy*, vol. 33, no. 7, pp. 1455-1460, 2008.
7. S. A. Grady, M. Y. Hussaini and M. M. Abdullah, “Placement of wind turbines using genetic algorithms,” *Renewable Energy*, vol. 30, no. 2, pp. 259-270, 2005.
8. A. Emami and P. Noghreh, “New approach on optimization in placement of wind turbines within wind farm by genetic algorithms,” *Renewable Energy*, in press, 2010.
9. B. Saavedra-Moreno, S. Salcedo-Sanz, A. Paniagua-Tineo, L. Prieto and A. Portilla-Figueras, “Seeding evolutionary algorithms with heuristics for optimal wind turbines positioning in wind farms,” *Renewable Energy*, vol. 36, pp. 2838-2844, 2011.
10. A. Kusiak and Z. Song, “Design of wind farm layout for maximum wind energy capture,” *Renewable Energy*, vol. 35, pp. 685-694, 2010.
11. M. A. Lackner and C. N. Elkinton, “An analytical framework for offshore wind farm layout optimization,” *Wind Engineering*, vol. 31, no. 1, pp. 17-31, 2007.
12. C. N. Elkinton, J. F. Manwell and J. G. McGowan, “Algorithms for offshore wind farm layout optimization,” *Wind Engineering*, vol. 32, no. 1, pp. 67-84, 2008.
13. R. A. Rivas, J. Clausen, K. S. Hansen, and L. E. Jensen, “Solving the turbine positioning problem for large offshore wind farms by simulated annealing,” *Wind Engineering*, vol. 33, no. 3, pp. 287-297, 2009.
14. B. Pérez, R. Mínguez and R. Guanche, “Offshore wind farm layout optimization using mathematical programming techniques,” *Renewable Energy*, vol. 53, pp. 389-399, 2013.
15. http://www.awsopenwind.org/
16. A. E. Eiben and J. E. Smith, “Introduction to evolutionary computing,” Springer-Verlag, 2003.
Construction of a happiness index using polytomous item response theory models in a survey

Diana Carolina Gamboa and Alvaro Mauricio Montenegro

1 Universidad Nacional de Colombia, Sede Bogotá, Departamento de Estadística (e-mail: dcgamboap.edu.co)
2 Universidad Nacional de Colombia, Sede Bogotá, Departamento de Estadística (e-mail: ammontenegrod@unal.edu.co)

Abstract. In this paper, we present an alternative to derive univariate indices by using polytomous item response theory models in data from surveys. Particularly, the Samejima’s graded response model [12] was used. A real data set from the Social European Survey database was fitted. In order to interpret the index a multiple correspondence analysis and a clustering was performed.

Keywords: Univariate index, polytomous item response theory, the European Social Survey, multiple correspondence analysis.

1 Introduction

Latent traits, are random variables which cannot be observed directly, Baker and Kim [1]. Some known examples of latent traits are intelligence, feelings and preferences. In this paper we work with happiness. To measure a latent trait, a test is designed to obtain observable data which are related with it. Each question in those tests is called an item. The item response theory (IRT) models have been developed for estimating continuous latent traits from categorial data. IRT models are really factor analysis of discretized variables, Takane and Leeuw [13]. In the general case, latent traits are multidimensional, Reckase [10]. However, when a latent trait is approximately unidimensional, unidimensional (UIRT) models may be used, Montenegro and Cepeda [7]. In this case, a unidimensional scale is adopted for the latent trait. The key idea when we use UIRT models is that the categories of the items and the latent traits are in the same scale. Consequently, if the items in a test are ordinal, say by using a Likert scale, the categories and the latent traits are sorted in the real line such that people with low latent trait have high probability of choosing low values of Likert scale and vice versa.

Two general types of IRT models are studied separately, even though one of them is a subclass of the other. To the first class belong the dichotomous or dichotomized models. In this case, all the variables in a test a dichotomic. To the second class belong the models where the items are categorial, and where each item may have more than two categories. This kind of models are more adequate for using in surveys. The members of the second class are called polytomous IRT models. In this paper, we focus on the graded response model.
(GRM) developed by Samejima [12]. This is a polytomous IRT model which performs with rating scale response and frequently is used in attitude test, assessing student performance and marketing studies. For a general review of polytomous IRT models see for example Ostini and Nering [8].

Since the main objective of UIRT models is to state a scale and sort the people according with such scale, the estimated latent traits are synthetic indices, Montenegro and Cepeda [7]. Some times, principal component analysis (PCA) or multiple correspondence analysis (MCA), Lebart [4], are used to obtain synthetic indices. The results are relatively similar, because all of these are factorial methods. In this paper we use this fact, and apply MCA as an exploratory tool in order to help in the interpretation of the index.

In surveys, the most common type of variables are categorical, and more specifically ordinal. In these cases, Likert scales are used. The data used along this paper are from the European Social Survey (the ESS). The ESS, is a biennial multi-country survey covering over 30 nations. The first round was fielded in 2002/2003, the fifth in 2010/2011. The project is funded jointly by the European Commission, the European Science Foundation and academic funding bodies in each participating country. For details about the EES go to http://www.europeansocialsurvey.org/.

For the construction of the index we used the human values section of the Spanish sample. Some of the statements of the survey that we used were: I am always optimistic about my future; my life is exactly as I want it to be; I like to plan and prepare for the future; I feel I can decide for myself how to live; All the time I feel that I will fail. The common response categories for this kind of statements in the ESS are: strongly disagree, disagree, neither agree nor disagree, agree, strongly agree. The complete data and the R procedures used in this work can be asked to the authors, via e-mail.

2 The graded response model

The GRM by Samejima [12] is based on the successive boundaries procedure proposed by Thurstone. The ordered response categories are viewed as regions of a continuum separated by boundaries, Masters [5].

The categories are ordered and denoted by $g$, where $g = 1, 2, \ldots, m$. The probability of person $j$ responding in or above $g$ to item $i$ is specified as

$$P^{*}_{ig} = \frac{e^{[a_i(\theta_j) - b_{ig}]}}{1 + e^{[a_i(\theta_j) - b_{ig}]}}$$

(1)

Where $a_i$ is called the discrimination parameter of item $i$ and $b_{ig}$ is the boundary location parameter which corresponds to the boundary between categories $g$ and $g - 1$ associated with item $i$, Ostini and Nering[8].
The probability $P_{ij}$ of person $j$ responding in category $g$ is given by

$$P_{ig} = P^*_i - P^*_{i(g+1)}$$

(2)

To completely define the model in terms of the available boundaries, two additional definitions are required.

- The probability of person $j$ responding in or above the lowest category to item $i$: $P_{i0} = 1$.
- The probability of person $j$ responding in or above the highest category to item $i$: $P_{i(m+1)} = 0$.

On the other hand, the category response function is given by $P_{ig}(	heta)$, i.e., $P_{ig}(	heta)$ is the probability of choosing category $i$ to item $g$ as a function of $\theta$. The figure 2 shows the graphs of the response functions of an item.

3 Data description

The data belongs to the ESS, applied in 2006. The test includes a wide range of social variables, social and public trust; social exclusion, national, ethnic and religious allegiances; well-being; health and security; human values; demographics and socio-economics. For the construction of the index we used the human values section of the Spanish sample. The sample size was $N=1876$ and the test size was $K=27$. In the selected subtest the items 5 to 19 have 4 categories. On the other hand, the other items have 5 categories.

There were 218 missing data in the test. This is not a problem with IRT models. For calibration, only the non missing responses are used in the likelihood function. However, for using MCA it is necessary to complete the data. Thus, we first fitted the data using the graded Samejima’s model. Item calibration was used to compute the latent trait estimation. Once the latent trait estimations were obtained, the missing data was estimated item to item by categories with highest probability.

4 Assessing the unidimensionality assumption

There exist different procedures for assessing the unidimensionality of the data. We use a recent procedure based on the Cronbach’s $\alpha$ (alpha) coefficient. This coefficient is commonly used to measure the internal consistency or reliability of a test. The coefficient is defined as

$$\alpha = \left( \frac{k}{k-1} \right) \left( 1 - \frac{\sum S_i^2}{S_{total}^2} \right),$$

where $k$ is the number of items, $S_i^2$ is the variability of item $i$ and $S_{total}^2$ is the total variability of the test. The coefficient may be interpreted as follows.
Cronbach’s alpha will generally increase as the intercorrelations among test items increase. Large values of $\alpha$ implies high reliability and consequently low error. A second interpretation useful for unidimensionality assessing is given by other authors, see for example Cortina[3]. According to those authors, alpha is said to be measure of first-factor saturation (i.e. the extend to which a certain factor is present in all items. Under the unidimensionality assumption, the alpha coefficient increases if we omit an item. When the coefficient decreases, the item is pointing towards a different direction of the test. In other words, the assumption of unidimensionality fails. Based on this idea, the Cronbach-Mesbah curve was designed. For details see Mesbah [6], Cameletti [2]). The key idea is that those items which decreases the curve when they are omitted, are problematic and must be omitted to assure the unidimensionality assumption. According to the curve in figure 1 we decided to omit the items 23, 22, 25, 21, 16, 27. The curve was obtained, using the CMC procedure of R [9].

![Fig. 1. The Cronbach- Mesbah curve](image)

5 Fitting the data

To fit the data we use the marginal likelihood method implemented in the package ltm of R. To define a scale, it is assumed that the latent traits $\theta_j$ of the people are a sample from a random variable $\Theta \sim N(0,1)$. In this method, the likelihood is integrated with respect to the latent trait to calibrate the item parameters. The scoring for the latent traits is running using the posterior expectation of the latent traits. For details, see Rizopoulos [11].

The gradual response model provides information about the discrimination power for each item $a_i$ and the location parameters $b_{ig}$. Estimations of some item parameters are showed in the table 1. The category response functions are showed in the figure 2 for the item 14. In this case, the discrimination parameter is $a= 1.51$. For this item, the people with trait level less than -3.8 have more chance to select the first category; people with latent trait values
between -2.5 and -0.8 have higher probability for selecting the second category response; finally people with trait level greater than -0.8 will have high chance to answer the last category.

| Item | $b_{i1}$ | $b_{i2}$ | $b_{i3}$ | $a_i$ |
|------|----------|----------|----------|-------|
| 5    | -2.877   | -1.845   | -0.214   | 2.52  |
| 6    | -2.774   | -1.689   | 0.084    | 1.873 |
| 7    | -2.93    | -1.742   | -0.047   | 1.384 |
| 8    | -2.694   | -0.856   | 0.773    | 1.872 |
| 9    | -3.054   | -2.142   | -0.654   | 1.606 |
| 10   | -2.464   | -0.567   | 1.224    | 1.594 |
| 11   | -2.771   | -1.74    | 0.113    | 2.442 |
| 12   | -2.772   | -1.738   | -0.043   | 2.447 |
| 13   | -1.261   | 0.442    | 2.266    | 1.411 |
| 14   | -3.791   | -2.427   | -0.66    | 1.51  |
| 15   | -3.136   | -1.454   | 1.35     | 1.219 |
| 16   | -2.403   | -0.184   | 1.751    | 1.299 |
| 17   | -4.752   | -2.91    | -0.439   | 1.133 |
| 18   | -1.807   | -0.215   | 1.535    | 1.15  |

Table 1. Estimates of some item parameters

![Item Response Category Characteristic Curves - Item: V14](image)

Fig. 2. Response Functions for the Item 14

Once the item parameters were calibrated, we estimated the latent traits $\theta_j$ for the people. We named this latent trait as people’s happiness.

6 Goodness of fit

In this work, we used a goodness of fit test, to identify problematic items (ambiguous or flawed) and to detect people with responses that not follow
the general pattern responses.[14]. Given a person \( j \) with latent trait \( \theta_j \), the probability to obtain the \( x_{ij} \) score is denoted by \( P_{ijk} \). The expected value and variance of \( x_{ij} \) are,

\[
E_{ij} \equiv E(x_{ij}) = \sum_{k=0}^{m_i} kP_{ijk} \tag{3}
\]

and

\[
W_{ij} \equiv V(x_{ij}) = \sum_{k=0}^{m_i} (k - E_{ij})^2P_{ijk} \tag{4}
\]

The standardized differences between score of the person \( j \) and its expected response for the item \( i \) is given by

\[
z_{ij} = \frac{(x_{ij} - E_{ij})}{\sqrt{W_{ij}}} \tag{5}
\]

The standardized residuals were computed based on the equations 3, 4 y 5. We detected that the number of residuals such \(|z_{ij}| > 2\) was 1556, that correspond to 3.9% of the responses of the responses. As this percent was less than 5%, there is no reason to worry.

7 Multiple correspondence analysis

A multiple correspondence analysis was done with the test responses as active variables. The latent trait was projected as an illustrative variable. Figure 3 shows the first factorial map. We observed a parabolic structure (Guttman Effect). This effect revealed that the first factor summarizes the order structure of modalities; remaining factors are functions of the first factor, Lebart [4]. The first factor shows opposition among extreme categories. The second factor shows opposition among extreme categories and average categories. The latent trait variable has a high correlation with the first factorial axis (0.96) therefore both variables measure essentially the same.

8 Clustering

A clustering analysis was ran. Four clusters were obtained. Cluster 1 had the lowest happiness index. To this cluster belongs 7.26% of the sample. The cluster was characterized by people that the most of the time felt depressed, sad, anxious and bore; they never felt energetic or enjoyed their life and they never felt peaceful or calm; they always felt really tired when they woke up in the morning; they felt that their life was far from how they would like it to be. They had bad health, they were widowed or divorced; they usually did housework, they were looking after children or other people almost all the time and they rarely met with their friends.
Cluster 2 had a low happiness index. To this cluster belongs 33.96% of the sample. This cluster was characterized by people that some of the time felt depressed, sad and anxious; most of the time they did not feel calm, they did not enjoy their life, they never felt really rested when they woke up in the morning; For them, when things were wrong in their life, usually it took a long time to get back to normal. Their health was not too good, they did not finish their primary education and they did not have not a close friend to hear their problems.

Cluster 3 had a high happiness index. To this cluster belongs 40.90% of the sample. This cluster was characterized by people who never felt sad or depressed or anxious; they were positive about their future, they felt positive about themselves and their life was close to how they would to like it to be. They have peaceful sleep and they hardly ever felt alone; their highest level of education is a university degree, and they had good health.

Cluster 4 had the highest happiness index. To this cluster belongs 17.89% of the sample. The cluster was characterized by people very hopeful about their future, they had a high level of self-confidence, and they felt that their life was close they would like it to be. They felt free to decide about their life, and most days they feel a sense of accomplishment about they do; they always felt a lot of energetic, peaceful and calm. They had very good health, everyday they met with their friends and they were Catholics.

9 Conclusion

In this paper, it was proposed a methodology to build a happiness index by using a polytomous item response theory. This is a synthetic index based on qualitative information. In contrast to the common methods to construct indexes, the happiness index is not a linear combination of other indexes. Instead, the happiness index is a latent trait. We illustrated, how polytomous IRT models may be useful for analyzing data from surveys. We show the main issues that the researchers most consider before using the models. In other areas like marketing, the ideas are useful if the researcher understand the models well and the assumptions about them.
References

1. Baker, F. B. and Kim Seok-Ho (2004), *Item Response Theory*, second edition, Marcel Decker Inc.
2. Cameletti, M & CAviezel, V. (2012), The cronbach-mesbah curve for assessing the unidimensionality of an item set: The r package cmc.
3. Cortina, J. (1993), 'What is Coefficient Alpha? An examination of Theory and Applications', *Journal of Applied Psychology* 78 (1), 98-104.
4. Lebart, L., M. A. & Piron, M. (1995), *Statistique Exploratoire Multidimensionnelle*, Dunod.
5. Masters, G. (1982), 'A Rasch model for partial credit scoring', *Applied Measurement in Education*, 1, 279-297.
6. Mesbah, M. (2010), 'Statistical Quality of Life', In N. Balakrishnan (ed.), *Method and Applications of Statistics in the Life and Health Sciences*, Wiley, 839-864.
7. Montenegro, A. and Cepeda, E., 'Synthesizing the Ability in Multidimensional Item Response Theory Models', *Revista Colombiana de Estadística* 33, 127-147.
8. Ostini, R. & Nering, M. (2006), *polytomous item response models*, SAGE publications.
9. R Development Core Team (2012), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL: http://www.R-project.org
10. Reckase, M., (2009), *Multidimensional Item Response Theory*, Springer
11. Rizopoulos, D. (2006), 'ltm: An R package for latent variable modeling and item response theory analyses', *Journal of Statistical Software* 17, 1-25.
12. Samejima, F (1969), Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, Monograph Supplement No. 17.
13. Takane, Y. and Leeuw, J. (1987), ‘On the relationship between item response theory and factor analysis of discretized variables’, *Psychometrika* 52(3), 393-408.
14. Van der Linden, W. & Hambleton, R. (1998), *Handbook of modern item response theory*, Springer.
On the Application of an Evolutionary Algorithm for the Optimization of Kitting Areas in Car Manufacturing Production Chains

C. A. Garcia-Santiago\textsuperscript{1}, A. Gonzalez-Gonzalez\textsuperscript{1}, I. Landa-Torres\textsuperscript{1}, J. Del Ser\textsuperscript{1}, S. Gil-Lopez\textsuperscript{1}, I. Diaz-Iriberri\textsuperscript{2}, F. Duran-Limon\textsuperscript{2}, and S. Salcedo-Sanz\textsuperscript{3}

\textsuperscript{1} TECNALIA. OPTIMA Unit, E-48160 Derio, Spain
(E-mail: carlosalberto.garcia@tecnalia.com)
\textsuperscript{2} Nissan Motor Iberica S.A. 08040 Barcelona, Spain
(E-mail: isabel.diaziriberri@nmisa.com)
\textsuperscript{3} Universidad de Alcala, 28871 Alcala de Henares, Spain
(E-mail: sancho.salcedo@uah.es)

Abstract. This research work gravitates on the use of evolutionary meta-heuristics as an efficient computational means to improve the saturation level in car manufacturing production chains, where material is prepared and buffered in Auto Guided Vehicles (AGVs). A thorough mathematically modelling of the problem is presented, whose complexity motivates the use of computationally-efficient optimization methods such as the herein proposed evolutionary meta-heuristics. Additional side methods are also incorporated to the main evolutionary solver so as to account for operational constraints such as the number of sub-paths to be allowed for the AGV in its path through the kitting area, or the maximum number of parts in any of the installed shelves. The paper also discusses several numerical experiments aimed at verifying the saturation decrease rendered by the proposed algorithm with respect to strategies where the picking frequency of parts is adopted as a criterion for their deployment in the shelves.

Keywords: Optimization, automotive, kitting, operator saturation, genetic algorithm.

1 Introduction

Recently a flurry of research has been invested towards assessing and reducing the work saturation of operators in massive car production plants. Saturation not only incurs severe drawbacks in the effective productivity of the personnel at hand, but may also cause risky situations due to the reduced awareness of the operator with respect to the surrounding machinery. In order to prevent such risks, and to improve the operators work conditions, a maximum saturation level (measured in % with respect to the daily working schedule of the operator) is defined. For this reason, the companies in the sector are obliged to improve their processes as much as possible in order to ensure an optimum saturation level yielding, as a direct consequence, that their resources must be accurately allocated. Based on this rationale, the adoption of innovative approaches for the cost-less optimization of operational processes on-site is deemed of utmost importance for the viability of the manufacturing company at hand.
In a company of the car manufacturing sector, the proper measurement and management of times is the key factor for modelling and designing logistics flows, processes and areas. A correct distribution of the workload in any team, process, area, etc. provides the best results with minimum resources while guaranteeing good working conditions for the operators. In this context, the saturation of one operator measures the working time (which involves all displacements and process times) against the total duration of the labour day. This concept has become one of the most important metrics to care about by the industry: due to the high competitiveness in the sector, the companies are obliged to improve as much as possible their processes in order to achieve an optimum saturation level with their resources accurately designated. There are many physical and human related factors to be considered for a proper calculation of the saturation level. In what relates to human factors, a maximum % of saturation is established in order to avoid any dangerous situation in the operators’ working conditions.

Interestingly for the scope of this work, modern car manufacturing chains require dynamic support systems in order to cope with the heterogeneity of parts and pieces of different product models. One of such machinery is the so-called Automated Guided Vehicle (AGV), which essentially consists of a moving trolley looping through a kitting area which serves as a buffer for the operator to pick and deploy the necessary parts to the assembly line. As shown in Figure 1, the kitting area is further composed by shelves of different capacity and size, which are set on specific reserved slots or PDLs (Point of Delivery at Line) containing different types of parts. Once the parts have been assigned and deployed in the shelves, the AGV is filled with items as a function of the model to be produced at a certain time. Interestingly for the scope of this work, the model series to be produced is reflected in the production forecast, which is usually fixed for a long term period. In this scenario, saturation comes from the fact that the operator in the kitting area must pick and deploy parts in the AGV before the trolley finishes the path, an intensive activity no matter if the production chain is producing at its maximum rate.

Fig. 1. Diagram depicting a typical car manufacturing chain with kitting areas.
In this paper we focus on a particular use case located at NISSAN car manufacturing facilities, where the optimization of the saturation is made over existing kitting areas where the operators prepare the material required in the production line, where different car models are produced sequentially, one after another. In this scenario, the better the NISSAN engineers distribute the references within the shelves as a function of the planned production or forecast (i.e. car manufacturing planning), the lower the resulting saturation ratio will be. This is mainly due to the resulting reduced travelled distance within the area by both operators and the aforementioned AGV. Unfortunately, the high diversity of references and the constantly changing production forecast add complexity to the engineers in charge for designing those areas in terms of references allocation, required number of operators, the optimum path for the AGVs, towards minimizing the kitting saturation level.

Among the technical strategies utilized for avoiding this undesired saturation regime, the use of avant-garde meta-heuristic optimization algorithms has lately gained momentum in the research community as an efficient means to solve complex optimization problems under a black-box approach. Their use in the context of the optimization of production chains is certainly not new, as many methods have been published in the literature dealing with production scheduling and lot storing, among many others (e.g. see [1,2] and references therein). However, less attention has been paid to the optimization of the physical configuration of the production machinery, mainly due to the difficulty of dynamically moving steady machinery from one location to another, and the inherent economic costs derived therefrom.

In this paper we take a step beyond the use of meta-heuristic solvers for the minimization of the saturation level in car manufacturing production chains. The contribution of this paper goes from the formal problem statement itself to the design of a genetically-inspired optimization algorithm that allows trading optimality for computational efficiency when handling the huge dimensionality of the mixed integer-continuous solution space. Although planned after the submission of this paper, some insights of the validation of the designed algorithm will be also provided.

2 Problem Statement

From a general perspective, the optimization problem can be formulated as to decrease the operator saturation by finding the optimum arrangement of PDLs, references and AGV paths inside the kitting area. A set of assumptions is further made with regards to the setup:

• The cycle time of the assembly line (i.e. the time taken to fill an AGV with references of any given model) is set fixed and imposed by the overall cycle time of the production chain.
• Two different types of PDLs are assumed: trolleys or shelves, each capable of storing a different number of references. Shelves are used for smaller references, whereas trolleys can store bigger ones.
• The type of references.
The routes that the AGVs may follow are affected by both the physical layout of the kitting and the minimum distance between two consecutive paths.

With the above set of assumptions in mind, let us introduce the following notation:

- \( v \): number of vehicles to be assembled.
- \( V_i \): index identifying each vehicle to be assembled, where \( i \in \{1, 2, \ldots, v\} \).
- \( \mu \): total number of references in the kitting area.
- \( R_j \): identification of each reference, where \( j \in \{1, 2, \ldots, \mu\} \).
- \( p \): number of available PDLs in the kitting area at hand.
- \( P_k \): identification of all PDLs in the kitting area, with \( k \in \{1, 2, \ldots, p\} \).
- \( P_{dk} \): distance (in meters) between the AGV entry point to the kitting area and the position of the \( k-th \) PDL.
- \( RP_{j,k} \): indicator function that equals one if reference \( j \) is located in PDL \( k \), and zero otherwise.
- \( RV_{j,i} \): indicator function that equals one if reference \( j \) is assembled in vehicle \( i \), and zero otherwise.
- \( T \): number of possible AGV sub-paths to be considered in the optimization process. The number of sub-paths, or AGV routes, is given by spatial considerations depending on the particular configuration of each kitting.
- \( \tau \): maximum number of AGV paths, i.e. \( T \leq \tau \).
- \( L_q \): distance (in meters) between the AGV entry point to the kitting area and the farthest point of \( q-th \) AGVs path, where \( q \in \{1, 2, \ldots, T\} \).
- \( \Gamma \): daily labour time of an operator.
- \( \alpha \): maximum number of references in a shelf.
- \( \beta \): maximum number of references in a trolley.
- \( S_j \): indicator function that equals one if reference \( j \) is kept in a shelf-type PDL, and zero otherwise.
- \( T_j \): indicator function that equals one if reference \( j \) is kept in a trolley PDL, and zero otherwise.
- \( D \): distance travelled by the AGVs inside the kitting.

By using this notation we refer as \( Rtp_j \) to the time taken to pick reference \( j \). This time is calculated by resorting to MTM (Methods-Time Measurement). Besides, \( Vtp_i \) will stand for the sum of times to pick all references needed for vehicle \( i \), where \( i \) can take on integer values \( i \in \{1, 2, \ldots, v\} \). This aggregate time can be expressed as

\[
Vtp_i \triangleq \sum_{j=1}^{\mu} Rtp_j \cdot RV_{j,i}.
\]  

The time spent by the operator in displacement inside the kitting is calculated between the start point of the kitting and the position of the reference that is located farthest. To compute this time, let \( Rd_j \) denote the distance between the AGV entry point and the PDL storing reference \( j \), where \( j = 1, \ldots, \tau \). Mathematically this can be calculated as

\[
Rd_j \triangleq \sum_{k=1}^{p} P_{dk} \cdot RP_{j,k},
\]
whereas the distance $RV_{d_i}$ between the entry point of the AGV to the kitting area and the farthest reference corresponding to vehicle model $i$ is given by

$$RV_{d_i} \triangleq \max_{1 \leq j \leq \tau} Rd_j \cdot RV_{j,i}.$$  

(3)

Based on this, the shortest AGV path for vehicle $i$ is then obtained as

$$Ad_i \triangleq \min_{1 \leq q \leq T} \max_{1 \leq q \leq \tau} [RV_{d_i} - L_q],$$  

(4)

which yields the total time $Vtd_i$ that an operator spends moving at a speed $Sd$ (meters per second) while picking all references for vehicle $i$

$$Vtd_i = \frac{Ad_i}{Sd}.$$  

(5)

As aforementioned in the introduction, the main objective in the optimization scenario at hand is the minimization of the saturation level undergone by the operator. Nonetheless, we also add an economic operation objective, this being the minimization of the number of changes to be done in order to decrease the saturation. If $CT \triangleq \Gamma / v$ denotes the cycle time of the assembly line (in seconds), one may estimate the saturation level $Sat_i$ associated to vehicle $i \in \{1, 2, \ldots, v\}$ as

$$Sat_i = Vtp_i + Vtd_i,$$  

(6)

from where the optimization problem considered in this manuscript can be formally cast as

$$\text{Minimize} \quad \left\{ \sum_{i=1}^{v} Sat_i \right\}$$  

subject to

$$\sum_{k=1}^{p} RP_{j,k} = 1, \forall j = 1, 2, \ldots, \mu,$$  

(8)

$$\sum_{j=1}^{\mu} RP_{j,k} S_j \leq \alpha, \quad k = 1, 2, \ldots, p,$$  

(9)

$$\sum_{j=1}^{\mu} RP_{j,k} T_j \leq \beta, \quad k = 1, 2, \ldots, p,$$  

(10)

$$T_j = 1 - S_j, \quad j = 1, 2, \ldots, \mu,$$  

(11)

where (8) denotes that each reference is stored in one and only one PDL; (9) and (10) express that each type of PDL can store a maximum number of references; and (11) poses that every reference fits in a shelf or a trolley, but not in both at the same time.

3 Optimization Algorithm

A Genetic Algorithm is employed to optimize the position of every reference in the different shelves and trolleys, as well as the optimum distance and number of AGV paths inside the kitting area. The genetic algorithm, which was
first proposed by Holland in [3] is a stochastic global search method based on the principles of natural evolution and the Darwinian “survival of the fittest” concept.

In short, the genetic algorithm iterates on a population of potentially good solutions (under a fitness criteria), which is successively refined via a set of recombination and mutation operations so as to approach the global optimal solution. This evolutionary algorithm encodes each one of these solutions as a simple data structure commonly referred to as a chromosome (which constitutes an individual), whereas their constituent elements or genes represent the parameters of the problem to be solved. In the present report, numeric coding is used in the genetic algorithm to solve the problem of PDL optimization in the kitting area. In the solution, the order indicates how relevant every reference is when it is located closer to the start of the kitting area. Each chromosome vector is composed by two parts. The first part is encoded as a vector of pointers to each reference of the same length as the total number of references to position in the kitting. The second ones holds the distances of every possible path for the AGVs inside the kitting area.

The Nissan chromosome is a set of parameters which define the solution to the problem at hand. The chromosome is defined as an mixed integer-real number vector $[REF_1, REF_2, \ldots, REF_\mu, L_1, L_2, \ldots, L_\tau]$, where $REF_n \in \{R_1, R_2, \ldots, R_\mu\}$ encodes the optimum reference order with respect to the input to the kitting area, and $L_n \in \mathbb{R}^+$ denotes the distance (in meters) between the AGV entry point to the kitting area and the farthest point of $n$-th AGVs path. The fitness function used by the genetic solver to evaluate the fitness of a given chromosome is given by the operator saturation averaged over a forecasted production, i.e. Expression (7).

The basic Genetic Algorithm operators include selection, crossover and mutation operators. The selection operator is responsible, at every iteration – or, as it is also known, a generation – for selecting the parents to produce the next generation of solutions under a probabilistic parameter directly depending on their fitness value (crossover probability). The higher its fitness, the higher is
the probability of being selected for the next population. The crossover operator is the basic operator for producing new chromosomes in the GA, called offspring. It produces new individuals by exchanging portions of chromosomes at randomly chosen points, known as crossover points. In the present study we use the simple one-point crossover, because of its simplicity.

On the other hand, this study uses two types of mutation operators: the first swaps 2 references, always subjected to being of the same type (shelf or trolley) and being in different PDLs. In reference to Figure 3, if the algorithm selects references P1 and P22, only these references will be swapped. Mutation of type 2 swaps all the references from one PDL to another, always within the same PDL type. Figure 4 shows the result of the algorithm selecting PDL1 - PDL3 and PDL2 - PDL4. The mutation operator is not applied on every newly produced offspring, but it operates under a certain probability, also known as mutation rate.

4 Status on Experimental Validation

Validation tests are being currently performed by using realistic data from 2 different kitting areas located at the Nissan Motor Iberica plant in Barcelona, Spain, taking as an input the production forecast of 3 months ($v = 10320$ vehicles to be assembled). In this production plan there are 2 main Nissan models, each with 3 different sub-models. Computer simulations run beforehand predict an average reduction of about 40% of the distance travelled by the operator when optimizing the reference arrangement in the considered kitting areas via the herein proposed meta-heuristic approach. The results obtained in these computer simulations will be contrasted with those arising from the practical implementation of the produced optimized solutions on the Nissan plant. This will allow debugging the initially outlined model and the algorithm derived for its efficient solving.

5 Concluding Remarks

The main scope of this research is the optimization of operator saturation in a car manufacturing production chain. To this end we have drafted a model of several kitting areas that are currently feeding the assembly line. Based on this model, a Genetic Algorithm is tailored so as to find the best location for all references and, at the same time, optimizing the number and length of the routes that the AGVs follow inside the areas at hand. Currently, validation of results is under way at the Nissan Motor Iberica plant at Barcelona, Spain. Future work includes the addition of economic optimization, which will find the most economic decision to proceed to a redistribution of parts and PDL types in the kitting area.

The presented work has been funded by the European Union FP7 Factories of the Future research project KAP: Knowledge, Awareness and Prediction. The authors would like to acknowledge the support and contributions of KAP industrial and academic partners in the development of this work.
References

1. J. Xie, J. Dong. Heuristic Genetic Algorithms for General Capacitated Lot-sizing Problems. *Computers & Mathematics with Applications*, 44, 1–2, 263–276, 2002.
2. Y. Li, J. Chen, X. Cai. Heuristic Genetic Algorithm for Capacitated Production Planning Problems with Batch Processing and Remanufacturing. *International Journal of Production Economics*, 105, 2, 301–317, 2007.
3. J. H. Holland. Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, Michigan; re-issued by MIT Press, 1992.
Reliability Evaluation of Custom 2D Barcode OCR by using Monte-Carlo Simulation

Evgeny Gershikov\(^1\) and Samuel Kosolapov\(^2\)

Signal and Image Processing Laboratory, ORT Braude Academic College of Engineering, Karmiel, Israel
\(^1\) E-mail: eugeny11@braude.ac.il
\(^2\) E-mail: ksamuel@braude.ac.il

Abstract. Modern smartphones enable fast and easy recognition of a single two-dimensional (2D) Barcode presented to the smartphone’s camera. In our current research, which deals with a camera-based Instant Feedback System (IFS), it is required to recognize multiple 2D barcodes presented to the camera. Additional IFS requirement is to evaluate the orientation of the 2D barcode relative to the camera. In order to make recognition of a plurality of barcodes, presented at the same time to the digital camera, feasible and reliable, special custom format of 2D barcode was created. Well-known OCR (Optical Character Recognition) algorithms were adapted to the multiple target recognition. In order to evaluate the feasibility of the selected IFS approach and selected 2D barcode design, MATLAB-based simulations were performed. The simulations covered the whole process starting from the acquisition of an IFS image and up to the recognition of the 2D barcodes. Considering a big number of the simulation parameters, the usage of the classical Monte-Carlo approach is necessary in order to evaluate the accuracy and reliability of the selected 2D barcode design and selected OCR algorithms. A number of practically interesting IFS configurations were analyzed and compared with respect to reliability. We conclude that the proposed IFS method is feasible at simulation level, but requires higher reliability OCR techniques for practical use. The proposed and similar techniques are also examined in real images.

Keywords: Image Processing, Instant Feedback System, OCR, 2D Barcode, Monte-Carlo Simulation

1 Introduction

2D Barcodes are widely used to store short alphanumeric information. A number of 2D barcode standards are known. A typical compact barcode reader contains a small size digital camera, a digital signal processor and an alphanumeric display or certain means for transferring the extracted barcode information to a computer for future processing. Modern iPhones, Android or Windows8 smartphones and tablets have a high quality camera and processing means powerful enough to recognize 1D barcodes and 2D barcode of different formats. Recently, a number of applications using 2D barcodes were created for smartphones and are available for free download. Most applications of that kind require manual positioning of the camera so that a target containing a 2D barcode will be located inside a certain rectangle of the smartphone’s monitor. This approach enables recognition of a single barcode only.
In our current research, which deals with a camera-based Instant Feedback System (IFS), it is required to recognize multiple 2D barcodes presented to the camera.

In the frames of our exemplary IFS, every student in the class gets a flat card containing a 2D barcode. When asked to answer to a multiple-choice question, presented by the lecturer, students raise their cards in a specific orientation (orientation of the label encodes the number of the chosen answer). A digital camera in the lecturer’s smartphone (or tablet) is used to take an image of the class and the smartphone’s processor is expected to process the image in order to generate a “list of grades” (the list format is student ID, number of the answer, grade).

In order to make recognition of a plurality of barcodes, presented at the same time to the digital camera, feasible and reliable, a special custom format of 2D barcodes was created and well-known OCR algorithms were adapted to the task of multiple target recognition. The barcode format used is described next.

2 Custom 2D Barcode Label

Specific design of our custom 2D barcode contains a 2-digit number in a human-readable form, denoted as Short ID (SID) and a number of color elements (markers) designed to assist the Optical Character Recognition algorithm. The 2D barcode label (2DBCL) is printed on the surface of a flat card, denoted the IFS card. The OCR algorithm must isolate the surface of every 2DBCL from the background and from other labels, register the image of the isolated 2DBCL by using well-known image registration algorithms with the help of the markers and, finally, recognize the SID. Additionally, the orientation of each 2DBCL must be evaluated since it encodes a vote of the card holder, for example, an answer to a multiple choice question.

The 2DBCL used in the first experiments is shown in Fig. 1. As can be seen, the label consists of an outer circle of cyan color with an inner yellow circle. The SID is printed in the center of the barcode label and in this example the SID is 82. Note that the inner circle is truncated on its top side to create a straight line, which is horizontal when the barcode is not tilted. This straight line is used to detect the 2DBCL orientation, which encodes the student vote (chosen answer).

Fig.1 Appearance of the 2DBCL (front side of the IFS card).
Note that the chosen colors were taken due to their simplicity, the ease of the yellow-cyan edge detection and the assumption that these colors are rarely seen in natural photographs taken indoors. However, sometimes the yellow-cyan edge on the side of the 2DBCL is recognized as a straight line due its distortion by the camera optics when positioned at certain angles relative to the camera axis. Thus, we propose using an improved barcode label with another color on sides, as shown in Fig. 2. In this 2DBCL design the magenta color is used for the sides.

The 2DBCL is printed on the front side of the square IFS card. The back side of the card is shown in Fig. 3. Note that this card design corresponds to the case of a multiple-choice question with four possible answers. By positioning the card, so that the chosen answer number points vertically upwards, the student makes his vote. Since this side of the card is towards the student, the camera cannot see the chosen answer and has to decode it using the detected card orientation.

3 Simulation

In order to evaluate the feasibility of the selected IFS approach, MATLAB-based simulations were performed. The simulations covered the whole process starting from the acquisition of an IFS image and up to the recognition of the SIDs and the votes present in it. The block diagram of the simulation units is shown in Fig. 4. As can be seen, there are four main simulation units, which provide creation of synthetic IFS images, and all the steps of the recognition of the SIDs of the simulated voters and their votes. A detailed description of each simulation unit follows.

![Fig.2 Improved 2DBCL (front side of the IFS card).](image-url)
The first implemented simulation unit creates images of the selected design as seen by a digital camera. Simulation parameters are: the distance from the camera to the 2DBCL (on the Z-axis), the position of the IFS card in the camera’s field of view (in the XY plane) and its three 3D orientation angles. In addition to these, typical camera distortions can be also taken into account in the simulation module. For simplicity, we do not include their effects in our simulation. The resulting images can be stored as uncompressed BMP file or as compressed JPG file, whereas image resolution and compression parameters can be controlled by the user. An example of a simple simulated IFS image with two barcodes is shown in Fig. 5.
Note that in Fig. 5 the upper 2DBCL is simulated as further from the camera and both patterns are both rotated (simulating the student votes) and distorted by a projective transform simulating a certain angle of view of the camera.

The second simulation unit we created searches for the markers that allow identifying the barcodes and isolates available 2DBCLs from the background. The simplest way to locate the barcode labels of the design proposed in Fig. 2 is to search for cyan-yellow edges, which are straight lines in the image. For this purpose the edges of the IFS image were calculated (e.g., by Canny's algorithm [1]) and then the straight lines in the edge image were located using the Hough transform [2]. The next stage was to iterate through these lines and keep only those that correspond to an edge between yellow and cyan regions by comparing the real colors on either side of each checked line and the expected colors. A simple way to compare the real and expected colors is to take the normalized inner product of the two RGB color vectors $\text{RealColor}$ and $\text{ExpColor}$, i.e.,

$$\frac{\text{RealColor} \cdot \text{ExpColor}}{\|\text{RealColor}\| \|\text{ExpColor}\|}$$

and check if it is above a predefined threshold (for a match) or not. Such a threshold should be a number close to 1, e.g., 0.93.

Once the correct lines were detected, the areas of the 2DBCLs can be located and processed in the following units.

Our third simulation unit recognizes the orientation of the barcode labels and isolates the areas of the digits. The recognition of the orientation of each barcode is simple once the yellow-cyan straight edge has been detected in the previous step. For each such edge there are two possible orientations (for example, the yellow side is at the top and the cyan side is at the bottom or vice versa) and we check the colors to choose the correct one of the two possibilities. Once the orientation is extracted, we can rotate the barcodes so that the digits become horizontal.

To detect the digit area correctly without knowing the distance to camera, its view angle and its resolution we apply a simple method based on watersheds [4] to locate the circular area of each 2DBCL (see Figs. 1-2) and use it to determine the bounding box of the digits. The watersheds algorithm is applied to the edges of the blue color component since this was found to provide good performance considering the colors chosen for the 2DBCL design.

The fourth simulation unit provides the last stage of OCR by using simple methods available in MATLAB programming environment. First, the bounding box containing the digits is converted to a binary image and the text in it is separated into its connected components. Each connected component is expected to be a digit and is compared to a database of the 10 digits using the normalized 2-D cross-correlation [3] after resizing all the digits to be of a predefined size. The examined digit is recognized by choosing the database digit providing the highest cross-correlation in the comparison. Two OCR algorithm versions were considered: without image registration of the recognized digit and the compared database digits prior to the calculation of the normalized cross correlation (OCR-1 method) and with it (OCR-2).
The results of the full simulation chain are the SIDs of the 2DBCLs and their original orientations.

4 Monte-Carlo Simulation

Considering a big number of practically required simulation parameters, the classical Monte-Carlo approach was necessary to evaluate the accuracy and reliability of the IFS approach. In the stage of Monte-Carlo simulations, the full chain of simulation units was operated a number of times, where at every simulation run the values of the selected set of setup parameters were modified in a pseudo-random way. The main setup parameters of the simulation are the number of 2DBCLs in the synthetic IFS image, the SIDs used, the orientations of the barcodes (that encode the chosen answers) as well as the viewing angles of the barcodes relative to the camera and the distances from it.

We can summarize our conclusions from the numerous simulations performed in the following bullets.

- The angle of the barcodes relative to the camera should be small, so that the barcode label is not very distorted. Otherwise, while it is still possible to recognize the circular area of the label, the SID digits recognition is not reliable.
- The distance of the barcode to the camera should not be too big and it depends on the camera resolution. Each barcode label should be at least of the size of around 40×40 pixels.
- Any orientation of each 2DBCL is allowed corresponding to the chosen answer of the simulated voter. The algorithm performance does not depend much on the orientations.
- Any SID of two digits can be used. This covers the case of up to 100 students. 3 digits SIDs for bigger classrooms were not tested, but we expect a similar performance since each digit is recognized separately.
- The number of the barcodes simulated in the IFS image does not affect the recognition performance, but has significant influence on the processing time.
- The reliability of the orientation recognition of the barcodes is high enough. Errors up to almost 90 degrees are theoretically allowed in the recognition when considering multiple-choice questions with four possible answers. In our simulations the errors were less than 15 degrees.
- The use of image registration techniques to improve the accuracy of the SID recognition did not justify the added complexity and running time. Although for individual images sometimes the OCR-2 method with registration provided better performance than OCR-1, on average both methods provided similar accuracy of the results.
- The simple OCR methods used are influenced by small distortions of the form of the digits. Other more complex OCR techniques can be considered for future research.
Conclusions

In this work we examined a new camera-based method for Instant Feedback, that can be used in classrooms in institutions of education, such as schools, colleges and universities. The proposed simple IFS system consists of a camera used to acquire a photo of the classroom and a collection of IFS cards with a specifically designed 2D barcode labels printed on them. The barcode was designed to allow reliable detection and recognition of the SID of the cardholder as well as the orientation of the label. This orientation encodes the student’s vote when answering a multiple-choice question, asked by the lecturer. Thus, both the SIDs and the orientations have to be reliably extracted from the IFS image captured by the camera. The reliability of our proposed system was tested using Monte-Carlo method of simulations. We conclude that while the reliability of the orientation recognition was good enough, the reliability of the SID recognition was very dependent on the position of the barcodes relative to the camera – when the barcodes were simulated at a big angle relative to the camera optical axis, the recognition accuracy was low. We believe that this result can be improved by applying other more advanced OCR techniques or considering other barcode designs that do not include digits. These ideas are currently under research [5].

References

[1] J. Canny, “A computational approach to edge detection”, IEEE Transactions on PAMI, 8(6), 679-697, 1986.
[2] R. O. Duda and P. E. Hart, “Use of the Hough Transformation to Detect Lines and Curves in Pictures”, Comm. ACM, 15, 1, 11-15, 1972.
[3] R. M. Haralick and L. G. Shapiro, “Computer and Robot Vision”, Volume II, Addison-Wesley, 316-317, 1992.
[4] F. Meyer, ”Topographic distance and watershed lines”, Signal Processing, 38, 113-125, July 1994.
[5] S. Kosolapov, E. Gershikov and N. Sabag, “Feasibility of Camera-Based Instant Feedback System”, Think Mind, Content 2013 Proceedings, 23-29, 2013.
Nonparametric Identification of ARX-Processes *

Irina Yu. Glukhova1 and Gennady M. Koshkin2

1 Department of Applied Mathematics and Cybernetics, Tomsk State University, 36 Lenin ave., 634050 Tomsk, Russia
(E-mail: win32.86@mail.ru)
2 Department of Applied Mathematics and Cybernetics, Tomsk State University, 36 Lenin ave., 634050 Tomsk, Russia
(E-mail: kgs@mail.tsu.ru)

Abstract. The principal parts of mean square errors for kernel plug-in estimators of the functions defining ARX-process are found. We use simulation to compare parametric and nonparametric identification and forecasting algorithms. To investigate the dependence of Russian Federation's Industrial Production Index on the dollar exchange rate, direct investments, and export for the period from September 1994 to January 2013, the proposed algorithms of identification and forecasting are applied.

Keywords: Kernel plug-in estimator, conditional mean, mean square error (MSE), ARX-process, nonparametric identification, forecasting algorithm.

1 Introduction

Suppose that a sequence \((Y_t)_{t=\ldots,-1,0,1,\ldots}\) is generated by ARX\((m, p, d)\)-process

\[
Y_t = \Psi(Y_{t,m}, X_{t,s}) + \xi_t,
\]

where \(Y_{t,m} = (Y_{t-i_1}, \ldots, Y_{t-i_m})\), \(X_{t,s} = (X_{t-j_1}, \ldots, X_{t-j_r}, \ldots, X_{t-j_k})\), \(s = r + \ldots + k, \ d = \max(r, \ldots, k)\), \(1 \leq i_1 < \ldots < i_m \ll n, 0 \leq j_1 < \ldots < j_k \ll n\) are known subsequences of natural numbers, \((\xi_t)\) is a sequence of i.i.d. random variables with zero mean, finite variance, zero third, and finite fourth moments, \(\Psi(Y_{t,m}, X_{t,s})\) is an unknown non-periodic function bounded on compact.

Models (1) are widely used on identification of economic systems and financial time series analysis. By identifying model (1) we mean the problem of parametric or nonparametric estimation of the function \(\Psi\). In this paper, we assume that the process \((Y_t)_{t=\ldots,-1,0,1,\ldots}\) is a strictly stationary process and satisfies the strong mixing (s.m.) condition with s.m. coefficient (Masry[9], Masry and Tjostheim[10], Kitaeva and Koshkin[3], [4])

\[
\alpha(\tau) \approx e^{-\delta \tau}, \ \delta > 0, \ \tau \to \infty.
\]

Let \(Y_1, Y_2, \ldots, Y_n\) be observations generated by the process (1). As a model of the structure of \(\Psi\) in (1), we take the conditional expectation

\[
b(y, x) = E(Y_t|Y_{t,m} = y, X_{t,s} = x) = E(Y|y, x), \ (y, x) \in \mathbb{R}^{m+s}.
\]

* Supported by Russian Foundation for Basic Research, project no. 13-08-00744
According to Dobrovidov et al.\cite{1} the integrals
\[ a_g(y, x) = \int q^g f(q, y, x) dq, \quad g = 0, 1, \]
are basic functionals, where \( f(q, y, x) \) is an unknown probability density function (p.d.f.) of a random vector \((Y_t, Y_{t,m}, Y_{t,s})\) in stationary conditions. Since
\[ a_0(y, x) = \int f(q, y, x) dq = p(y, x), \]
where \( p(y, x) \) is p.d.f. of the vector \((Y_{t,m}, Y_{t,s})\), then the conditional expectation can be written as
\[ b(y, x) = \frac{a_1(y, x)}{a_0(y, x)} = \frac{a_1(y, x)}{p(y, x)} = \int Y_t f(Y_t|y, x) dY_t. \]
We take the kernel estimators of basic functionals \( a_g(y, x) \) at the point \((y, x)\) in the following form:
\[ a_{gn}(y, x) = \frac{1}{n - Q} \sum_{i=Q+1}^{n} \frac{Y_{i}^g}{h_j} K_m \left( \frac{y - Y_{i,m}}{h_y} \right) K_s \left( \frac{x - X_{i,s}}{h_x} \right) \frac{r}{\prod_{j=1}^{k} h_{pj}}, \]
where \( Q = \max(i_m, \max(j_r, \ldots, j_k)), h^y = (h_1, \ldots, h_m), h^x = (h_1^x, \ldots, h^x_k), h^z = (h_1^z, h_1, \ldots, h_{pk}^z), h^y = (h_1^y, \ldots, h^y_k) \) are suitable bandwidths (positive numbers), \( K_m \) and \( K_s \) are \( m \)- and \( s \)-dimensional kernels. Thus, the kernel plug-in estimator of conditional functional \( b(y, x) \) at the point \((y, x)\) and, hence, the function \( \Psi(y, x) \) in (1) is the ratio
\[ b_n(y, x) = \Psi_n(y, x) = \frac{\sum_{i=Q+1}^{n} \frac{1}{h_j} K_m \left( \frac{y - Y_{i,m}}{h_y} \right) K_s \left( \frac{x - X_{i,s}}{h_x} \right) \frac{r}{\prod_{j=1}^{k} h_{pj}}}{\sum_{i=Q+1}^{n} \frac{1}{h_j} K_m \left( \frac{y - Y_{i,m}}{h_y} \right) K_s \left( \frac{x - X_{i,s}}{h_x} \right) \frac{r}{\prod_{j=1}^{k} h_{pj}}}. \quad (3) \]

The problem of identifying model (1) is a problem of estimating function (cf. Kitaeva and Koshkin\cite{3}, Dobrovidov et al.\cite{1})
\[ H(A) = H(a_0, a_1) = \frac{a_1}{a_0}, \quad a_0 > 0, \quad (4) \]
where \( A = (a_0, a_1), a_g = a_g(u) = \int q^g f(q, u) dq, g = 0, 1, f(q, u) = f(z), z \in R^{m+s+1}, \) is p.d.f. of the random vector \((Y_t, U_z) = (Y_t, Y_{t,m}, X_{t,s}) = Z_t \) in stationary conditions.

In this paper, we study the mean square convergence of estimator (3) to the function \( \Psi \) determining ARX-process (1).
2. The principal part of MSE for plug-in estimator of $\Psi$

We introduce the following notation:

$$f_1(i_1(i+1) \ldots (i_jk+1)z, u, u, u)$$

is $4(m+s+1)$-dimensional p.d.f. of sample vectors $Z_1, Z_2, \ldots, Z_n$.

$$a_1^{(i_1(i+1) \ldots (i_jk+1)z, u, u, u)} = 
\int_{R^4} |\nu\nu'\nu'|^p f_1(i_1(i+1) \ldots (i_jk+1)z, u, u, u) \, d\nu \, d\nu' \, du,'
Q + 1 \leq i, j, k < n, \ i + j + k \leq n - 1;
$$

$$a_1^{(i_1(i+1) \ldots (i_jk+1)z, u, u, u)} = 
\int_{R^4} |\nu\nu'\nu'|^p f_1(i_1(i+1) \ldots (i_jk+1)z, u, u, u) \, d\nu \, d\nu' \, du',
\nu, z, u, w, u, u, u,
$$

Assume that for the function $H(\cdot) : R^L \rightarrow R^1$ belongs to the class $N_0(z)$

$$a_0(u) > 0, a_0(\cdot) \in N_0(R), \sup_u |a_0(u)| < \sup_u \left| \frac{\partial a_0(u)}{\partial u_1 \ldots \partial u_q} \right| < \infty;$$

Thus, we can construct an estimator for $\Psi$ and study its MSE. For each variable from $m + s$ variables in Theorem and the product of one-dimensional kernels as multidimensional kernels of proper dimensions.

$$\omega_{g+p} = \frac{T_0}{T_0} \sum_{j=1}^{L} \frac{\partial^\nu a_p(u)}{\partial u_j}, \quad H_g = \frac{\partial H(A)}{\partial (a_p)}$$

Definition 1. A function $K(u)$ belongs to the class of one-dimensional kernels $K(\cdot) \in A_\nu$, if $|K(u)|du < \infty, \int K(u)du = 1, \int |u''K(u)|du < \infty,$

$$T_j = \int u'' K(u)du = 0, j = 1, \ldots, \nu - 1, \nu - 1, T_\nu \neq 0, \text{ and } K(u) = K(-u).$$

Below, to study convergence of estimators, we use the same bandwidth $h_n$ for each variable from $m + s$ variables in Theorem and the product of one-dimensional kernels as multidimensional kernels of proper dimensions.
Then $E[H(A_n) - H(A)]^2 =$
\[
\sum_{g, p=0} H_g H_p \left\{ \frac{a_t + p(x)}{n h_n^L} \left( \int K^2(u) du \right)^L + \omega_{2\nu}(z) \omega_{2\nu}(z) h_n^{2\nu} \right\} + O \left( \left[ h_n^{2\nu} + \frac{1}{nh_n^L} \right]^2 \right).
\]

Note that in this formula according to (4) $H_0 = -\frac{a_1}{a_0}$, $H_1 = \frac{1}{a_0}$. The proofs of Theorem is based on the results, presented in Masry[9], Dobrovidov et al.[1], Kitaeva and Koshkin[4], Koshkin[5], Koshkin and Glukhova[6].

3 Comparison of parametric and nonparametric algorithms

Computer modeling is started by generation sequences of dependent observations, using the following processes:

$M(1) : Y_n = 0.2Y_{n-1} + 0.11X_1^n + 0.15X_1^{n-1} + 0.3X_2^n + 0.2X_2^{n-2} + \varepsilon_n,$

$M(2) : Y_n = 0.01Y_{n-1}Y_{n-2} + 0.2X_1^n + 0.03X_2^nX_1^{n-1} + 0.7X_3^n + \varepsilon_n,$

$M(3) : Y_n = e^{0.11n^{-1} + 0.2X_1^n + 0.1X_1^{n-1} + 0.01X_2^n + 0.03X_3^n + \varepsilon_n}.$

Here variables $X_1, X_2,$ and $X_3$ take values from uniform distribution laws on the corresponding intervals $[2, 2.5]$, $[5, 6]$, $[8, 10]$, and random variables $\varepsilon_n$ are distributed according to normal distributions with zero mean and variances, which are calculated for models $M(1)$–$M(3)$ by the formula

$$\sigma^2 = \frac{Y_{\max} - Y_{\min}}{6} \cdot a,$$

where the multiplier $a$ is a level of noise $\varepsilon_n$, and $a$ takes the values $0.01, 0.05, 0.1, 0.15, 0.2, 0.5$. We use in simulation the following sizes of observations: 50, 100, 200, and 500.

Note that condition (2) holds for models $M(1)$ and $M(2)$ (see subsection 3.5.9 in Dobrovidov et al.[1]).

Identification algorithms for functions $\Psi$ in (1) were obtained by the least squares method (LSM), iterative weighted least squares method (WLSM), and nonparametric approach. The LSM and WLSM estimators are computed by making use of MATLAB built-in functions.

Simulation of nonparametric algorithms is also based on MATLAB. As a kernel $K(u)$, we use the standard Gaussian density. The bandwidths are defined in two ways. In accordance with Stone[12], Hall[2], Li and Racine[8], Leung[7], the bandwidths are calculated by the cross-validation (CV) method. The second method of finding the bandwidth uses the estimate $Y_{\delta}$, based on the following empirical criteria:

$$h_{Empirical} = C_0\sigma_{\delta}n^{-\frac{1}{4}}$$,

$$C_0 = \arg\min_{0<C<\infty} \left| \sum_{i=Q}^{n-1} Y_{i}K_{L} \left( \frac{U_{n} - U_{i}}{h} \right) \right|$$,

$$Y_{\delta} = \frac{1}{n} \sum_{i=Q}^{n-1} \frac{K_{L} \left( \frac{U_{n} - U_{i}}{h} \right)}{h}.$$
\( j = 1, L \), where \( L \) is the dimension of function \( \Psi \), \( \sigma_j^2 \) is the sample variance of observations for the \( j \)-th variable, \( h = (h^y, h^x) \).

**Fig. 1.** Averaged identification error for \( M(1) \), \( n = 50 \)

**Fig. 2.** Averaged identification error for \( M(2) \), \( n = 50 \)

**Fig. 3.** Averaged identification error for \( M(3) \), \( n = 50 \)

**Fig. 4.** Dependence of the identification quality of \( M(1) \) on the size of observations for the noise level 0.15

For models \( M(1)–M(3) \) the values of the identification errors

\[
A_{n,j} = \frac{1}{n - Q} \sum_{i=Q+1}^{n} \left| \frac{Y_i - \bar{Y}_i}{Y_i} \right|, \quad j = 1, 2, 3,
\]

\( n = 50 \), are presented in Figures 1–3. The results for other sizes of observations are shown in Figures 4–6. Note that all the simulation results are averaged over 20 different samples of the same size.

According to Figures the identification quality for all the models and the methods decrease with increasing the level of noise. Further, for non-linear model \( M(3) \) nonparametric algorithms have advantages over parametric algorithms because of their adaptability. There is a tendency of reduction of identification errors for all models by increasing sizes of observations.
4 Real data processing

We will examine the dependence of Russian Federation’s Industrial Production Index (IPI) $Y$ (see Figure 7) on the dollar exchange rate $X^1$, import $X^2$, and direct investment $X^3$ for the period from September 1994 to January 2013. The data are available from the following cites: http://www.gks.ru and http://sophist.hse.ru/. Apply (3) under

$$U_i = (Y_{i,1}, X_{i,4}) = (Y_{i-1}, X^1_{i}, X^2_{i}, X^3_{i}, X^3_{i-1}).$$

Due to the fact that the classification principles of economic activities were changed in 2002, we consider two series of the data from September 1994 to December 2002 and from January 2003 to January 2013.

We take the Gaussian density as the kernel $K(u)$, the bandwidths

$$h_j = 1.1\sigma_j n^{-1/9}$$
for the data from September 1994 to December 2002, and

\[ h_j = 1.23 \sigma_j n^{-1/9} \]

for the data from January 2003 to January 2013, \( j = 1, 2, 3, 4, 5 \), where the constants 1.1 and 1.23 are obtained by the above empirical criteria. To compare the nonparametric algorithms (3) and the LSM-estimators, we have calculated the relative errors \( A_n \) and the relative average annual errors \( A(t) \), \( t = 1994, \ldots, 2013 \), for both the approaches:

\[ A_n = \frac{1}{n} \sum_{i=1}^{n} \frac{|Y_i - \hat{Y}_i|}{Y_i}, \quad A(t) = \frac{1}{12} \sum_{i=1}^{12} \frac{|Y_i(t) - \hat{Y}_i(t)|}{Y_i(t)}, \]

where \( Y_i \) is the true value of the IPI and \( \hat{Y}_i \) is its estimate. The results of such a comparison are given in Figure 8 and Figure 9.

![Fig. 8. Identification relative errors \( A(t) \)](image)

![Fig. 9. Identification relative errors \( A(t) \)](image)

![Fig. 10. Forecasting relative errors \( A(t) \)](image)

The results of 1998 and 2009 can be explained by Russian financial crisis ("Ruble crisis") in August 1998 and Global financial crisis in 2009.
To predict the IPI $Y$ for the data from 2002 to 2013 (cf. Simakhin[11]), we apply (3) under

$$U_i = (Y_{i,1}, X_{i,4}) = (Y_{i-1}, X_{i-1}^1, X_{i-1}^2, X_{i-1}^3, X_{i-2}^3).$$

Here the bandwidths are equal to

$$h_{jt} = 0.94\sqrt[3]{\sigma_j t^{-1/3}},$$

where $\sigma_j$, $j = 1, 2, 3, 4, 5$.

The similarity of identification algorithms and forecasting algorithms leads one to expect the both should behave similarly. For the relative average annual errors $A(t)$, seen in Figure 9 and Figure 10, one indeed observed that.

References

1. A.V. Dobrovidov, G.M. Koshkin and V.A. Vasiliev. Non-Parametric State Space Models. Kendrick Press, Inc., Heber, UT 84032, USA, 2012.
2. P. Hall. Asymptotic properties of integrated square error and cross-validation for kernel estimation of a regression function. Z. Wahrscheinlichkeitstheorie Verw, Geb. 67, 175–196, 1984.
3. A.V. Kitaeva and G.M. Koshkin. Semi-recursive nonparametric identification in the general sense of a nonlinear heteroscedastic autoregression. Automation and Remote Control, 2, 92–111, 2010.
4. A.V. Kitaeva and G.M. Koshkin. Nonparametric semirecursive identification in a wide sense of strong mixing processes. Problems of Information Transmission, 46, 1, 22–37, 2010.
5. G.M. Koshkin. Deviation moments estimates of substitution and its piecewise-smooth approximations. Siberian Mathematical Journal, 40, 3, 605–618, 1999.
6. G.M. Koshkin and I.Yu. Glukhova. Nonparametric identification of nonlinear ARX-processes. Journal of Control and Computer Science. Tomsk State University, 3, 20, 55–61, 2012 (in Russian).
7. D. Leung. Cross-validation in nonparametric regression with outliers. Annals of Statistics, 33, 2291–2310, 2005.
8. Q. Li and J.S. Racine. Cross-validated local linear nonparametric regression. Statistica Sinica, 14, 485–512, 2004.
9. E. Masry. Probability density estimation from sampled data. IEEE Trans. Inf. Theory, IT–29, 5, 696–709, 1983.
10. E. Masry and D. Tjostheim. Nonparametric estimation and identification of nonlinear ARCH time series Econometric Theory, 11, 2, 258–289, 1995.
11. V.A. Simakhin. Adaptive robust nonparametric prediction algorithms. Journal of Control and Computer Science. Tomsk State University, 1, 1, 45–54, 2011 (in Russian).
12. C.J. Stone. Cross-validatory choice and assessment of statistical predictions (with discussion). Journal of Royal Statistical Society, 36, 111–147, 1974.
A Study on European Football Championships in the GLMM Framework with an Emphasis on UEFA Champions League Experience

Andreas Groll\textsuperscript{1} and Jasmin Abedieh\textsuperscript{2}

\textsuperscript{1} Department of Mathematics, Workgroup Financial Mathematics, Ludwig-Maximilians-University, Theresienstr. 39, 80333 Munich, Germany (E-mail: andreas.groll@math.lmu.de)
\textsuperscript{2} Jasmin Abedieh (E-mail: jasmin.abedieh@hotmail.de)

Abstract. This article has two major objectives. First, the results of a preceding article are revised, where all matches of the European football championship (EURO) 2012 have been predicted on the quite small data basis of the two preceding EUROs, resulting in a possible course of the tournament. There, a pairwise Poisson model for the number of goals scored by national teams competing in EURO matches was established, incorporating two approaches for variable selection, which was then used for prediction. Including the data of the EURO 2012, in the present article this analysis is replicated on a more reliable data basis and the set of selected influence variables is compared to the results of the preceding analysis. Besides, the course of the EURO 2012 suggests a positive correlation between a national team’s success at a EURO and the number of its players that have been successful in the preceding Union of European Football Associations (UEFA) Champions League (CL) season. Hence, a second objective of this article is to check, if in fact a significant influence of this covariate can be detected.

Keywords: Football, European football championships, UEFA Champions League, Sports tournaments, Generalized linear mixed model, Lasso, Variable selection.

1 Introduction

Recently, the statistical analysis of major soccer events such as the UEFA CL (see Leitner et al. \cite{15}, Eugster et al. \cite{4}), the EURO (see Leitner et al. \cite{12}, Leitner et al. \cite{13}, Zeileis et al. \cite{17} or Groll and Abedieh \cite{10}) or the Fédération Internationale de Football Association (FIFA) World Cup (see Leitner et al. \cite{14}, Stoy et al. \cite{16}, Dyte and Clarke \cite{3}) has gained more and more attention. A major and challenging objective in this context is to predict the respective tournament winner. In general, the existing approaches can be divided into two major categories: approaches based on the easily available source of “prospective” information contained in bookmakers’ odds (compare Leitner et al. \cite{12}, Leitner et al. \cite{14} and Zeileis et al. \cite{17}) and regression based models (compare Stoy et al. \cite{16}, Dyte and Clarke \cite{3} or Groll and Abedieh \cite{10}).

Based on the 62 matches of the EUROs 2004 and 2008, in Groll and Abedieh \cite{10} a pairwise Poisson model for the number of goals scored by national teams in the single matches of the tournaments is used for prediction of the EURO
2012. There, 32 potential influence variables are considered and team-specific random effects are included, resulting in a flexible generalized linear mixed model. Each match occurs in the data set in the form of two different rows, one for each team, containing both the variables corresponding to the team whose goals are considered as well as those of its opponent. The matched-pair design is accounted for by including a second match-specific random intercept, following Carlin et al. [2], which is assumed to be independent from the team-specific random intercept. Two different methods for the selection of relevant predictors, an $L_1$-penalization based technique (see Groll and Tutz [11]; implemented in the `glimmLasso` function of the corresponding R-package from Groll [9]) as well as subset selection, are used to obtain a sparse final model, which is then used for the prediction of the whole tournament outcome of the EURO 2012: it contains only the four variables market value (MV; of both teams), the maximum number of teammates and UEFA points. Note here that in contrast to other team sports such as basketball, ice-hockey or handball, in soccer pure chance plays a larger role, as in soccer fewer points (goals) are scored and thus single game situations can have a tremendous effect on the outcome of the match. Hence, the prediction of soccer tournaments is especially demanding. Nevertheless, the forecast of the EURO 2012 tournament outcome in Groll and Abedieh [10] shows surprisingly many accordances with the true one: seven of the eight teams that qualified for the knockout stage were predicted correctly, three of the four teams that qualified for the half-finals and finally, the tournament winner Spain.

However, their results base on the quite small data basis of only two preceding EUROs. Besides, intuitively it is somewhat surprising that the covariate maximum number of teammates has a significant effect at all and that this effect is negative. Therefore, a major objective of the present article is to revise the results of Groll and Abedieh [10], including the data of the EURO 2012. We replicate their analysis on a more reliable data basis and compare the set of selected influence variables with the results of the preceding analysis.

In Groll and Abedieh [10] it is already mentioned that for the half-final of the EURO 2012, with the national teams of Spain, Germany and Portugal, exactly those three teams qualified that had the largest proportion of players amongst their squad that reached at least the half-finals of the UEFA CL 2012: Spain with 14, Germany with 10 and Portugal with 4 players. All other national teams, except for France with 3 players, had only 2 or fewer players that reached at least the half-finals of the preceding UEFA CL season. Besides, also Frohwein [5] has already pointed out that there is a connection between the final rounds of the UEFA CL and the FIFA World Cup final. Though this

---

1 Several economic (GDP per capita, population size) and sportive factors (unfairness points, home advantage, odds, market value, FIFA points, UEFA points) as well as factors describing the team’s structure (maximum number of teammates, second maximum number of teammates, average age, number of CL players, number of Europa League players, age of the national coach, nationality of the national coach, number of legionnaires) have been included in the analysis, both of the team whose goals are considered and of its opponent. For a detailed description of these variables consult Groll and Abedieh [10].
coherence seems too distinct to be just a matter of chance, based on the data of the EUROs 2004 and 2008, no clear significance of the covariate number of CL players could be detected in Groll and Abedieh [10]. Hence, a second major objective of this article is to investigate, if the number of a national team’s players, which reached at least the half-finals in the preceding UEFA CL season, has now a significant influence on the team’s success at the subsequent EURO tournament, if the data of the EURO 2012 are included.

The rest of the article is structured as follows. In Section 2 a pairwise Poisson model for the number of goals is used to determine the covariates of a final model, which serves as control model with regard to the preceding analysis presented in Groll and Abedieh [10]. In Section 3 the explorative power of the covariate number of CL players with respect to the success of national teams at EURO tournaments is investigated, before we conclude in Section 4.

2 Poisson Regression on the EUROs 2004-2012

The following regression analysis is based on the mixed Poisson model presented in Section 2 of Groll and Abedieh [10], with 32 covariates and the number of goals scored by national teams in the single matches of the tournaments as response variable. Team-specific random intercepts are included in order to adequately account for different basis levels of the national teams, as well as match-specific random intercepts to model the matched-pair design. We use two different approaches that are both able to perform variable selection, an $L_1$-penalization technique, which is proposed by Groll and Tutz [11] and implemented in the glmmLasso function, and forward subset selection based on the glmer function (Bates and Maechler [1]), denoted by glmer-select.

For the Lasso approach we obtain different levels of sparseness by changing the determination procedure of the optimal tuning parameter. In the following we consider three techniques: AIC, BIC and leave-one-out cross-validation (LOOCV)$^3$. BIC leads to the sparsest models, followed by AIC, whereas the LOOCV yields models that include several covariates. The sparseness of the models obtained by the forward selection procedure glmer-select can be controlled directly by the level of significance $\alpha$ in the corresponding model testing, which is based on an analysis of deviance. Table 1 presents the corresponding results for $\alpha \in \{0.01, 0.05, 0.1\}$. We consider the same models as in Groll and Abedieh [10], decreasing step by step the number of given influence variables:

- **Model 1**: A model containing all covariates is fitted.
- **Model 2**: A model containing all covariates except for the variable ODDSET odds is fitted.
- **Model 3**: As the variable fairness is not available for the prediction of future EUROs, a model containing all covariates except for ODDSET odds and fairness is fitted.

$^2$ measured by the number of goals scored in the single matches of the next EURO.

$^3$ Due to the matched-pair design, not single observations but single matches are excluded from the training data.
Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013) International Conference, Mataró (Barcelona), Spain 25 - 28 June 2013

| BIC | AIC | LOOCV | \( \alpha = 0.01 \) | \( \alpha = 0.05 \) | \( \alpha = 0.1 \) |
|-----|-----|-------|----------------|----------------|----------------|
| ODSS | ODSS | ODSS | ODSS | ODSS | ODSS |
| - | - | - | MV opp. | MV opp. | MV opp. |
| - | - | - | MV opp. | MV opp. | MV opp. |
| - | - | - | UEFA pts. | UEFA pts. | UEFA pts. |
| - | - | - | # CL players | # CL players | # CL players |
| - | - | - | nat. coach | nat. coach | nat. coach |

| M1 | MV opp. | MV opp. | MV opp. | MV opp. | MV opp. |
|-----|---------|---------|---------|---------|---------|
| - | - | - | MV opp. | MV opp. | MV opp. |
| - | - | - | MV opp. | MV opp. | MV opp. |
| - | - | - | FIFA pts. | FIFA pts. | FIFA pts. |
| - | - | - | # CL players | # CL players | # CL players |
| - | - | - | nat. coach | nat. coach | nat. coach |

| M2 | MV opp. | MV opp. | MV opp. | MV opp. | MV opp. |
|-----|---------|---------|---------|---------|---------|
| - | - | - | MV opp. | MV opp. | MV opp. |
| - | - | - | MV opp. | MV opp. | MV opp. |
| - | - | - | FIFA pts. | FIFA pts. | FIFA pts. |
| - | - | - | # CL players | # CL players | # CL players |
| - | - | - | nat. coach | nat. coach | nat. coach |

| M3 | MV opp. | MV opp. | MV opp. | MV opp. | MV opp. |
|-----|---------|---------|---------|---------|---------|
| - | - | - | MV opp. | MV opp. | MV opp. |
| - | - | - | MV opp. | MV opp. | MV opp. |
| - | - | - | FIFA pts. | FIFA pts. | FIFA pts. |
| - | - | - | # CL players | # CL players | # CL players |
| - | - | - | nat. coach | nat. coach | nat. coach |

Table 1: Selected variables for \texttt{glmmLasso} and \texttt{glmer-select} for Model 1-3 and different levels of sparseness (covariates have been standardized).

Groll and Abedieh [10] suggest a possible goodness-of-fit criterion to assess the performance of these models, based on the “three-way” odds from the German state betting agency ODDSET for all 93 games of the EUROs 2004-2012, which can be directly transformed into (approximate) probabilities \( \hat{p}_i \). On the other hand, let \( G_k \) denote the random variables representing the number of goals scored by Team \( k \) in a certain match and \( G_l \) the goals of its opponent, respectively. Then we can compute the same probabilities by approximating \( \hat{p}_1 = P(G_k > G_l), \hat{p}_2 = P(G_k = G_l) \) and \( \hat{p}_3 = P(G_k < G_l) \) for each of the 93 matches using the corresponding Poisson distributions, whereas the estimates \( \hat{\lambda}_k \) and \( \hat{\lambda}_l \) are obtained by our regression models. Hence, we can provide a goodness-of-fit criterion by comparing the values of the log-likelihood of the 93 matches for the ODDSET odds with those obtained for our regression models. For \( \omega_j \in I, j = 1, \ldots, 93 \), the likelihood is given by the product \( \prod_{j=1}^{93} \hat{p}_{ij} \), with \( \hat{p}_{ij} \) denoting Kronecker’s delta. The log-likelihood scores for \texttt{glmmLasso} and \texttt{glmer-select} corresponding to Model 1-3 and different levels of sparseness can be found in Table 2. The results show that for all settings the fit obtained by our regression models outperforms the log-likelihood score corresponding to the ODDSET odds (which yields -94.25) and hence, the models seem reasonable.

Table 1 shows that the results for \texttt{glmer-select}, which serve as a control for our \( L_1 \)-penalization approach, generally agree with those obtained by the \texttt{glmmLasso} function, but are somewhat sparser. For each setting, the \texttt{glmmLasso} approach based on LOOCV chooses several more variables than the
other approaches. For Model 1 all methods select the ODDSET odds, while for Model 2 almost all methods select fairness and the MV of the opponent. The MV of both teams is selected by almost all methods for Model 3. Note that the covariate maximum number of teammates, which was surprisingly selected in the preceding analysis, has not been selected in any of the regarded models and thus will not be incorporated in our final model. The variables UEFA points and number of CL players (each for both teams) are selected for each setting by glmmLasso based on LOOCV, but for no other method. Similar to Groll and Abedieh [10], we focus on the contribution of those covariates that seem to be able to adequately replace the bookmakers’ odds and consider all covariates from Model 2 and 3 that have been selected at least three times (and for at least two of the six approaches), except for the variable fairness (as it cannot be observed before the start of the tournament and thus cannot be used for prediction). This yields the following sparse predictor

$$\log(\lambda_{it,ji}) = \beta_0 + (\text{MV})_{it,ji} \beta_1 + (\text{MV opp.})_{it,ji} \beta_2 + b_i + c_j,$$

where $\lambda_{it,ji}$ denotes the expected number of goals scored by team $i$ at its $t$-th game with match number $j$, $b_i \sim N(0, \sigma^2_b)$ represent team-specific random intercepts and $c_j \sim N(0, \sigma^2_c)$ represent match-specific random intercepts in order to account for the matched pair design with $\tilde{t} \in \{1, 2\}$. This model coincides with the models selected by glmer and glmmLasso with AIC for Model 3. The corresponding fit is easily obtained, e.g. by using the glmer function. The results are presented in Table 3. As expected, the variable MV

| glmmLasso | glmer-select |
|------------|---------------|
| BIC | AIC | LOOCV | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
| M1 | -90.23 | -85.85 | -83.92 | -90.29 | -88.12 | -88.12 |
| M2 | -88.55 | -86.52 | -85.35 | -89.36 | -88.14 | -88.14 |
| M3 | -88.55 | -87.14 | -86.35 | -90.17 | -90.17 | -90.17 |

Table 2: Log-likelihood scores for glmmLasso and glmer-select for Model 1-3 and different levels of sparseness.

has a clear positive effect on the number of goals a national team scores, while the effect of the opponent’s MV is negative. Both effects are clearly significant and the final model from equation (1) yields a rather respectable fit with an “in sample” log-likelihood score of -90.17. Both variables have already been selected in Groll and Abedieh [10], but on the basis of the larger data set of the EUROs 2004-2012, now the maximum number of teammates and the UEFA points of the opponent are not incorporated anymore.
### 3 Explorative Power of the Number CL Players

As already mentioned in the introduction, a second major objective of this article is to investigate, if the number of a national team’s players, which reached at least the half-finals in the preceding UEFA CL season, has a significant influence on the team’s success at the subsequent EURO tournament, if the data of the EURO 2012 are included. On the one hand, we found in Section 2 that the number of CL players for both teams is selected for each setting by `glimLasso` based on LOOCV, but for no other method. On the other hand, there is a substantial positive correlation (0.827) between the MV and the number of CL players for the data based on the EUROs 2004-2012.

The results from Section 2 suggest that the only relevant variable for predicting games of a EURO tournament is the MV of both teams. Though the variable has gained increasing importance and newly approaches for the prediction of the most renowned soccer events have been based on it (see for example Gerhards and Wagner [6, 7], Gerhards et al. [8]), there are some drawbacks, as its realizations are usually based on estimates. Any registered user of the web-site [http://www.transfermarkt.de](http://www.transfermarkt.de), e.g., is allowed to rate the MVs of single international players, and a player’s MV then essentially results as an average of these ratings. Beside the transfer value of a player, the user ratings also cover aspects such as experience, future perspective or prestige of a player. Consequently, there is a certain amount of subjective valuation contained in these estimated MVs and it may be preferable to consider an alternative variable, which is fixed and easy to obtain, such as the number of CL players. Of course, such an alternative variable would only help, if it provides at least almost the same explorative power as the MV. In the following we show that the number of CL players serves a highly precious candidate.

In Table 4 we present the results for the Models 1-3 from Section 2, after excluding the MV from the set of potential covariates. We find that, now, in almost all settings for Model 2 and 3 the number of CL players is selected. In general, the resulting log-likelihood scores presented in Table 5 are almost indistinguishable compared to those in Table 5. Hence, the variable number of CL players seems to be a promising competitor for the MV. Following the selection criteria from the preceding section, we end up with the model

\[
\log(\lambda_{ij}) = \beta_0 + (UEFA \ points \ opp.)_{i,j} \beta_1 + (# \ CL \ players)_{i,j} \beta_2 + b_i + c_j, \tag{2}
\]

which achieves almost the same goodness-of-fit (-90.46) as model (1). The corresponding estimates are presented in Table 6. Again, both effects are clearly significant and, as expected, the opponent’s UEFA points have a negative effect on the number of goals a national team scores, while the number of CL players has a distinct positive effect.

### 4 Conclusion

In the article two major objectives are treated. First, the results of Groll and Abedieh [10] are revised by replicating their analysis on a more reliable data
Table 4: Selected variables for \textit{glmmLasso} and \textit{glmer-select} for Model 1-3 and different levels of sparseness (excluding the MV from the set of potential covariates).

| Model | glmmLasso | glmer-select |
|-------|-----------|--------------|
| BIC   | AIC       | LOOCV α = 0.01 | AIC α = 0.05 | AIC α = 0.1 |
| M1    | ODDS opp. | ODDS opp. | UEFA pts. opp. | UEFA pts. opp. | UEFA pts. opp. |
|       | UEFA pts. opp. | fairness | fairness | fairness |
|       | # CL players | # CL players | # CL players | # CL players |
| M2    | UEFA pts. opp. | fairness | UEFA pts. opp. | UEFA pts. opp. | UEFA pts. opp. |
|       | # CL players opp. | # CL players | # CL players | # CL players |
|       | FIFA pts. | - | - | - |
| M3    | UEFA pts. opp. | # CL players | UEFA pts. opp. | UEFA pts. opp. | UEFA pts. opp. |
|       | UEFA pts. opp. | # CL players | # CL players | # CL players |

Table 5: Log-likelihood scores for \textit{glmmLasso} and \textit{glmer-select} for Model 1-3 and different levels of sparseness (excluding the MV from the set of potential covariates).

| Model | glmmLasso | glmer-select |
|-------|-----------|--------------|
| BIC   | AIC       | LOOCV α = 0.01 | AIC α = 0.05 | AIC α = 0.1 |
| M1    | -90.23    | -85.32 | -84.85 | -89.98 | -87.60 |
|       | -87.60    | -87.60 |
| M2    | -88.52    | -86.09 | -85.33 | -89.21 | -87.82 |
|       | -87.82    | -87.82 |
| M3    | -88.52    | -87.18 | -87.18 | -90.46 | -90.46 |
|       | -90.46    | -90.46 |

Table 6: Estimates for the final model from equation (2) with \textit{glmer}.

| Coefficients | Standard errors |
|--------------|-----------------|
| Intercept    | 0.171           | 0.069           |
| UEFA pts. opp. | -0.236         | 0.071           |
| # CL players | 0.179           | 0.056           |
| $\hat{\sigma}_b$ | $6.94 \cdot 10^{-7}$ |
| $\hat{\sigma}_c$ | $4.73 \cdot 10^{-7}$ |

basis, including the data of the EURO 2012. Still, the MV of both competing teams plays a major role for the teams’ success, but on basis of the larger data set of the EUROS 2004-2012, the variables maximum number of teammates and UEFA points opponent are not incorporated anymore. Secondly, it is shown that the variable number of CL players provides almost the same explorative power with respect to the number of goals scored by national teams at EUROS as the MV, and hence serves as a reliable and competitive alternative.

References

[1] D. Bates and M. Maechler. \textit{lme4}: Linear mixed-effects models using \textit{S4} classes, 2010. URL \url{http://CRAN.R-project.org/package=lme4}. R package version 0.999999-0.

[2] J. B. Carlin, L. C. Gurrin, J. A. C. Sterne, R. Morley, and T. Dwyer. Regression models for twin studies: a critical review. \textit{International Journal of Epidemiology}, B57:1089–1099, 2005.
[3] D. Dyte and S. R. Clarke. A ratings based Poisson model for World Cup soccer simulation. *Journal of the Operational Research Society*, 51 (8), 2000.

[4] M. J. A. Eugster, J. Gertheiss, and S. Kaiser. Having the second leg at home - advantage in the UEFA Champions League knockout phase? *Journal of Quantitative Analysis in Sports*, 7(1), 2011.

[5] T. Frohwein. Die falschen Pferde. In: e-politik.de (08.06.2010), available at: http://www.e-politik.de/lesen/artikel/2010/die-falschen-pferde (12.06.2012), 2010.

[6] J. Gerhards and G. G. Wagner. Market value versus accident - who becomes European soccer champion? *DIW-Wochenbericht*, 24:236–328, 2008.

[7] J. Gerhards and G. G. Wagner. Money and a little bit of chance: Spain was odds-on favourite of the football worldcup. *DIW-Wochenbericht*, 29:12–15, 2010.

[8] J. Gerhards, M. Mutz, and G. G. Wagner. Keiner kommt an Spanien vorbei - außer dem Zufall. *DIW-Wochenbericht*, 24:14–20, 2012.

[9] A. Groll. *glmLasso: Variable Selection for generalized linear mixed models by L1-penalized estimation*, 2011. URL http://CRAN.R-project.org/package=glmLasso. R package version 1.1.1.

[10] A. Groll and J. Abedieh. Spain retains its title and sets a new record - generalized linear mixed models on European football championships. *Journal of Quantitative Analysis in Sports*, 9(1):51–66, 2013.

[11] A. Groll and G. Tutz. Variable selection for generalized linear mixed models by L1-penalized estimation. *Statistics and Computing*, 2012. To appear.

[12] C. Leitner, A. Zeileis, and K. Hornik. Who is going to win the EURO 2008? (A statistical investigation of bookmakers odds). Research report series, Department of Statistics and Mathematics, University of Vienna, 2008.

[13] C. Leitner, A. Zeileis, and K. Hornik. Forecasting sports tournaments by ratings of (prob)abilities: A comparison for the EURO 2008. *International Journal of Forecasting*, 26 (3):471–481, 2010.

[14] C. Leitner, A. Zeileis, and K. Hornik. Forecasting the winner of the FIFA World Cup 2010. Research report series, Department of Statistics and Mathematics, University of Vienna, 2010.

[15] C. Leitner, A. Zeileis, and K. Hornik. Bookmaker consensus and agreement for the UEFA Champions League 2008/09. *IMA Journal of Management Mathematics*, 22 (2):183–194, 2011.

[16] V. Stoy, R. Frankenberger, D. Buhr, L. Haug, B. Springer, and J. Schmid. Das Ganze ist mehr als die Summe seiner Lichtgestalten. Eine ganzheitliche Analyse der Erfolgschancen bei der Fußballweltmeisterschaft 2010. Working Paper 46, Eberhard Karls University, Tübingen, Germany, 2010.

[17] A. Zeileis, C. Leitner, and K. Hornik. History repeating: Spain beats Germany in the EURO 2012 final. Working paper, Faculty of Economics and Statistics, University of Innsbruck, 2012.
Beta kernel regression estimator for censored data: Some comparisons

Zohra Guessoum * & Sara Ghettab **

* Lab. M.S.T.D., Faculté de Math., U.S.T.H.B., BP 32, El Alia, 16111, Algeria. 
zguessoum@usthb.dz
** Lab. M.S.T.D., Faculté de Math., U.S.T.H.B., BP 32, El Alia, 16111, Algeria. 
sara.a3mal@gmail.com

Abstract. We investigate the asymptotic behavior of the beta kernel regression estimator in censored model. Chen (1999) and Bouezmarni et al (2003) introduced and studied the beta kernel density estimator with known compact support $[0, 1]$, for complete data, in order to remove the boundary bias of the standard kernel introduced by Rosenblatt (1956). We compare by simulation the performance of this beta kernel regression estimator with the symmetric kernel, performed by Guessoum and Ould-Said (2008).

Keywords. kernel estimator, nonparametric regression, survival data, censored model, beta kernel.

1 Introduction

In many survival practical applications, censored dependent data appear: in the clinical trials domain (Lipsitz and Ibrahim (2000)), In economic duration data (Wei et al. (1989)), duration study of unemployment (Cai and Prentice (1995))...

Consider a real random variable ($Y$) and a sequence of strictly stationary rv’s $(Y_i)_{i \geq 1}$ with common unknown absolutely continuous distribution function (df) $F$ and let $(C_i)_{i \geq 1}$, be a sequence of censoring rv’s with common unknown df $G$. In contrast to statistics for complete data studies, censored model involves pairs $(T_i, \delta_i)_{i = 1,...,n}$ where only $T_i = Y_i \wedge C_i$ and $\delta_i = I_{\{Y_i \leq C_i\}}$ are observed.

Let $X$ be an $\mathbb{R}^d$-valued random vector. Let $(X_i)_{i \geq 1}$ be a sequence of copies of the random vector $X$ and denote by $X_{i,1}, \cdots, X_{i,d}$ the coordinates of $X_i$. The study we perform below is then on the set of observations $(T_i, \delta_i, X_i)_{i \geq 1}$. In regression analysis one expects to identify, if any, the relationship between the $Y_i$’s and $X_i$’s. This means looking for a function $m^*(X)$ describing this relationship that realizes the minimum of the mean squared error criterion. It is well known that this minimum is achieved by the regression function $m(x)$ defined on $\mathbb{R}^d$ by

$$m(x) = \mathbb{E}(Y \mid X = x).$$

1This work was partially supported by the PNR/MESRS/DGRSDT/ANDRU grants MoDIF 8/u160/996 and MSPDE 8/u160/5126.
Our goal is to establish the uniform strong convergence with rate for the kernel regression estimate for independent data when the symmetric condition on the kernel is relaxed. In particular, we take a beta kernel and we draw some simulations to lend further support to our theoretical results. Then we compare the shape obtained in this case with those obtained for symmetric kernel.

2 Notations, definitions and assumptions

Suppose that \((Y_i)_{i \geq 1}\) and \((C_i)_{i \geq 1}\) are two independent sequences of stationary random variables. Then

\[
m(x) = \frac{\int_{\mathbb{R}} yf_{X,Y}(x,y)dy}{\ell(x)} =: \frac{r_1(x)}{\ell(x)}
\]

with \(f_{X,Y}(\cdot, \cdot)\) being the joint density of \((X,Y)\) and \(\ell(\cdot)\) the density function of \(X\).

A feasible estimator of \(m(x)\) is given by:

\[
m_n(x) = \frac{\frac{1}{nh_n^d} \sum_{i=1}^n \delta_{x_i} K \left( \frac{x-x_i}{h_n} \right)}{\frac{1}{nh_n^d} \sum_{i=1}^n K \left( \frac{x-x_i}{h_n} \right)} =: \frac{r_{1,n}(x)}{\ell_n(x)}
\]

where \(\overline{G}_n\) is the Kaplan-Meier estimator (KME) of \(\overline{G}\), and \(\ell_n(x)\) is the kernel estimator of \(\ell(x)\).

Let \(\tau_H = \sup \{ y, \bar{F}(y) > 0 \}\) and \(\tau_G = \sup \{ y, \bar{G}(y) > 0 \}\) be a upper endpoints of \(\bar{F}\) and \(\bar{G}\), respectively. We assume that \(\tau_F < \infty, \bar{G}(\tau_F) > 0\) (this implies that \(\tau_F \leq \tau_G\)) and that \(C\) and \((X, T)\) are independent.

Let \(C\) be a compact set of \(\mathbb{R}^d\) which is included in \(C_0 = \{ x \in \mathbb{R}^d / \ell(x) > 0 \}\).

First we deal with the standard kernel, studied by Guessoum and Ould-Said (2008), and we recall their first theorem (for \(d = 1\), under the following conditions:

\begin{itemize}
    \item[A1.] The bandwidth \(h_n\) satisfies: \(\lim_{n \to +\infty} h_n = 0, \lim_{n \to +\infty} \frac{nh_n}{\log n} = +\infty\) and \(\log \log n = o\left(\frac{1}{h_n}\right)\) where \(0 < \mu < 1\).
    \item[A2.] The kernel \(K\) is bounded, symmetric and has compact support. It is also Hölderian of order \(\gamma > 0\).
    Furthermore \(\int_{\mathbb{R}} t |K(t)| dt < +\infty\) and \(\int_{\mathbb{R}} t^2 K(t)dt < +\infty\) hold.
    \item[A3.] The function \(r_1(x)\) defined in (9) is twice differentiable and \(\sup_{x \in C} |r''_1(x)| < +\infty\).
    \item[A4.] The marginal density \(\ell(.)\) is twice differentiable and satisfies the Lipschitz condition. Furthermore \(\ell(x) > \Gamma\) for all \(x \in C\) and \(\Gamma > 0\).
    \item[A5.] The integral defined by \(\int_{\mathbb{R}} \frac{y^2}{\ell(y)} f_{X,Y}(x,y)dy =: r_2(x)\) is twice differentiable and \(\sup_{x \in C} |r''_2(x)| < +\infty\).
\end{itemize}
Theorem 1 Under Assumptions A1-A5, we have for $n \to \infty$:

$$\sup_{x \in C} |r_n(x) - r(x)| = O\left(\max\left\{\sqrt{\frac{\log n}{nh_n}}, h_n^{-1}\right\}\right) \text{ a.s.}$$

3 Main result

For $d > 1$ and a non symmetric kernel, the assumptions above become:

**A1)** The bandwidth $h_n$ satisfies:

- $\lim_{n \to +\infty} \frac{nh_n^d}{\log n} = +\infty$,
- $\sqrt{\frac{\log \log n}{n}} = o(h_n)$

**A2)** The kernel $K$ is bounded and satisfies:

- $\int_{\mathbb{R}^d} ||t|| K(t)dt < +\infty$,
- $\int_{\mathbb{R}^d} (t_1 + t_2 + \ldots + t_d)K^2(t)dt < +\infty$ and $\int_{\mathbb{R}^d} K^2(t)dt < +\infty$,
- $\forall (t, s) \in C^2 |K(t) - K(s)| \leq ||t - s||^\gamma$ for $\gamma > 0$.

**A3)** The function $r_1(\cdot)$ defined in (??) is continuously differentiable. and $\sup_{x \in C} \left| \frac{\partial r_1}{\partial x_i}(x) \right| < +\infty$ for $i = 1, \ldots, d$.

**A4)** The function $r_2(x) := \int_{\mathbb{R}^d} \frac{y^2}{G(y)} f_{X,Y}(x, y)dy$ is continuously differentiable. and $\sup_{x \in C} \left| \frac{\partial r_2}{\partial x_i}(x) \right| < +\infty$ for $i = 1, \ldots, d$.

**A5)** $\exists D_1 > 0$ and $\exists D_2 > 0$ such that $\sup_{u,v \in C} |\ell_{ij}(u, v)| < D_1$ and $\sup_{u \in C} |\ell(u)| < D_2$, where $\ell_{ij}$ is the joint distribution of $(X_i, X_j)$.

**A6)** The marginal density $\ell(.)$ is continuously differentiable and $\sup_{x \in C} \left| \frac{\partial \ell}{\partial x_i}(x) \right| < +\infty$ for $i = 1, \ldots, d$. and there exists $\xi > 0$ such that $\ell(x) > \xi \ \forall x \in C$.

Then we have

Theorem 2 Under Assumptions A1, H2, A3-A5, we have

$$\sup_{x \in C} |m_n(x) - m(x)| = O\left(\sqrt{\frac{\log n}{nh_n^d}}\right) + O(h_n) \text{ a.s as } n \to \infty.$$
Proof
The basic idea is to write the process $|m_n(x) - m(x)|$ as

$$
|m_n(x) - m(x)| = \left| \left( \frac{r_{1,n}(x)}{\ell_n(x)} - \frac{\tilde{r}_{1,n}(x)}{\ell_n(x)} \right) + \left( \frac{\tilde{r}_{1,n}(x) - \mathbb{E}(\tilde{r}_{1,n}(x))}{\ell_n(x)} \right) \right|
$$

Then we have

$$
\sup_{x \in C} |m_n(x) - m(x)| \leq \frac{1}{\inf \ell_n(x)} \left\{ \sup_{x \in C} |r_{1,n}(x) - \tilde{r}_{1,n}(x)| + \sup_{x \in C} |\tilde{r}_{1,n}(x) - \mathbb{E}(\tilde{r}_{1,n}(x))| \right. \\
\left. \sup_{x \in C} |\mathbb{E}(\tilde{r}_{1,n}(x)) - r_1(x)| + \sup_{x \in C} (|r_1(x)| \xi^{-1}) \sup_{x \in C} |\ell(x) - \ell_n(x)| \right. \\
\leq \frac{1}{\inf \ell_n(x)} \left\{ I_1 + I_2 + I_3 + \sup_{x \in C} (|r_1(x)| \xi^{-1}) I_4 \right\}
$$

For $I_3$ observe that,

$$
\mathbb{E} \left( \frac{\delta_i T_1}{G(T_1)} | X_1 = u \right) = \mathbb{E} \left[ \frac{Y_i}{G(Y_1)} \mathbb{E} \left[ 1 = I_{\{Y_i \leq C; 1\}} | Y_1 \right] \ | X_1 = u \right] \\
= \mathbb{E} \left[ Y_1 | X_1 = u \right] = m(u)
$$

then we have

$$
\mathbb{E}(\tilde{r}_{1,n}(x)) - r_1(x) = \int_{\mathbb{R}^d} K_d(t) \left[ r_1(x - h_n t) - r_1(x) \right] dt
$$

A Taylor expansion gives

$$
r_1(x - h_n t) - r_1(x) = -h_n(t_1 \frac{\partial r_1}{\partial x_1}(x') + ... + t_d \frac{\partial r_1}{\partial x_d}(x'))
$$

where $x'$ is between $x - h_n t$ and $x$. Then

$$
\sup_{x \in C} |\mathbb{E}(\tilde{r}_{1,n}(x)) - r_1(x)| = \sup_{x \in C} \left| \int_{\mathbb{R}^d} K_d(t) \left[ r_1(x - h_n t) - r_1(x) \right] dt \right|
$$

$$
\leq h_n \sup_{x \in C} \int_{\mathbb{R}^d} \left| K_d(t)(t_1 \frac{\partial r_1}{\partial x_1}(x') + ... + t_d \frac{\partial r_1}{\partial x_d}(x')) \right| dt.
$$
Assumptions A1 i), A2 i) and A3, give

\[
\sup_{x \in \mathcal{C}} |\mathbb{E}(\tilde{r}_{1,n}(x)) - r_1(x)| = O(h_n) \quad a.s.
\]

For \( I_1 \) we can see that

\[
|r_{1,n}(x) - \tilde{r}_{1,n}(x)| \leq \frac{1}{nh_n} \sum_{i=1}^{n} Y_i K \left( \frac{x - X_i}{h_n} \right) \frac{\tau_F}{\tau_F} \frac{\tau_F}{n h_n} \sum_{i=1}^{n} \left| \frac{x - X_i}{h_n} \right|.
\]

Then by using the strong law of large numbers and the law of iterated logarithm (Deheuvels and Einmahl 2000):

\[
\sup_{x \in \mathcal{C}} |r_{1,n}(x) - \tilde{r}_{1,n}(x)| \leq \frac{\tau_F}{G^2(\tau_F)} \mathbb{E} \left( \left| \frac{1}{h_n} K \left( \frac{x - X_1}{h_n} \right) \right| \right) \sqrt{\frac{\log \log n}{n}}.
\]

By Assumptions A1 and A2.

To control \( I_2 \), we use a covering of \( \mathcal{C} \) (since it is compact) by a finite number \( s_n \) of balls \( B_k(x^*_k, a^d_k) \) centered at \( x^*_k = (x^*_{1,k}, \ldots, x^*_{d,k}), \quad k \in \{1, \ldots, s_n\} \). Then for all \( x \in \mathcal{C} \) there exists \( k \in \{1, \ldots, s_n\} \) such that \( \|x - x^*_k\| \leq h_n^\eta \), where \( \eta > \frac{1}{d} + \frac{1}{\gamma} \), (\( \gamma \) is the same as in Assumption A2). Since \( \mathcal{C} \) is bounded there exists a constant \( M > 0 \) such that \( s_n \leq M h_n^\eta \).

Then we set, for \( x \in \mathcal{C} \):

\[
\Delta_i(x) = \frac{1}{nh_n^d} \frac{\delta_i T_i}{G(T_i)} K_d \left( \frac{x - X_i}{h_n} \right) - \mathbb{E} \left( \frac{1}{nh_n^d} \frac{\delta_i T_i}{G(T_i)} K_d \left( \frac{x - X_i}{h_n} \right) \right).
\]

It is obvious that

\[
\sum_{i=1}^{n} \Delta_i(x) = \tilde{r}_{1,n}(x) - \mathbb{E}(\tilde{r}_{1,n}(x)).
\]

Writing \( \Delta_i(x) = \Delta_i(x_k^*) = \tilde{\Delta}_i(x) \), we have clearly \( |\Delta_i(x)| \leq |\tilde{\Delta}_i(x)| + |\Delta_i(x_k^*)| \).

By assumption A2 we have
\[ \sup_{x \in \mathcal{C}} \left| \sum_{i=1}^{n} \Delta_i(x) \right| \leq \sup_{x \in \mathcal{C}} \left( \frac{2 \mathbb{E}(|Y_1|)}{G(\tau_F)} \frac{1}{h_n^d} \left\| x - x^*_k \right\|^{\gamma} \right) \leq \frac{\mathbb{E}(|Y_1|) h_n^{d\gamma}}{G(\tau_F) h_n^{\gamma+d}} \]

Assumption A1 implies that \( \sup_{x \in \mathcal{C}} \left| \sum_{i=1}^{n} \Delta_i(x) \right| = o(1) \) a.s.

On the other hand, for all \( \epsilon > 0 \), we have
\[
\mathbb{P} \left\{ \max_{k=1,\ldots,s_n} \left| \sum_{i=1}^{n} \Delta_i(x^*_k) \right| > \epsilon \right\} \leq s_n \mathbb{P} \left\{ \left| \sum_{i=1}^{n} \Delta_i(x^*_k) \right| > \epsilon \right\} . \tag{3}
\]

Then let \( U_i = nh_n^d \Delta_i(x^*_k) \). We have \( \mathbb{E}(U_i) = 0 \) and \( |U_i| \leq 2\tau H \overline{K} \nu =: M_1 \), where \( \overline{K} \) is the upper bound of \( K \) and \( \nu = \frac{1}{G(\tau_H)} \). In order to apply Bernstein’s inequality, we calculate \( S^2 = \text{Var}(U_i) \), that is
\[
S^2 = \int K^2 \left( \frac{x^*_k - u}{h_n} \right) r_2(u)du - \left( \int K \left( \frac{x^*_k - u}{h_n} \right) r_1(u)du \right)^2.
\]

By a Taylor expansion around \( x^*_k \), under Assumptions A2 ii) and A4, we obtain
\[
S^2 = O(h_n^d).
\]

Then applying Bernstein’s inequality, we have
\[
\mathbb{P} \left\{ \left| \sum_{i=1}^{n} \Delta_i(x_k) \right| > \epsilon \right\} = \mathbb{P} \left( \left| \sum_{i=1}^{n} U_i \right| > \epsilon h_n^d n \right) \leq 2 \exp \left\{ - \frac{\epsilon^2 h_n^d n}{2(c + \epsilon M_1)} \right\} =: A
\]

so
\[
\mathbb{P} \left\{ \max_{k=1,\ldots,s_n} \left| \sum_{i=1}^{n} \Delta_i(x_k) \right| > \epsilon \right\} \leq s_n A \leq M h_n^{-d\gamma} A = M \left( nh_n^d \right)^{-\gamma} n^{-\frac{\epsilon^2}{2(c + \epsilon M_1)}} n h_n^d/\log n. \tag{4}
\]

From Assumption A1 i), the last term in (4) is the general term of a convergent series. Then by Borel-Cantelli’s Lemma the first term in (4) goes to zero almost surely. Now
if we replace $\varepsilon$ by $\varepsilon_0 \sqrt{\log \frac{n}{nh_n^d}}$ we can choose $\varepsilon_0$ so that the term $n^{-\frac{2}{2 + \gamma \log \frac{n}{nh_n^d}}}$ is the general term of a convergent serie and then

$$\sup_{x \in \mathbb{C}} |\tilde{r}_{1,n}(x) - \mathbb{E}(\tilde{r}_{1,n}(x))| = O \left( \sqrt{\frac{\log n}{nh_n^d}} \right)$$

For $I_4$, we use the same steps as in $I_2$ and $I_3$ for getting

$$\sup_{x \in \mathbb{C}} |\ell(x) - \ell_n(x)| = O \left( \sqrt{\frac{\log n}{nh_n^d}} \right) + O(h_n) \text{ a.s}$$

In final, assumption $A_4$ conclude the proof of the main result.

4 Simulation study

First we consider the linear regression model: $Y_i = \Upsilon X_i + \Upsilon_0 + b \varepsilon_i$, for which we simulate $X_i$ and $\varepsilon_i$ as two independent iid sequences distributed as $\mathcal{N}(0,1)$ and where $b$ is a constant appropriately chosen (here we take $b = 0.2$).

We also simulate $n$ iid rv $C_i \sim \mathcal{N}(0,1)$, and then take $T_i = \min(Y_i, C_i)$ with indicative rv $\delta_i = 1 = I\{Y_i \leq C_i\}$

We calculate our estimator based on the observed data $(X_i, T_i, \delta_i), i = 1, \ldots, n$,

- by choosing a Gaussian kernel
  $$K_g(\frac{x - X_i}{h_n}) = \frac{1}{\sqrt{2\pi}} \exp\left(- \left( \frac{x - X_i}{h_n} \right)^2 / 2 \right)$$

- by choosing a beta kernel
  $$K_{x/h_n, (1-x)/h_n}(X_i) = \text{beta}(x/h_n, (1-x)/h_n)(X_i).$$

We compare the shape obtained in the two cases to the true one for 10% percentage of censoring and for a size $n = 100, n = 500$. 

7439
We can see that the behavior of the beta kernel is better on the bounds, which is confirmed by Table 1, where we built the mean square error table of the estimator with respect to the theoretical value. For each values $n$ we replicate $B = 100$ times. In Table 1, we give the MSE for each $n$ at the bounds, that is, $x = 0$ and $x = 1$.

|          | $n = 50$ | $n = 100$ | $n = 1000$ |
|----------|----------|-----------|------------|
| $x = 0$  | $MSE_{Gauss}$ | 1.544     | 1.601      | 1.294      |
|          | $MSE_{Beta}$  | 1.307     | 1.335      | 1.281      |
| $x = 1$  | $MSE_{Gauss}$ | 0.391     | 0.249      | 0.153      |
|          | $MSE_{Beta}$  | 0.275     | 0.228      | 0.060      |

Table 1: MSE for $x = 0$ and $x = 1$ the limit bounds.

We also give, in Table 2, the MSE by taking the median over $x \in [0, 1]$. We note that, in this case, the symmetric kernel have better behavior which is due to a faster speed of convergence.

|          | $n = 50$ | $n = 100$ | $n = 1000$ |
|----------|----------|-----------|------------|
|          | $MSE_{Gauss}$ | 0.1044    | 0.0888     | 0.0748     |
|          | $MSE_{Beta}$  | 0.1219    | 0.1092     | 0.0889     |

Table 2: The median of MSE over $x \in [0, 1]$

Now we consider the case of non linear regression by choosing the following model:

$$Y_i = \sin(aX_i) + \varepsilon_i$$
\( n = 50 \quad \text{and} \quad n = 500 \)

which is the sinus case, with \( \text{perc.cens.} = 10\% \), \( n = 50 \) and \( n = 500 \)

Once again we can notice a best behavior of the beta kernel at the bounds.

**Bibliographie**

[1] Bouezmarni, T. et Rolin, J. M. (2003) *Consistency of the beta kernel density function estimator*. The Canadian J. of Statis. Vol 31, n1, 89 – 98.

[2] Cai, J., Prentice, R. L. (1995). *Estimating equations for hazard ratio parameters based on correlated failure time data*. Biometrika, Vol. 82 n1, pp 151 – 164.

[3] Chen, S. X. *Beta kernel estimators for density functions* (1999) Computational Statistics & Data Analysis 31, 131-145.

[4] Guessoum, Z. et Ould Saïd, E. (2008). *On the nonparametric estimation of the regression function under censorship model*. Statist. & Decisions 26:159-177.

[5] Lipsitz, S. R. et Ibrahim, J. G. (2000). *Estimation with correlated censored survival data with missing covariates*. Biostatistics. 1: 315-327.

[6] Wei, L. J., Lin, D. Y. (1989) *Regression analysis of multivariate incomplete failure times by modelling marginal distribution*. J. Amer. Statist. Assoc. 84: 1064-1073.
Neural network modeling and fractional model identification of a thermal process

Sofiane Hammouche, Rabah Mellah, Tounsia Djamah, and Rachid Mansouri

Laboratoire de Conception et Conduite des Systemes de Production (L2CSP), University M. Mammeri of Tizi Ouzou, Algeria
(E-mail: sofham10@yahoo.fr, mellah_rabah@yahoo.fr, djamah.jummo@yahoo.fr, Rachid_mansouri.jummo@yahoo.fr)

Abstract. Diffusive phenomena exhibit complex dynamics that can be described by fractional models; a typical example is the heat transfer process. Its is characterized by the hereditary property and its internal behaviour contains unknown non-linearities, difficult to model and varying with the operating frequency band. The aim of this paper is the modeling and identification of the thermal process. To achieve this goal, two models are investigated: a fractional order model and a neural network (NN). Numerical simulations on a thermal benchmark are performed in order to analyze the two models fitting ability to approximate the diffusive phenomena.

Keywords: Modeling, identification, heat transfer, diffusive process, fractional order system, multi-layer neural network.

1 Introduction

Modeling and identification of a complex dynamical system is an important task for its simulation, control design, prediction and fault diagnosis. Various non-linear model structures have been proposed in the literature such as Volterra series, Wiener and Hammerstein models and neural networks based models [1–3]. On the other hands, studies of diffusive phenomena, such as thermal [4]... exhibit complex fractional behaviour characterized by long memory transients and infinite dimensional structure, and the use of classical integer models is inappropriate; thus, a further class based on the fractional derivative called fractional models has been developed for their modeling.

The well known heat transfer process is addressed in this paper. Different approaches based on fractional systems have been proposed for its modeling [4–6]; although, they give satisfactory approximation in the time domain, for some methods, a priori choice of 0.5 on the fractional order is required [4]. Another major limitation evolves, when temperature ranges in a wide interval, causing that certain thermophysical parameters of the medium (conductivity and diffusivity) to vary and depend on the temperature. In this context, the process internal behaviour contains unknown non-linearities, difficult to model and not fully understood, and the use of intelligent computational methods such as neural networks may be attractive. They offer the advantages of the learning and adaptation mechanisms to adjust the non-linear model parameters, while conventional approaches tend to fail or become cumbersome.

Artificial Neural Networks have gone through a rapid development and grown past the experimental stage to become implemented in a wide range
of engineering applications [3]. It is a parallel distributed information processing system which provide a distinctive computational paradigm by exploiting the massively parallel processing performed by the neurons.

Thus, a neural network is a strong tool for replacing complex non-linear dynamics with a simple numeric database [7], [8]. However, (NN) are notoriously difficult to train, especially when the system dynamics involve many changing factors such as unstructured uncertainties, and the NN parameters have to be adjusted, to composite for these effects. The use of the powerful learning algorithm, in occurrence Levenberg-Marquardt algorithm for training the networks, allows to overcome this drawback. This algorithm is a blend of the steepest descent method and the Gauss-Newton algorithm, and it inherits the stability of the former and the speed advantage of the latter [9].

The aim of this paper is to present two methods for the modeling of the heat transfer process. The first one is based on parsimonious fractional models, with no a priori choice on the fractional order, and the second one uses the multi-layer neural networks to reproduce the thermal process behaviour. These methods are applied to a thermal benchmark and their fitting ability to approximate the diffusive phenomena is analyzed.

The paper is organized as follows: Section 2 recalls the heat transfer process properties, and its fractional behavior is pointed out. In section 3, some preliminaries on the fractional models are introduced, and the identification method for fractional systems is presented. Section 4, is dedicated to the modeling with neural network and basic definitions concerning the multi-layer neural networks are given. In Section 5, the thermal benchmark is described, and the simulation results obtained using the fractional model and the neural network are analyzed. Finally, we conclude this paper with the evaluation of the two methods efficiency, and some remarks on future research.

2 Dynamical Behaviour of a Thermal Process

One dimension heat transfer in a plane surface of thickness L is governed by the following partial differential equation [5]:

\[
\begin{cases}
\frac{\partial T(x, t)}{\partial t} = \alpha_c \frac{\partial^2 T(x, t)}{\partial x^2}, & 0 < x < \infty, \ t > 0 \\
\lambda_c (\frac{\partial T(x, t)}{\partial x}) = \phi(x, t) & \text{when} \ x = 0, \ t > 0
\end{cases}
\]  

(1)

Where the finite medium is homogeneous (thermal conductivity \(\lambda_c\), diffusivity \(\alpha_c\), \(x\) being the abscissa of the measurement slot in the medium. The boundary conditions are:

\[
\begin{cases}
\phi(0, t) = u(t) \\
\phi(L, t) = T(L, t)/R
\end{cases}
\]

A dynamical model, linking the temperature \(T(x, t)\) and the heat flux \(\phi(x, t)\) at the diffusion interface \((x = 0)\) is described by the transfer function:
Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013)
International Conference, Mataró (Barcelona), Spain 25 - 28 June 2013

\[ H_p(s) = \tanh \left( \frac{L}{S_{m} \lambda_c} \sqrt{s} \alpha_c \right) \]

\[ \lim_{s \to \infty} H_p(s) = \frac{1}{S_{m} \lambda_c} \sqrt{\alpha_c} \]

\[ \lim_{s \to 0} H_p(s) = \frac{L}{S_{m} \lambda_c} \frac{1}{1 + \frac{L}{\alpha_c} s} \]

At low frequencies an integer model is sufficient to describe the diffusion interface, while, for high frequencies, it behaves like a fractional integrator of order 0.5; thus fractional models are appropriate to describe the diffusive interface behaviour in this frequency band [4, 6]. Also, from the study of thermodynamics, it can be shown that some thermophysical coefficients are directly related to specific state vector components, thus are time varying parameters which induce non-linear uncertainties. In order to circumvent this difficulty, two models can be considered: a fractional model with at least two fractional orders to composite with the effects of the varying parameters with the frequency band [6, 10], and a neural network known to be able to picture the unknown uncertainties of dynamical systems.

3 Fractional order models

Fractional calculus is the generalization of integration and differentiation to fractional (non integer) order \( \alpha \), where \( D^\alpha f(t) \) is the fractional derivative of order \( \alpha \) of the function \( f(t) \), and \( r - 1 < \alpha < r \), with \( r \in \mathbb{N} \) [5, 11–13]. A fractional single input \( u(t) \), single output \( y(t) \), linear system can be described by the following fractional state space model [13], with: \( x \in \mathbb{R}^n \):

\[
\begin{align*}
D^{(\alpha)}(x) &= Ax + Bu \\
y &= Cx + Du
\end{align*}
\]

with \( D^{(\alpha)}(x) = [D^{\alpha_1}x_1, \ldots, D^{\alpha_n}x_n]^T \) (3)

The matrices \( A, B, C, D \) are of appropriate dimensions. In the particular case, where all the states \( x_i(t) \) are differentiated to the same order, the system is called a commensurate order system with: \( D^{(\alpha)}(x) = D^\alpha [x_1 x_2 \ldots x_n]^T \)

3.1 Fractional system identification

Given the fractional model state space representation (3), the controllable canonical form (4) can be considered, with the matrices \( A, B, C, D \) as follows:

\[
A = \begin{bmatrix}
0 & 1 & \ldots & 0 \\
\vdots & & & \\
0 & 0 & \ldots & 1 \\
-a_1 & -a_{n-1} - a_n \\
\end{bmatrix} \quad B = \begin{bmatrix}
0 \\
\vdots \\
0 \\
1 \\
\end{bmatrix} \quad C = [c_1 c_2 \ldots c_n] \quad D = [d] \quad (4)
\]
The objective of the identification consists in estimating the model matrices \( A, C, D \), with: \( \hat{\theta} = [a_1 \ldots c_1 \ldots c_n d] \) and the non integer order \( (\alpha) = [a_1 \ldots a_n] \). The vector \( \theta \) of parameters to be estimated is given by: 

\[ \hat{\theta} = \theta(\alpha) \]

The model output being non-linear in \( \theta \), an output error method based on the Levenberg-Marquardt (LM) algorithm has been developed in preceding work [6,10]. The sampled data set is composed of \( K \) observations \( \{u_k, y_k^*\} \). Let \( \hat{\theta} \) be the estimate of the exact parameters vector \( \theta \), the output prediction error is given by: 

\[ \varepsilon_k = y_k^* - \hat{y}_k, \]

where \( \hat{y}_k \) is the corresponding output estimate. The optimal value of \( \hat{\theta} \) is obtained by minimizing the quadratic criterion using LM algorithm [14–16]: 

\[ J(\theta) = \frac{1}{2} \sum_{k=1}^{K} \varepsilon_k^2 \]

with the update rule:

\[ \theta^{(i+1)} = \theta^{(i)} - \left[ J''_{\theta\theta} + \lambda I \right]^{-1} J'_\theta \]

\[ J'_\theta = -2 \left( \sum_{k=1}^{K} \varepsilon_k \sigma_{k/\theta} \right) \]

\[ J''_{\theta\theta} = -2 \left( \sum_{k=1}^{K} \sigma_{k/\theta} \sigma_{k/\theta}^T \right) \]

\[ \sigma_{k/\theta} = \frac{\partial \hat{y}(k, \theta)}{\partial \theta} \]

with \( \lambda \) a monitoring parameter. This algorithm is based on the calculation of the gradient and Hessian, which depend on the sensitivity functions. A multivariable fractional order model [6, 10] is used to compute these functions with respect to \( \hat{\theta} \); the fractional order \( (\alpha) \) sensitivity is calculated numerically.

\section{4 Neural network modeling}

An artificial neural network is a powerful data-modeling tool that is able to capture and describe NL input/output (I/O) relationships of a complex system. The motivation for the development of NN technology stemmed from the desire to develop an artificial system that could perform "intelligent" tasks similar to those of the human brain [7,8].

Given a set of inputs and desired outputs or targets \( (u(t-j), j = 1, n_u \) and \( y(t-i), i = 1, n_y ) \) of the NL system (6).

\[ y(t) = f(y(t-1),...,y(t-n_y),u(t-1),...,u(t-n_u)) \] (6)

The aim is to approximate the underlying dynamics \( f \) using NN. This modeling task can then be formulated: \( \hat{y}(t) = \hat{f}(x(t)) \), with \( x(t) \) the NN input:

\[ x(t) = [y(t-1)...y(t-n_y)u(t-1)...u(t-n_u)]^T \] (7)

The NN consists of many interconnected units or neurons, organized in layers and operating in parallel to attain the desired targets. The last layer is the NN output layer and all the others below it are called hidden layers. The behaviour of a NN depends on both the weights \( w^{(k)}_{ij} \) which define the strength of connection between the nodes and the (I/O) activation function \( a(z) \) that is
specified for the units. This last can be chosen as hyperbolic tangent, sigmoid...

The relationship for a generic \( i^{th} \) neuron, in the \( k^{th} \) layer, is expressed as:

\[
\begin{align*}
\zeta_i^{(k)} &= \sum_{j=1}^{(n_k-1)} w_{ij}^{(k)} x_j^{(k-1)} + b_i^{(k)} \\
a(z) &= \tanh(z) = \frac{1 - \exp(-2z)}{1 + \exp(-2z)}
\end{align*}
\]

where \( b_i^{(k)} \) are the node connection threshold (bias). The weights are adapted during use to yield good performance and the NN acquires knowledge through learning (training) which is the process of 'capturing' the unknown information.

### 4.1 Learning Algorithm

The process of learning involves tuning the weights and biases values of the network such that the outputs match the desired targets. A class of learning algorithms such as steepest descent methods (gradient), backpropagation algorithm... [9, 16] have been developed to optimize the performance function. They are guaranteed to converge but their convergence can be very slow. On the other hand, second order methods based on the Hessian such as Gauss-Newton methods are much more efficient in terms of convergence; however, they require more computation per iteration, and an initialization around the optimum is necessary. The use of the powerful Levenberg-Marquardt algorithm allows to overcome these drawbacks. This last blends the steepest descent method and the Gauss-Newton algorithm, and it inherits the stability of the former and the speed advantage of the latter [9, 16]. Its basic idea is that it performs a combined training process: around the area with complex curvature, it switches to the gradient algorithm (with \( \lambda \) very large), until the local curvature is proper to make a quadratic approximation; then it becomes the Gauss-Newton algorithm (with \( \lambda \) very small), which can speed up the convergence significantly.

Let \( \theta \) be the parameters vector containing all the weights and thresholds of the NN, and \( \varepsilon_k \) the output prediction error between the target and the NN output, the optimal value of \( \theta \) is obtained by minimizing the quadratic criterion:

\[
J(\theta) = \frac{1}{2} \sum_{k=1}^{N} \varepsilon_k^2 \text{ where } \varepsilon_k = y_k^d - \hat{y}_k
\]

The update rule of LM algorithm is given in equation (5), the Hessian matrix which is the second-order derivatives of the performance function has to be calculated. It is approximated using the Jacobian matrix \( J_a \) as follows:

\[
H \approx J_a^T J_a.
\]

This algorithm is used for the heat transfer modeling.

### 5 Modeling and identification of the thermal process

#### 5.1 Description of the thermal benchmark

The heat transfer in an aluminium rod is considered. The input signal is a thermal flux applied at one end, and the output signal is the rod’s temperature measured at a distance \( d = 0.5 \) cm from the heated point. The input: a pseudo-random binary sequence and the output sets are depicted in Fig.1 and
Fig.2 respectively; a first parts of the data is used for identification and the second for validation. As shown by the theory, the thermal process requires for its description different models, depending on the frequency band. In this purpose, two models are investigated: a fractional model and a multi-layer neural network model. The simulation results are given in the next subsection.

5.2 Fractional model Identification of the thermal process

A fractional state space model of dimension two, with $(\alpha) = [\alpha_1 \alpha_2]$ is used to describe the thermal benchmark and its parameters are estimated using the presented identification method. The model output response depicted in Fig.3, fits the data with a good accuracy and the prediction error $(\varepsilon \simeq 0.3)$ is plotted in Fig.4. The fractional order vector: $(\alpha) = [0.54 \ 0.93]$, characterizing the diffusion process is recovered, with a quadratic criterion $J = 5.11$. The use of non commensurate orders ($\alpha_1 \neq \alpha_2$), gives a supplementary degree of freedom that allows to describe the effects of the frequency bands. This highlights the interest of fractional models to describe the complex thermal behaviour.

5.3 Neural network modeling of the thermal benchmark

The architecture of the proposed multi-layer NN is shown in Fig.5. It is a three layer architecture with eleven nodes in the input layer (one of control and ten past output values of the system); the hidden layer contains three neurons and the output layer contains one neuron. An hyperbolic tangent transfer function (th) is used as the activation function for all the neurons of the hidden layer.
The weight matrices are initialized with random values and are adjusted during training based on LM algorithm.

The thermal system database is split in two parts, the first one is devoted to the system learning, while the second is used to the system validation. The NN training sequence of the thermal process and the validation sequence are depicted in Fig.6 and Fig.7 respectively. We see that the output of the proposed NN matches accurately the thermal process data for the two phases: learning and validation. The NN prediction error of the test sequence is plotted in Fig.8, and the tracking performance is satisfactory with a value of order \( \approx 10^{-3} \). From these results, the NN potential to model adaptively the thermal process behaviour is verified and the use of LM algorithm has efficiently speed up the NN training phase.

6 Conclusion

In this paper we have investigated two methods for the heat transfer process modeling. A first system of fractional order of dimension two, to composite for the frequency effect has been tested. The simulation results show that the model matches the thermal data with a good accuracy, and the estimated fractional orders, characteristics of the diffusive process are in adequation with the theory. A second model based on the multilayer neural network approach
has been used; The LM algorithm adjusts the NN parameters and optimizes the convergence rate of the training phase.

The simulation results show the efficiency of the neural network model to adapt its internal structure to achieve the fitting of the benchmark behaviour. Thus, we verify that a multilayer neural network can be a universal approximator for complex non-linear systems, in this case with fractional behaviour.

The comparison of the two methods, shows that a better performance ($\varepsilon \approx 10^{-3}$) is obtained for the NN than that of the fractional model. However, the drawback of this approach, is that the NN model parameters cannot give any information about the physical system, in opposition to the fractional model that relates its parameters (orders) to the real process characteristics.

The studied benchmark outputs (temperatures) are distributed on a relatively limited range, so that the thermophysical properties of the material vary very slightly. Thus, a further step will be to test these two models efficiency for an industrial thermal process where the temperatures range in a wide interval.

References

1. R. Haber and L. Keviczky, *Nonlinear System Identification, Input Output Modeling Approach*, Kluwer Academic Publishers; 1999.
2. A. Isidori, *Nonlinear Control Systems*, Springer Verlag Berlin, Heidelberg; 1989.
3. H. White, Learning in artificial neural networks: A statistical perspective. *Neural Computation*, Vol. 1, 1989, pp 425-464.
4. A. Benchellal, T. Poinot and J.C. Trigeassou, Approximation and Identification of Diffusive Interfaces by Fractional Systems, *Signal Processing*, Vol. 86, (10), 2006, pp. 2712-2727.
5. K. B. Oldham, J. Spanier, *The Fractional Calculus*, Academic Press, N.Y, 1974.
6. T. Djamah, R. Mansouri, S. Djennoune and M. Bettayeb, Heat Transfer Modeling and Identification using Fractional Order State Space Models. *J. European des Systemes Automatises*, vol. 42, 2008, pp. 939-951, Lavoisier, Hermes sciences.
7. K. Hornik, Some New Results on Neural Network, *Neural Networks*, Vol. 6, 1993, pp. 1069-1072.
8. G. Dreyfus, *Neural Networks: Methodology And Applications*, Springer Verlag; 2005.
9. B. M. Wilamowski and H. Yu, Improved computation for Levenberg Marquardt training, *IEEE Transactions on Neural Networks*, Vol. 21, pp. 930-937, 2010.
10. T. Djamah, S. Djennoune, M. Bettayeb, Diffusion Process Identification in Cylindrical Geometry using Fractional Models. *Physica Scripta*, IOP journals, 2009.
11. S. G. Samko, A.A. Kilbas and O.I. Marichev, *Fractional Integrals and Derivatives*, Gordon and Breach Science Publishers; 1993.
12. K. S. Miller and B. Ross, *An Introduction to the Fractional Calculus and Fractional Differential Equations*, A Wiley Interscience Publication, San Francisco; 1993.
13. A. Oustaloup *La Derivation Non Entiere: Theorie, Synthese et Application*, Editions Hermes, Paris 1995.
14. L. Ljung, *System Identification: Theory For the User*. Prentice Hall, London; 1987.
15. D. W. Marquardt, An Algorithm for Least Squares Estimation of Non-Linear Parameters. *J. of the Soc. Industrial and Applied Mathematics*. Vol. 11, (2), 1963, pp. 431-441, Paris France.
16. Y.Q. Zhang and A. Kandel, Compensatory Neurofuzzy Systems with Fast Learning Algorithms, *IEEE Transactions on Neural Networks*. Vol.9, (1), pp. 83-105, 1998.
Fantasy Leagues and Decision making under uncertainty. An approach

Javier J. Hernández G., María J. García G., and José G. Hernández R.

1 Gerencia de Investigación, Minimax Consultores, Bristol England
   (E-mail: Minimax@cantv.net)
2 Gerencia General, Minimax Consultores, Caracas Venezuela
   (E-mail: Marimininagarcia@yahoo.com)
3 Departamento de Gestión de la tecnología, Universidad Metropolitana, Caracas Venezuela
   (E-mail: jhernandez@unimet.edu.ve)

Abstract. A new aspect of the sports business is the fantasy leagues. These fantasy leagues where participants compete against other fans, following the performance of athletes in real life, making them feel within the pitch, although each one of them is in different places, very far from the sports arena. Among the problems to be solved by each participant in a fantasy league is your squad of players. The creation of this squad is usually very complicated in sports like soccer where you must satisfy many constraints as: Budget, teams, positions, among others. If one takes into account only the fantasy leagues performance of a given player, the expected performance of that player for next season can be seen as a problem of decision making under uncertainty. Thus a hierarchy of the best players can achieve by using a model of decision making under uncertainty, in particular The Amplitude Model (TAM). The way to use this model raises the aim of this work: To present The Amplitude Model and how to use it to create a squad of players to participate in a fantasy league.

Keywords: Fantasy League, Sport, Soccer, Making decision under uncertainty, The Amplitude Model (TAM).

1 Introduction

Sports are a great business (Dwyer[5]; Pillay et al.[19]), is a growing business, but its growth is not only limited to the courts, in actuality most sports have captivated a large group of followers, that usually move large amounts of money through the fantasy leagues (Burke[2]; Dwyer and Kim[6]; Kwak et al.[15]; Massari[17]).

To participate in the fantasy leagues fans are converted into the owners, managers or coaches of their fantasy leagues teams (Davis and Duncan[3]; Duncan[4]; Wann et al.[20]) by which its interest in the sports is increased and thus there is a market expansion.

Nevertheless these fans, which by participating in the fantasy leagues must continually be making decisions, in most cases have no training in the decision making processes. One of the first decisions that the participants of the fantasy leagues should make it is to select the players that will form its squad.

To help these fantasy league fans in this and all its decisions, the experts analyze the possible performance of each of the sportsmen who will be in the court and they do recommendations to the players of fantasy leagues.
Perhaps one of the most difficult sports to perform recommendations is soccer, mainly because the performance of a player depends heavily on many circumstances, some related to himself, with his own colleagues and even with the team and the players which is facing.

Thus to find a good model that permit to complete a squad of players, that have high performance is not easy task. Especially because not only one player is expected to contribute fantasy leagues points, but besides it has a budget to invest, players of different positions should be integrated and in general of different teams.

One way to explore is to view this selection as a decision making under uncertainty process and use for that The Amplitude Model (TAM), which takes into account the dispersion of the data at the moment to select the best alternative (García et al.[10], [11]; Hernández and García[13]). In this case the players will be the alternatives and states of nature are given by the possible variations of their performance over the previous season. This integration of the squad to participate in a fantasy league represents the objective of this work: To present The Amplitude Model and how to use it to create a squad of players to participate in a fantasy league.

To reach this objective it was followed the Integrated-Adaptable Methodology for the development of Decision Support System (IAMDSS, in Spanish, Metodología Integradora-Adaptable para desarrollar Sistemas de Apoyo a las Decisiones [MIASAD]). In its condition of adaptability this methodology permits to take the steps which are necessary to reach the objectives (García et al.[9]).

In similar way to the followed in other works (García et al.[9],[10],[11]; Hernández et al.[14]) for the present study have been taken some of the most prominent steps of the methodology:

a) To define the problem, this as stated in the objectives is to create a squad of players to participate in a fantasy league, based on TAM,

b) To prepare the first prototype, where the users of the final product, the main readers of the article, are identified. Who are all persons interested in the application of quantitative tools in ludic that will join those soccer fans typically involved in fantasy leagues. Also there was established the structure of the article, which in addition to the introduction will consist of three central chapters, in the last and main one of them the steps will be shown to conform the squad,

c) Obtaining data, in this case on soccer, fantasy leagues and TAM,

d) Establishing the alternatives, that would be the different options that it could be used TAM to conform the squad of a participant in a fantasy league,

e) Evaluate alternatives, according to the performance expected of the squad and to the facility of implementation,

f) Selecting the best alternative, as product of previous evaluation process, and based on the secondary objectives, tacit or explicit, being considered,

g) Implementing the best alternative, illustrating through a simple example that the model chosen permits to select an “adequate” squad of players and
h) Establishing controls, Establishing the mechanisms, probably indicators, that permit to recognize if the alternative selected, continues being valid in the course of the time.

With regard to limitations and reaches, especially by reasons of space, in this paper does not give greater details on the criteria used by the experts to get the payments expressed in TAM. It will be a stronger emphasis on how TAM can be a good starting point for the elaboration of the squad of the participant in the fantasy league.

2 The Fantasy Sports

Although Oates[18] comments that the fantasy leagues arises between sports journalists and publicists in the decade of the 60s, it can be said that The Fantasy Sports Leagues have their origins in the eighties. The growth of the fantasy leagues is increasing, already for the year 2003, as referred Davis and Duncan[3] and ratifies Duncan[4], more of 15.2 millions of adult Americans participated in the fantasy sports. Although this amount of followers according to Lee[16]. Practically doubled in six years, to go to near the 27 million in 2009, however, this number exceeds 31 million, according to Akilli[1], based on a source from 2008. But the growth of the fantastic leagues is not unique to United States; its impact has come to other countries and to many sports. In addition, the fantastic leagues are used even in education (Gillentine and Schulz[12]), from there it is necessary to know what a fantasy league.

It has a fantastic sports definition: “A fantasy sport is a game in which players (i.e., public individuals) compete against one another by maintaining rosters of actual professional athletes and using the real game statistics of those athletes to score points” (Massari[17], 444-445).

And even directly spoken of fantasy leagues: “Fantasy sports leagues are one way fans can enjoy their favorite sports away from the stadium or arena. A fantasy sports league is made up of a dozen or so participants who compete against each other based on statistics from real-world competitions” (Farquhar and Meeds[8], 1208).

In addition, these authors state that fantasy leagues normally begin with a draft of some sort, where the participants (owners) either select their players or are randomly assigned players (Farquhar and Meeds[8]).

This decision to form his team is the first and perhaps most important activity a participant in a fantasy leagues.

Although a large number of participants in the fantasy leagues, not usually have a quantitative background and in decision making processes, the participate in one of them, entails a series of decision making, most of them chained with each other.

Among the processes most important of decisions making which must resort to a participant in a fantasy league is, as just noted, the selection of its initial squad of players.
It is in this aspect, the selection of the initial squad will be the main focus of this work. And although it will seek to make this selection in such a way that it can be adapted to the great generality of the leagues, the illustration will be made through a sport and a specific league.

In particular this work will focus on soccer and although the model will try to be built as general as possible, here it will be used La Liga (Spanish football).

For the conditions that regulates the fantasy leagues it will be followed the ones set on “Desafío la liga”, of which not greater comments will be given and it is recommended that its rules be revised in ESPN[7], of them, by being essential in this work it will be indicated: There will be use up to 15 players, the total investment cannot exceed US$ 100 millions, you cannot have more than three players from the same team and with regard to positions will not have more than 2 goalkeepers, 5 defenders, 5 midfield players and 3 forwards.

These rules seek to implement an approach to decision making under uncertainty, by which subsequently some comments are done, on decision making under uncertainty, in particular on The Amplitude Model.

3 The Amplitude Model (TAM)

What characterizes a problem decision making under uncertainty is known states of nature (SN), but not its probabilities. The interception of the alternatives with these SN produces some payments, that they are reflected in the matrix of payments (MP) on which the analysis is done.

It is important to revise that this MP have not alternatives dominated, neither dominant alternatives, but in case to exist these seconds, they would be without doubt the winning alternative; since an alternative is dominant when their payments are equals or over all the payments of the remaining alternatives and at least one of them is strictly greater than.

The Amplitude Model (TAM, in Spanish El Modelo de Amplitud [EMA]) Emerges as a reinforcement to two of the traditional models for decision making under uncertainty, Hurwicz and Laplace (Hernández and García[13]), taking from this last one the expected value and the first philosophy of working with a multiplicative factor.

The relevant characteristic of TAM is that takes into account the dispersion of the data and use the multiplicative factor β to integrate the dispersion of the data in the analysis. This dispersion of the data is represented by the amplitude, which at the same time gives name to the model.

For problems of maximization, in its general expression, TAM choose the optimal alternative (Alternative*) following (1).

\[
\text{Alternative*} = \frac{\text{Alternative (i)}}{\text{Maxi } \{\text{VEi} + (\beta \times \text{VEi} - (1 - \beta) \times \text{Ami})\}} \quad (1)
\]

Where VEi is the expected value calculated through the expression of Laplace, Ami the amplitude of the alternative and β an multiplicative factor whose value should be around 0.8 (García et al.[10],[11]).
For the case to minimize the expression of TAM will be very similar to (1), only that the act the amplitude as penalization, instead of subtracting the factor that includes it, will be added.

4 The Amplitude Model to conform a squad in fantasy soccer

As indicated above, the players will be the alternatives and to the states of nature (SN) will be used the expected performance in this campaign, compared to the previous season. In the table 1, the five SN are show, that the player have a performance: very low (MB); low (BA); equal (IG); superior (SU) o very superior (MS) to their previous performance.

This SN they will obey major injuries, personal and team problems; minor injuries, bad starter of the team, in the two first cases and to maturity of the player, good start of the team or very good start of the team, full development of the player, in the two last states and for the state IG, that the changes be smaller. In every case the payments will be a percentage of the performance of the previous season, date that are usually available or can be acquired across the experts in fantasy leagues.

The latter performances, which are expressed in points, become payments and take on a base 100, to have a more homogeneous and to calculate the value of each player simply applies (2).

\[ P_{ty_i} = \left( \frac{Pts_i}{Pt_{max}} \right) * 100 \quad (2) \]

Where \( P_{ty_i} \) represents the points in base 100 of the player \( i \), \( Pts_i \) the points that player \( i \) had achieved in the previous fantasy league and \( Pt_{max} \) the points that reached the player, which is still available, with the greatest amount of points achieved in the previous season.

For payments to be used by TAM, will work with a percentage of the \( P_{ty_i} \) and just as it indicate Zadeh[21] it will be used precision in meaning and imprecise in value, since it will be handled the following percentage ranges for each state of nature:

- For MB until 70% under the payment of the previous season.
- BA until 30% under the payment of the previous season.
- IG until 10% over or under previous season.
- SU until 30% above the payment of the previous season.
- And for MS until 50% above the payment of the previous season.

Note that performance is not considered less than 70%, because a player with too low performance, for example suffering a serious injury beginning the season, can be substituted.

Equally it is not expected that a player increases its performance beyond 50% from last season value, even if is a matter of returning after serious injury, in which case basis points for their performance will be taken before the injury.
For players who do not have history, either by promotion or come from the quarry or other league, can take a range between 75% and 125% of the midpoints to its position, or estimate it by players of the same position and that have similar price in the fantasy leagues.

The prices of each player in the fantasy leagues, give some measure of what is hoped it will be its performance.

If all players are taken in their initial value and a fixed value for different SN, in general it would be working with dominant alternatives.

To avoid this permanent dominance at the time to calculate the payments, for each player are handled different percentages, in general the better players will tend to be penalized more hardly, simultaneously that are rewarded more softly.

This is justified, on one hand because a low performance of an excellent player, in the fantasy leagues will have much greater impact than a normal player and moreover for a great player is very difficult to overcome their great performance and not expected so excessively high values, with respect to its previous achievements.

Equally forwards and goalkeepers sometimes tend to contribute most points in fantasy leagues than midfields and defenders, from there to a low performance of goalkeepers and especially the forward may be more costly in points for fantasy leagues squad.

In the table 1, the final percentages for four players are reflected, whose points for the fantasy leagues, in the previous season (Pts), they are here expressed:

A forward (J01034) with 250 points, a midfields (J01193) with 102.5 points, a goalkeeper (J01111) with 115 points and a defender (J02112) with 95 points.

To facilitate information management, all players are identified with the letter J followed by five digits, the first two identify the player, usually according to their performance against other teammates, the second two digits represent the team and go according to alphabetical order from the 01 to 20 and the last digit represents the position: Goalkeeper (GK) = 1, Defender (Df) = 2, Midfields (Mi) = 3 and Forward (Fo) = 4.

From there J01034 (J 01 03 4), represents the first player (01) of the team 03, whose position is Fo (4); J01193 (J 01 19 3), corresponds to the first player (01) of the team 19, which plays as Mi (3); J01111 (J 01 11 1) correspond to a goalkeeper (1) of the team 11 and In this case, it is the first player (01) of the team, while J02112 (J 02 11 2) identifies the second player (02) of this team 11, that is a Df (2).

Although all they had an excellent performance in their respective positions, nevertheless their points in base 100 (Ptyi) are: J01034 = 100, J01193 = 41, J01111 = 46 and J02112 = 38.

Where it is reflected that there are positions, in particular the forward, which for the fantasy leagues generate more points than others.

For the values of the SN to be carried to the table 1, based on the Ptyi and in accordance with previously indicated, they will come given for:

\[
\text{For MB: } [\text{Ptyi} \times (1 - 0.70 \times \text{[Ptyi/100]})] \times \text{Ffi} \tag{3}
\]
\[
\text{For BA: } [\text{Ptyi} \times (1 - 0.30 \times \text{[Ptyi/100]})] \times \text{Ffi} \tag{4}
\]
For IG: \[ \text{Ptyi} \times (1 + 0.10 \times \left\lfloor \frac{\text{Ptyi}}{100} \right\rfloor) \times \text{Ffi} \text{ If Ptyi} \geq 50 \] (5)
For IG: \[ \text{Ptyi} \times (1 - 0.10 \times \left\lfloor \frac{\text{Ptyi}}{100} \right\rfloor) \times \text{Ffi} \text{ If Ptyi} < 50 \] (6)
For SU: \[ \text{Ptyi} \times (1 + 0.30 \times [1 - \text{Ptyi}/160]) \times \text{Ffi} \] (7)
For MS: \[ \text{Ptyi} \times (1 + 0.50 \times [1 - \text{Ptyi}/175]) \times \text{Ffi} \] (8)

Where Ffi be a factor that the fantasy leagues fanatic apply to player i, on each estimated realized, this factor will have a rank between 0.95 and 1.05.

The idea is to achieve a greater differentiation when the players have obtained very similar points; this factor is justified, because in general the fantasy leagues participant usually has his favorite team and favorites players, even to the point of not taking a player on the rival team.

The values obtained by the expressions (3) to (8) are shown in Table 1. In this case, for simplicity, were taken Ffi all equal to 1.

In table 1 and below in tables 2 and 3: The first column identifies the alternatives, the Players, following the nomenclature previously clarified; the second column, Team, corresponds to the team that owns the player; the third column, Po, identifies the position of the player; the four column, PN, is the performance normalized, which corresponds to the Pty, as indicated in (2); the fifth and sixth columns correspond to the cost, Normalized, CN and direct, DC; the columns of the seven to eleven, represent each of the states of nature: MB, BA, IG, SU and MS, that already have been defined and are calculated through the expressions (3) to (8). The last three columns, twelve, thirteen and fourteen, correspond to the assessments that as expressed in (1) are necessary for the model: the expected value, Ve, the amplitude, Am and the model output itself, TAM.

It is important to note in Table 1, that J01034, dominates all other alternatives, however, they are not deleted, as they relate to different positions. This is normal to happen, given that a player with higher performances, usually have all the best payments, that a player with lower performance, for this reason it is also normal that J01193 and J01111, also dominate J02112, that a normal problem of decision making under uncertainty, J02112, should be eliminated, a fact which was not performed, by what it said earlier that represents a different position. In any case, in a normal problem of decision making under uncertainty, would be expected, for all alternatives, what happens between J01193 and J01111, where none dominates the other.

Table 1. Payout table for four players.

| Alt. (Player) | Team | Po | PN | CN | DC | MB | BA | IG | SU | MS | Ve | Am | TAM |
|---------------|------|----|----|----|----|----|----|----|----|----|----|----|-----|
| J01034        | O3   | 100.00 | 100.00 | 19.97 | 70.00 | 110.00 | 111.25 | 121.42 | 88.53 | 91.42 | 141.08 |
| J01193        | Mi   | 41.00 | 15.00 | 3.01 | 26.65 | 34.85 | 45.10 | 51.25 | 33.55 | 32.64 | 71.01 |
| J01111        | Mi   | 46.00 | 70.00 | 14.02 | 16.10 | 34.50 | 41.86 | 52.90 | 41.49 | 46.50 | 65.49 |
| J02112        | Mi   | 38.00 | 38.00 | 7.51 | 13.30 | 28.50 | 34.58 | 45.90 | 33.86 | 32.28 | 54.10 |

In table 2 have been represented, sorted according to their PN, The first forty-five players of La Liga, in this table will be the selection of the fifteen players that at the end form the squad.
In any case, an efficient way of working is that the list appears only players that you have enough money to take it and that does not violate the requirements of the number of players of a team and a position.

In case 1 and 2, you can see that the first player to select will be the J01034, but before you integrate it into the squad should be checked that you have enough money to take it and that does not violate the requirements of the number of players of a team and a position.

In any case, an efficient way of working is that the list appears only players that meet all requirements. For this reason in tables 1, 2 and 3, have been added Team and Po columns, to know quickly which team belongs and which position a player plays.

To carry the analysis it is recommended that you have the players sorted by their points of the season and according to the four positions.
This way it will be easier to perform these checks and possibly discard players whose cost exceeds the remaining budget or belong to teams or play positions, whose quota has already been exhausted. For this reason has been added the column DC, to have on hand the direct cost of each player.

In the case of the CN and PN columns that relate to the normalized performance and costs, they are not used directly in the calculation, but serve the participant to give more or less weight to different SN.

In practice, when using the model, after selected a player, it should check if there are no players with Ptyi equal to 100, in which case you should calculate the new Ptyi equal to 100 and redo all calculations.

However, for simplicity, it is recommended that you continue to work with the values already calculated and only see if some of the players that until now were dominated alternatives have ceased to be dominated, to incorporate them into the table and begin to participate in the analysis.

In summary and as general as possible, the steps for forming the squad would be: Before you begin the implementation of the model, the Minimal indispensable money (DiMI) is calculated, to thereby form four increasingly listings, according to the costs, one for each one of the positions and there take the nj players who turn out to be more economic in each of these lists, with nj the number of players that play the position j needed to complete the team.

The bigger nj is Nj, that is the maximum permitted by the rules. That is to say is initiated with nj = Nj for each one of the four j. The DiMI it will be the sum of the costs of the players that occupy the first nj positions in each one of its lists.

Also the Remaining money (DiRe) is calculated, that is the difference of the initial available money less the invested in all the acquired players. It is initiated DiRe = to the available total money according to the rules (US$ 100 millions, in Desafio La Liga).

The Available money (DiDi) is calculated as the difference of DiRe - DiMI.

If the cost of a player i is represented Ctji, every player with Ctji < DiDi, it can be integrated to the list of available players.

All this information is what has made possible to create the Table 2, with the first forty-five players of La Liga.

Equally a counter for the number of players on the team k (JEqk) that already part of the squad and for the number of players by position j (JPoj) is begun, these two variables are initiated in zero and they will not be able to surpass the maximum of players of a team (MJEqk), neither the maximum of players of a position (MJPoj) respectively.

It starts a list of players ordered decreasingly according to their Ptsi (Table 2), it is recommended to take four sub-lists, one as every position.

Here not to have to evaluate all SN of all the players, it is recommended that to the table where it will be applied TAM, only they be taken 3*Nj players for each one of the four types j of players.

It makes the evaluation of these players and the table is created to resolve TAM, in this case the table 2.
As already said for this table that only the players will go whose $C_{tji}$ is less than $D_{tDi}$ and that besides comply that $J_{Poji}$ it is minor to $M_{JPoji}$ and $J_{Eqk}$ minor to $M_{JEqk}$, according to the position $j$ and team $k$ of the player $i$. It is chosen, just as was done in the Table 1, the player of major TAM, that in this case, for the conditions of entry to the table, it will be a player that can be selected.

Updated $DiMI$, doing $DiMI = DiMI - Ctjj$, with $Ctjj$, the cost of the player of the position $nj$ in the list of cost of the position of the last player selected. After updating $DiMI$, it is updated $nj = nj - 1$; $DiRe = DiRe - Ctji$. It is calculated $DiDi$ and they are eliminated of the table all those players with $Ctji$ greater than $DiDi$; $J_{Eqk} = J_{Eqk} + 1$ for $k$ the team of the player $i$; If $JEqk = MJEqk$ they are eliminated of the table all the players of the team $k$; $J_{Poj} = J_{Poj} + 1$ for the position $j$ of the player $i$ selected; If $JPoj = MPoj$ they are eliminated of the table all the players of the position $j$.

If not working with the full list of players, new players of the different positions are incorporated until having in the list $3*nj$ players of each group.

Note that $nj$ is zero no other positions players will be incorporated. The process is restarted until the squad is complete.

Without giving detail of the incorporation of each player, but if sorted according to their selection, in the Table 3, shows the fifteen players selected.

There, in table 3, making the sum of the elements of the column $DC$, It can be seen that was made very good use of the available budget, which was used, almost entirely.

To finish, in the table 4 presents a summary of the players selected according to their positions and their teams.

In Table 4, taking the information in Table 3, shows that investment of U.S.$ 99.74 from U.S.$ 100 million, that is to say it was a use of almost 100% of the available capital.

Also in this table 4 reflecting the average expected performance of the fifteen players selected. For the calculation of the average performance, use was made of the PN of the players selected.

While this performance is only a 41.73 % can be considered high, since the majority of the players are unable to offer a superior performance to 35 %.

Other aspects that can be seen in tables 3 and 4, is that only eleven of the twenty possible teams provided players.

Also of note is that only a team, the 11th, was arrived to take the limit of three players.

Too be observed, particularly in Table 3, although TAM did not have a maximum number to be reached, the difference between the best and lowest player selected TAM is almost one hundred points.

Of course it can also be seen in tables 3 and 4, which not only meet the budget and the limit of players per team, but used the number of players allowed for each of the four positions.

Although no details are given as to how they were entering to the squad each player, it must be clarified that not always been followed the strict order established by TAM.
### Table 3. The fifteen players selected.

| Team | Fo | GK | Mi | Df | Total |
|------|----|----|----|----|-------|
| 01   | 1  |    |    |    | 1     |
| 03   | 1  | 1  |    |    | 2     |
| 04   | 1  |    |    |    | 1     |
| 06   |    | 1  |    |    | 1     |
| 07   |    |    | 1  |    | 1     |
| 10   |    |    | 1  |    | 1     |
| 11   |    |    |    | 1  | 2     |
| 12   |    |    |    | 1  | 3     |
| 14   |    |    |    | 1  | 1     |
| 16   |    |    |    | 1  | 2     |
| 19   |    |    |    | 1  | 1     |
| Total| 3  | 2  | 5  | 15 |       |

**Performance average expected:** 41.73%  
**Percentage of Money invested:** 99.74%

### Conclusions

First of all it was demonstrated, the particular characteristic that has The Amplitude Model (TAM), to resolve problems of decision making under uncertainty, which is to make use of the dispersion of data, taken as a measure of the dispersion the amplitude, which is used as a penalty.

As shown in tables 1, 2 and 3, you can see that starting from a matrix of decisions under uncertainty, where applicable The Amplitude Model can form a squad of players to participate in a fantasy league.
In particular illustration was made for soccer, following La Liga (soccer Spanish league) and the rules of the fantasy leagues promoted by ESPN for this league. The results achieved, respected all the conditions required in the league rules being followed, in other words, not abusing the budget respecting the maximum number of players per team and position and, of course, the maximum number of players in total.

As a step forward in the integration of statistical and mathematical models in the world of the fantasy sports, for a future is recommended to work with completely real data.

A possibility of having these real data, it would be the case to take the results of the season 2011-2012, generating the squad for the 2012-2013 and to verify the results obtained with the select squad. Making like that a finished ratification of the followed strategy and of the use of TAM.

References

1. G. K. Akilli. Knowledge exchange patterns in ESPN’S Fastbreak community, Doctoral dissertation, The Pennsylvania State University, USA, 2010.

2. E. J. Burke. “Quasi-property” Rights: Fantasy or Reality? An Examination of C.B.C. Distribution & Marketing Inc. v. Major League Baseball Advanced Media, L.P. and Fantasy Sports Providers’ Use of Professional Athlete Statistics, Washington University Journal of Law & Policy, 161-190, 2008.

3. N. W. Davis and M. C. Duncan. Sports Knowledge is Power. Reinforcing masculine privilege through Fantasy Sport League participation. Journal of Sport & Social Issues, 30, 3, 244-264, 2006.

4. M. C. Duncan. Bodies in Motion: The Sociology of Physical Activity. Quest, 59, 55-66, 2007.

5. B. Dwyer. Fantasy sport consumer behavior: An analysis of participant attitudes and behavioral intentions, Doctoral dissertation, University of Northern Colorado, USA, 2009.

6. B. Dwyer and Y. Kim. For love or money: Developing and validating a motivational scale for fantasy football participation, Journal of Sport Management, 25, 70-83, 2011.

7. ESPN. Official Rules. Desafío La liga. In: http://games.espn.go.com/desafio-la-liga/es/story?pageName=desafio-la-liga/reglas, consulted January 2013, 2012.

8. L. K. Farquhar and R. Meeds. Types of fantasy sports users and their motivations, Journal of Computer-Mediated Communication, 12, 4, 1208-1228, 2007.

9. M. J. García, J. G. Hernández and G. J. Hernández. Una Metodología Integradora-Adaptable para desarrollar Sistemas de Apoyo a las Decisiones (MIASAD). In J. Valderrama (Ed.), Congreso Interamericano de Computación Aplicada a la Industria de Procesos CIAP’2011, Universidad de Girona, España, 753-760, 2011.

10. M. J. García, J. G. Hernández and G. J. Hernández. Making decisions under risk and uncertainty of the Virtual channel Manager of the Logistic Model Based on Positions. In G. Dukic (Ed.), Proceedings ICIL’2012, University of Zagreb, Croatia, 448-455, 2012.

11. M. J. García, G. J. Hernández and J. G. Hernández. Quantification of the Ethical-Legal decision. In N Delener et al. (Ed.), International Conference Global Business And Technology Association, Reading Book, GBATA, USA, 250-257, 2012.
12. A. Gillentine and J. Schulz. Marketing the Fantasy Football League: Utilization of Simulation to Enhance Sport Marketing Concepts, *Journal of Marketing Education*, 23, 3, 178-186, 2001.

13. J. G. Hernández and M. J. García. Análisis del factor BETA (β), en El Modelo de Amplitud (EMA), *Revista de Matemática: Teoría y Aplicaciones*, 10, 1-2, 187-199, 2003.

14. J. Hernández R., M. García G. and G. Hernández G. Enterprise logistics, indicators and Physical distribution manager. *Research in Logistics & Production*, 3, 1, 5-20, 2013.

15. D. H. Kwak, C. H. Lim, W. Y. Lee and J. Mahan III. How Confident Are You to Win Your Fantasy League: Exploring the Antecedents and Consequences of Winning Expectancy. *Journal of Sport Management*, 24, 416-433, 2010.

16. J. Lee. The effects of fantasy football participation on team identification and NFL fandom, Master dissertation, Louisiana State University, USA, 2011.

17. M. G. Massari. When fantasy meets reality: The clash between on-line fantasy sports providers and intellectual property rights, *Harvard Journal of Law & Technology*, 19, 2, 443-465, 2006.

18. T. P. Oates. New Media and the Repackaging of NFL Fandom. *Sociology of Sport Journal*, 26, 31-49, 2009.

19. U. Pillay, R. Tomlinson and O. Bass. Development and dreams. The urban legacy of the 2010 football world cup, HSRC Press, South Africa, 2009.

20. D. L. Wann, F. G. Grieve, R. K. Zapalac, J. A. Partridge and P. M. Parker. An Examination of Predictors of Watching Televised Sport Programming, *North American Journal of Psychology*, 15, 1, 179-194, 2013.

21. L. A. Zadeh. Generalized theory of uncertainty (GTU)—principal concepts and ideas. *Computational Statistics & Data Analysis*, 51, 5-46, 2006.
Binary dynamic response model with structural changes

Šárka Hudecová

Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University in Prague, Sokolovsk 83, CZ-186 75 Prague 8, Czech Republic (E-mail: hudecova@karlin.mff.cuni.cz)

Abstract. Binary time series play an important role in many fields of application, including macroeconomics and financial econometrics. A common approach to modelling a binary series is based on a dynamic probit model (also binary dynamic response model).

We study dynamic binary response models with possible changes in their parameters. A cumulative sum (CUSUM) type test statistic is proposed for testing a change. The asymptotic distribution of the test statistic is derived under the null hypothesis. The performance of the test is illustrated by a simulation study.

Keywords: binary time series, binary dynamic response model, dynamic probit model, change point analysis.

1 Introduction

Binary time series play an important role in many fields of application, including macroeconomics and financial econometrics. They are typically observed when one is concerned with an occurrence of an event in a time period. For example, daily occurrences of precipitation can be modeled as binary time series, see Wilks and Wilby[12]. In a financial area, one might be interested in recession indicators, see Kauppi and Saikkonen[6] or Startz[10], or in a series of direction-of-change of stock returns.

In practice, the data-generating mechanism may change at some unknown time point (change point). The main objective is then to decide whether a change has occurred and to detect the change point. There is a broad statistical literature on the change point topic, see Csörgő and Horváth[2] for an overview. In the time series setting, various results have been derived for ARMA models, see Aue and Horváth[1] for an extensive overview on the recent work in this field. Within the framework of discrete time series, we mention Weiss and Testik[11], who investigate detection of changes in INAR processes, and Franke et al.[4], where cumulative sums (CUSUM) type test statistic is derived for the Poisson autoregression of order one.

In this paper we study dynamic binary response models with possible changes in their parameters. The considered model is introduced in Section 2. A cumulative sum (CUSUM) type test statistic is proposed for testing a change. The asymptotic distribution of the test statistic is derived under the null hypothesis in Section 3. A small simulation study is provided in Section 4, and an application to real data is shown in Section 5.
2 Considered Model

Let \( \{Y_t\} \) be a binary (0-1 valued) time series of interest and let \( \{x_t\} \) be \( k \times 1 \) vector process of explanatory variables such that \( x_t \) includes the constant term for all \( t \). Let \( \mathcal{F}_t = \sigma\{Y_s, x_s, s \leq t\} \) be the \( \sigma \)-field of the information available at time \( t \) and let \( \mathcal{F}_t^* = \sigma\{\mathcal{F}_t, x_{t+1}\} \). Assume that the conditional distribution of \( Y_t \) given \( \mathcal{F}_{t-1}^* \) is binary with the success probability \( \pi_t \) which depends on \( x_t \) and \( p \) previous values of the series \( Y_{t-1}, \ldots, Y_{t-p} \) via the model

\[
Y_t = I\left(\sum_{j=1}^{p} \gamma_j Y_{t-j} + \beta' x_t + u_t > 0\right),
\]

where \( I(\cdot) \) is an identifier function and \( u_t \) are iid with a standard normal distribution. Then we can write

\[
\Phi^{-1}(\pi_t) = \sum_{j=1}^{p} \gamma_j Y_{t-j} + \beta' x_t,
\]

where \( \Phi \) stands for the distribution function of the standard normal distribution. The model is referred to as a dynamic probit model (also a binary dynamic response model), see Kauppi and Saikkonen[6], de Jong and Woutersen[3].

Having data from (1), the parameters \( \theta = (\gamma_1, \ldots, \gamma_p, \beta') \) can be estimated using the maximum likelihood (MLE) method, see Kedem and Fokianos[7] or de Jong and Woutersen[3]. Under some regularity conditions (specified later) such estimator is consistent and asymptotically normal.

Let \( Y_{-p}, \ldots, Y_0, Y_1, Y_2, \ldots, Y_n \) be realizations of the model (1). This means that we have \( n \) realizations of \( (Y_t, Y_{t-1}, \ldots, Y_{t-p}, x_t) \). We would like to decide whether a change has appeared in the data generating process. Hence, we introduce the model

\[
\Phi^{-1}(\pi_t) = \begin{cases} 
\sum_{j=1}^{p} \gamma_j Y_{t-j} + \beta' x_t, & t = 1, \ldots, m \\
\sum_{j=1}^{p} \gamma_j^* Y_{t-j} + \beta'^* x_t, & t = m + 1, \ldots, n,
\end{cases}
\]

where \( \theta = (\gamma_1, \ldots, \gamma_p, \beta') \neq \theta^* = (\gamma_1^*, \ldots, \gamma_p^*, \beta'^*) \). Model (3) describes the situation where the first \( m \) observations follow the model (1) with the parameters \( \theta \), and the remaining \( n - m \) observation follow the model (1) with the parameters \( \theta^* \). The main objective is to test whether a change has occurred or not, i.e., to test

\[
H_0 : m = n \quad \text{against} \quad H_1 : m < n.
\]

We propose the test statistic for testing \( H_0 \) against \( H_1 \) which is based on the cumulative sums (CUSUMs) of the estimated residuals. Namely, let \( \hat{\pi}_t \) be given by (2) with \( \theta \) replaced by the MLE estimator \( \hat{\theta} \). Define

\[
T_n = \max_{1 \leq k \leq n} \sqrt{\frac{n}{k(n-k)}} \frac{\sqrt{\frac{1}{n} \sum_{t=1}^{n} |Y_t - \hat{\pi}_t|}}{\sqrt{\frac{1}{2} \sum_{t=1}^{n} \hat{\pi}_t(1 - \hat{\pi}_t)}}.
\]
Such a statistic is a common tool for a changepoint detection, see Csörgő and Horváth[2]. The hypothesis $H_0 : m = n$ is rejected if $T_n > c_n$, where $c_n$ is the corresponding critical value. In the next section we show that under some regularity conditions, the test statistic, appropriately normalized, converges under the null hypothesis to the Gumbel extreme value distribution.

3 The limiting distribution under the null hypothesis

Assume that the following assumptions hold:

(A1) The true parameter $\theta$ lies in an open subset of $\mathbb{R}^{p+k}$.

(A2) The sequence $\{x_t\}$ is strictly stationary and strong mixing with mixing coefficients decaying exponentially fast. In addition, $x_t$ has finite second absolute moments.

(A3) The random vector $x_t$ lies almost surely in a nonrandom compact subset $\Gamma \subset \mathbb{R}^k$ for all $t$. In addition, if we define $w_t = (x_t', y_{t-1}, \ldots, y_{t-p})'$ then $P(\sum_{i=1}^N w_i w'_i > 0) = 1$.

(A4) Denote $S = \{0, 1\}^p$ and define $G = \{g : g = \sum_{j=1}^p \gamma_j s_j \text{ for } s \in S\}$. Let $g_{\text{min}} = \min G$ and $g_{\text{max}} = \max G$. Assume that there exists $\delta > 0$ and $K > 0$ such that

$$P(g_{\text{max}} + \max_{1 \leq i \leq p} (\beta' x_{t-i} + u_{t-i}) > 0 | y_{t-p-K}, y_{t-p-K-1}, \ldots)$$

$$- P(g_{\text{min}} + \max_{1 \leq i \leq p} (\beta' x_{t-i} + u_{t-i}) > 0 | y_{t-p-K}, y_{t-p-K-1}, \ldots) < 1 - \delta.$$ 

It is shown in de Jong and Woutersen[3] that under the assumptions (A2) and (A4) there exists a strictly stationary solution of (1) such that $(y_t, x'_t)$ is strictly stationary and strong mixing with the mixing coefficients decaying exponentially fast. Note that the condition (A4) is rather technical and it limits the predictability of $y_t$ given a distant past, see de Jong and Woutersen[3] for more explanation.

Furthermore, under (A1)-(A3) there exists a unique maximum likelihood estimator $\hat{\theta}$ of $\theta$ which is consistent (in probability) and asymptotically normal, see Kedem and Fokianos[7, Chapter 1] and note that the assumptions are fulfilled due to the strong mixing property. Define

$$a(n) = \sqrt{2 \log \log n} \quad \text{and} \quad b(n) = 2 \log \log n + \frac{1}{2} \log \log \log n - \frac{1}{2} \log \pi.$$ 

We have the following limiting distribution for $T_n$

**Theorem 1.** Let assumptions (A1)-(A4) hold. Then under $H_0 : m = n$ we have

$$P(a(n)T_n - b(n) < t) \to \exp\{-2e^{-t}\}$$

as $n \to \infty$, $t \in \mathbb{R}$. 

467
The previous theorem can be used for the computation of the asymptotic critical value for the test of $H_0 : m = n$. If $H_0$ is rejected then

$$m^* = \arg\max_{1 \leq k \leq n} \sqrt{\frac{n}{k(n-k)}} \frac{|\sum_{t=1}^{k} Y_t - \hat{\pi}_t|}{\sqrt{\frac{1}{n} \sum_{t=1}^{n} \hat{\pi}_t(1 - \hat{\pi}_t)}}$$

can be taken as an estimate of the unknown change point $m$. The proof of Theorem 1 is postponed to Section 6.

Remark that it follows from the proof of Theorem that the limiting distribution of $T_n$ is the same as the limiting distribution of $U_n$ defined as

$$U_n = \max_{1 \leq k \leq n} \sqrt{\frac{n}{k(n-k)}} \frac{|\sum_{t=1}^{k} Y_t - \hat{\pi}_t|}{\sqrt{\frac{1}{n} \sum_{t=1}^{n} (Y_t - \hat{\pi}_t)^2}}$$

Simulations indicate that the performance of the two test statistics $T_n$ and $U_n$ is comparable with respect to the power of the test.

4 Simulations

In this section we present results of some simulations which illustrate the behavior the suggested test procedure.

In order to investigate the behavior under the null hypothesis, $N = 1000$ replicas of a binary time series $\{y_t\}$ following the particular model (1) (without a change) were simulated. For each replica, the empirical distribution of the normalized version of the test statistic

$$t_n = \left( T_n - \sqrt{2 \log \log(n) - \frac{\log \log \log n}{2 \sqrt{2 \log \log(n)}}} \right) \sqrt{2 \log \log n + \frac{1}{2} \log(\pi)}$$

was computed. Characteristics as distribution function and quantiles were compared to their theoretical counterparts. The simulations were conducted for various choices of $n$ and model parameters $\theta$.

As expected, the rate of the convergence to the extreme value distribution is quite low, which is a well-known fact. This is visible, for instance, from the comparison of the sample quantiles listed in Table 1 or from the comparison of the empirical and limiting distribution functions in Figure 1. The presented results were obtained for model (1) with $p = 1$, $k = 2$, $\gamma_1 = 2$, $\beta_1 = -2$, $\beta_2 = 1$, and $x'_t = (1, w_t)$ such that $\{w_t\}$ are iid with the uniform distribution on $[0,1]$.

The performance of the test procedure was studied under the alternative as well. As expected, the power of the test increases with an increasing sample size and with an increasing “size” of the change. Furthermore, the power is largest if the change occurs approximately in the middle of the sample. The dependence of the (estimated) power on the sample size $n$ is shown for several choices of $\lambda = m/n$ in Figure 2. The presented results are for model (3) with $p = 1$, $k = 2$, $x'_t = (1, w_t)$ such that $\{w_t\}$ are iid with the uniform distribution on $[0,1]$. The parameters were chosen as $\gamma_1 = 2$, $\beta_1 = -2$, $\beta_2 = 1$, and
\begin{tabular}{lcccc}
  asymp. & \( n = 100 \) & \( n = 200 \) & \( n = 500 \) & \( n = 1000 \) \\
 1\% & -0.8340 & -0.5661 & -0.6683 & -0.7746 & -0.8392 \\
 5\% & -0.4040 & -0.1639 & -0.2965 & -0.3711 & -0.3733 \\
 10\% & -0.1409 & -0.0005 & -0.0817 & -0.1373 & -0.1284 \\
 50\% & 1.0597 & 0.9367 & 0.8718 & 0.8776 & 0.9580 \\
 90\% & 2.9435 & 2.0767 & 2.0489 & 2.2064 & 2.5189 \\
 95\% & 3.6633 & 2.3833 & 2.5077 & 2.7220 & 3.0521 \\
 99\% & 5.2933 & 2.9618 & 3.3462 & 3.2361 & 3.9214 \\
\end{tabular}

\textbf{Table 1.} Comparison of the theoretical and sample quantiles

\begin{figure}
\centering
\includegraphics[width=0.5\textwidth]{fig1.pdf}
\caption{Comparison of the empirical distribution functions of \( t_n \) for \( n = 100, 200, 500, 1000 \) with the limiting distribution \( G(t) = \exp(-2e^{-t}) \) (solid line).}
\end{figure}

\( \gamma_1^* = 1, \beta_1^* = 0.5, \beta_2^* = -2 \). This change causes a change in the autocorrelation structure of the binary process, a change in the dependence of \( Y_t \) on \( x_t \), and also a change in the stationary (unconditional) probability of a success (an increase from 0.2 to 0.5).

\begin{figure}
\centering
\includegraphics[width=0.5\textwidth]{fig2.pdf}
\caption{Estimated power of the test (based on 300 replicas) as a function of the sample size \( n \) for various choices of \( \lambda = m/n \).}
\end{figure}
5 Real data analysis

We apply the test of a change to the time series of US quarterly recession indicators from 1855–2011 obtained from The National Bureau of Economic Research (n = 628 records). Here, $Y_t$ is coded as 1 if any month in the quarter $t$ is being in a recession, and $Y_t = 0$ otherwise, see [10].

Model (1) with $p = 3$ and $x_t \equiv 0$ was considered under the null hypothesis. The test statistic $T_n$ equals to $T_n = 4.2113$, and the corresponding critical value on level $\alpha = 0.05$ is $c_\alpha = 3.6926$. Hence, the null hypothesis of no change is rejected with an asymptotic $p$-value 0.0187. The maximum of $T_n$ is attained for $k = 313$, the first quarter of 1933, but the critical value $c$ is crossed in the whole period 1933–1939.

The procedure detected a single change in the series of US recession indicators with the change point in 1933. The change in the structure of the model led to a change in the autocorrelation structure of $\{Y_t\}$ as well as to a lower overall probability of a recession, cf. [5].

6 Proof of the main result

Proof. Let $Z_t = Y_t - \pi_t$ and $S_k = \sum_{t=1}^k Z_t$. It follows from the strong mixing property of $\{Y_t\}$ and $\{x_t\}$ that $\{Z_t\}$ is strong mixing with mixing coefficients decaying exponentially fast. It then follows from the strong invariance principle, see Kuelbs and Philipp[8], that we can replace $Z_t$ by iid normal variables without changing the asymptotic distribution of $\max_{1 \leq k \leq n} \sqrt{\frac{n}{k(n-k)}} |S_k - \frac{k}{n} S_n|$, see Schmitz[9, Theorem 2.1.5] for details. Namely, let $\sigma^2 = E(Y_t - \pi_t)^2 = E\pi_t(1 - \pi_t)$ (a constant from the strictly stationarity) then for $t \in \mathbb{R}$

$$
P \left( a(n) \max_{1 \leq k \leq n} \sqrt{\frac{n}{k(n-k)}} \frac{|S_k - \frac{k}{n} S_n|}{\sigma} - b(n) < t \right) \rightarrow \exp\{-2e^{-t}\} \quad (6)$$

as $n \rightarrow \infty$.

Let $\hat{S}_k$ be the empirical counterpart of $S_k$, that is $\hat{S}_k = \sum_{t=1}^k (Y_t - \hat{\pi}_t)$. It is possible to show after some computations that

$$
\max_{1 \leq k \leq n} \sqrt{\frac{n}{k(n-k)}} |\hat{S}_k - (S_k - \frac{k}{n} S_n)| = o_P((\log n)^{-1/4}).
$$

The approximation is derived analogously as in Hudecová[5]. It then follows that

$$
P \left( a(n) \max_{1 \leq k \leq n} \sqrt{\frac{n}{k(n-k)}} \frac{|\hat{S}_k|}{\sigma} - b(n) < t \right) \rightarrow \exp\{-2e^{-t}\}. \quad (7)$$

From the Taylor expansion and (A3) we get

$$
\frac{1}{n} \sum_{t=1}^n \hat{\pi}(1 - \hat{\pi}_t) = \frac{1}{n} \sum_{t=1}^n \pi_t(1 - \pi_t) + O_P(1/\sqrt{n}),
$$
and combining this with the strong mixing property of $\pi_t$ we get

$$\frac{1}{n} \sum_{t=1}^{n} \hat{\pi}(1 - \hat{\pi}_t) = \sigma^2 + O_p(n^{-1/4}).$$

This enables us to replace $\sigma^2$ in (7) by $\frac{1}{n} \sum_{t=1}^{n} \hat{\pi}(1 - \hat{\pi}_t)$ without changing the asymptotics. This finishes the proof.

7 Acknowledgments

The work of the author is supported by the Czech Science Foundation project DYME Dynamic Models in Economics No. P402/12/G097.

References

1. A. Aue and L. Horváth,. Structural breaks in time series. J. Time Ser. Anal. 34, 1–16, 2013.
2. M. Csörgő and L. Horváth. Limit Theorems in Change-Point Analysis, Wiley, New York, 1997.
3. R.M. de Jong and T. Woutersen. Dynamic time series binary choice. Econometric Theory 27, 673–702, 2011.
4. J. Franke, C. Kirch and J.T. Kamgaing. Changepoints in times series of counts. J. Time Ser. Anal. 33, 757–770, 2012.
5. Š. Hudecová. Structural changes in autoregressive models for binary time series. Submitted, 2013.
6. H. Kauppi and P. Saikkonen. Predicting U.S. recessions with dynamic binary response models. Rev. Econ. Stat. 90, 777–791, 2008.
7. B. Kedem and K. Fokianos. Regression Models for Time Series Analysis. Wiley, New York, 2002.
8. J. Kuelbs and W. Philipp. Almost sure invariance principles for partial sums of mixing $B$-valued random variables. Ann. Probab. 8, 1003–1036, 1980.
9. A. Schmitz. Limit Theorems in Change-Point Analysis for Dependent Data. Ph.D. thesis. University of Cologne, Germany, 2011. Available online at http://kups.ub.uni-koeln.de/4224/
10. R. Startz. Binomial autoregressive moving average models with an application to U.S. recession. J. Bus. Econom. Statist. 26, 1–8, 2008.
11. C. Weiss and M. Testik. CUSUM monitoring of first-order integer-valued autoregressive processes of Poisson counts. J. Quality Techn. 41, 389–400, 2009.
12. D. Wilks and R. Wilby. The weather generation game: a review of stochastic weather models. Prog. Phys. Geog. 23, 329–357, 1999.
TO RELIABILITY OF MORTALITY SHIFTS IN WORKING POPULATION IN RUSSIA

Alla Ivanova1, Tamara Sabgaida2, Viktoria Semyonova3, Elena Zemlyanova4

Department of Health Statistics, Federal Research Institute for Health Organization and Informatics of Ministry of Health of the Russian Federation, 11, Dobrolubov str., Moscow, Russia

1 e-mail: ivanova-home@yandex.ru
2 e-mail: tsabgaida@mail.ru
3 e-mail: vika-home@yandex.ru
4 e-mail: zemlianova@mednet.ru

1 Introduction

Russian population mortality during last 50 years appears to be widely studied problem both in comparative European context (J. Shapiro, 1995; VM Shkolnikov, F Mesle, J Vallin, 1996; VM Shkolnikov, VV Chervyakov, 2000; NS Gavrilova et al., 2008) and in relation to its internal regional variation.

Accuracy of mortality causes diagnostics progressively becomes more urgent. It’s possible to allocate several sources of mortality causes distortion. First, the cases when death cause is established by a paramedic or a doctor whose task was just to state a death, i.e. a person not having adequate qualification or information for establishing a diagnosis. This leads to increasing a number of death causes including in their definition such characteristics as “other” or “unspecified” (VG Semyonova et al., 2004; NS Gavrilova et al., 2008). Second, it is observed growth of mortality from diagnosis “Event of undetermined intent (Y10-Y34)” or “Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified (R00-R99)” which relates to concealment of violent death causes (VG Semyonova, GN Evdokushkina, 2003; DA Wiseman, EV Dubrovina, AN Redko, 2006; AE Ivanova, VG Semyonova, 2004). Third, it is low share of postmortem examinations in case of somatic death causes which leads to reduction of alcohol-related mortality components (DA Wiseman, EV Dubrovina, AN Redko, 2004; EV Dubrovina, 2006). Forth, it is data compilation and formation of database individual mortality data of Federal State Statistics Service only by primary death cause excluding associated diseases. This doesn’t allow to execute logic control of accuracy of death cause establishing (EI Pogorelova, 2007; VG Semyonova, OI Antonova, 2007).

Working population during several decades occurs to be the key group determining life expectancy (LE) level in Russia, its regional variance, trends and projections. And namely for this group the problem of accuracy of death cause establishing is the most urgent.

2 The aim of the study
The aim of the study is to analyze death causes structure of working population (15-59 years) in Russia, to determine its age-specific and gender regularities, to define their influence on nosological profile of mortality in working ages.

3 Methods and data

On the base of official statistics we studies age and gender peculiarities of mortality in working ages; structure of death causes and its shifts at the stages of both negative dynamics and mortality reduction at more than 40 years period (1965-2010). While dynamics analysis we used indicators recalculated with consideration for population censuses results including 2010 population census. Component analysis was used for determining of age groups and death causes which defined mortality changes during periods of “soviet evolutionary LE reduction” (1965-1980); “opportunistic LE growth as a result of anti-alcohol campaign” (1980-1987); “crisis LE decline at the stage of shock socio-economic reforms” (1987-1994); “opportunistic LE changes during reforms’ stagnation” (1994-2005); steady LE growth at the background of conducted policy” (2005-2010).

4 Results

Periods of mortality dynamics in Russia

Long-term life expectancy dynamics of Russian population is well-known (F Millet, M Shkolnikov, 2006; EM Andreev, AG.Vishnevksy, 2004; VM Shkolnikov et al., 2004; AE Ivanova, 2009). That’s why lets cover the aspects which are important for understanding of periods and their causes.

The middle of 1960s was selected as a starting point of research. At that moment previous positive mortality dynamics which allowed Russia to catch up European countries largely at life expectancy level was depleted. Since the middle of 1960s the negative trend took its shape and it continued till the edge of 1980s. During that period male LE decreased by 2.8 years, female LE – by 0.4 years. The majority of losses (2.5 years in males and 0.35 years in females i.e. about 88-92%) were determined by age groups 15-59 years. The causes which determined steady negative dynamics were comprehensively discussed in scientific publications. To resume them briefly – ideological reasons were principal and first of all – disrespect of life value both at state and individual levels. This is testified by similarity of mortality trends in the majority of socialist countries even levels of mortality could be rather different.

The beginning of 1980s was characterized by slight opportunistic fluctuations of life expectancy that transformed into evident positive fluctuations.
dynamics with the start of anti-alcohol campaign. It’s well known that during the short period of hard-edged measures life expectancy increased essentially. Up to 1987 losses during previous 15-years period were not only compensated but LE reached maximal level in Russian history exceeding maximum of 1965 by 0.6 years in males and by 0.9 years in females. During 1980-1987 the share of working population 15-59 years old accounted 2.8 and 0.8 years from 3.4 and 1.3 years in males and females correspondingly i.e. 82% and 67% of overall LE growth.

Following gradual rejection of hard-edged measures limiting accessibility to alcohol gradual mortality growth upraised and it accelerated since the beginning of 1990s at the background socio-economic reforms that led to sharp impoverishment, losing of social waymarks and perspectives for the majority of Russian population. In total, during 1987-1994 life expectancy losses estimated 7.2 years in males and 3.2 years in females; the share of 15-59 age group was 6.2 and 2.2 years or 80% and 63% in males and females correspondingly.

The period of the second half of 1990s and first half of 2000s was characterized of sharp mortality fluctuations: partial recovery of life expectancy after collapse in the middle of 1990s, new decline after 1998 and gradual way to stabilization in 2002-2005. As a result, the situation with mortality of Russian population occurred to be just slightly better than 10 years before – at the peak of socio-economic crisis. Both in males and females LE was high the level of 1994 by 1.3 years. But this small gain was to minimal extent due to risk groups. Age groups 15-59 years covered only 0.4 and less than 0.1 years in males and females or 29% and 2% correspondingly. In other words, situation with mortality of working population in 2005 was the same sharp as in catastrophic 1994.

New stage of Russian mortality dynamics started since the middle of new decade when positive life expectancy trend developed. It was determined by the whole age range, main death causes and the majority of regions. It was not instigated by some opportunistic interventions but developed at the background of previous mortality stabilization as a result of implementation of continuous measures aimed at mortality reduction both at federal and regional levels. In turn it gives a hope for continuation of positive dynamics conditional upon implementation of scheduled policy measures. In general, life expectancy in males increased by 4.1 years, in females – by 2.4 years; and 3.0 and 1.2 years (or 75% and 48%) in males and females were due to population in working ages. Thus, during the second half of decade it was possible to achieve essential positive shift in men in working ages; in women – only one half of increase was due to risk groups.

Death causes that determined life expectancy shifts at stages of its growth and decline

During 15 “soviet” years during 1965-1980 reduction of male LE in working ages was determined by all age groups from 15 to 59 years but one half of all losses was due to 45-59 year-olds. Among causes the first place was
occupied by traumas and poisonings (1.5 years in age interval 15-59 years) and cardio-vascular diseases (0.88 years). It’s important that aside from those causes the input into LE losses was made by respiratory diseases (0.2 years); digestive diseases (0.12 years) and partly – neoplasms (0.01 years). Only in class of neoplasms negative dynamics spread to selected age groups (over 45 years). All other causes had determined losses at the whole age range from 15 to 59 years. Only infections in the contrary to other main causes of death didn’t participate LE losses in men of working ages.

The situation with women in 1965-1980 was essentially different from men’s. First, the losses were formed due to age group over 45 years; in young women mortality trends from the majority of causes and in general were positive. Second, female mortality from neoplasms decreased in all age groups; that’s why neoplasms as well as infections didn’t participate in LE losses in working ages. At the same time, the similarity on formation of LE losses in males and females was in prevalence of external causes and cardio-vascular diseases as well as partial input of respiratory and digestive diseases.
During the period of essential LE growth in 1980-1987 positive
dynamics was characteristic to all age groups and main death causes.
Both in males and females equal input into LE growth was made by
age groups 30-44 years and 45-59 years. That’s why the role of external causes
in LE gain was critical: 1.91 from 2.8 years of total growth in 15-59 years old
males; and 0.42 from 0.84 years in females. In males the role of cardio-vascular
and respiratory diseases was relatively comparable (0.32 and 0.25 years);
digestive diseases – half as much (0.13 years). In females the role of cardio-
vascular diseases was half as much as traumas and poisonings (0.21 years); the
role of other causes – negligible.
The special role as during the previous period was taken by neoplasms.
In males neoplasms continued their negative trend apart of sinuosity of anti-
alcohol campaign; and it affected ages over 45 years similarly to the previous
period. In females reduction of mortality from neoplasms continued in all age
groups.
Life expectancy losses during 1987-1994 formed at the whole range of
working ages but the scales of losses were increasing with age and maximum
was observed in the group 45-49 years both in males (2.75 of 6.18 years) and
females (1.16 of 2.21 years). About one half of losses in working ages was due
to external causes: 3.31 and 1.0 years in males and females correspondingly;
losses due cardio-vascular diseases were twice less – 1.54 and 0.67 years in
males and females correspondingly. In contrast to the previous period mortality
growth was registered in all death causes and age groups. This emphasizes the
universalism of effect of social processes on health and mortality during that
period.
The new phenomenon of that period was manifested in increase of mortality in working ages from “Symptoms, signs and ill-defined conditions” which covered LE losses in working ages comparable to losses from respiratory and digestive diseases and higher than losses from infections.

Since the middle of 1990s till the middle of 2000s mortality dynamics became very diverse; it broke previous uniformity in respect to population response to social changes.

In men cumulative results occurred to be positive both for young population groups (15-24 years) and older groups of working ages (35-59 years). But the most active and productive groups (25-34 years) suffered from negative trends. In females gain were achieved only in teenagers (15-19 years) and groups over 40 years. Unfavourable ages covered wider age group than in males – 20-40 year olds.

It is important that slight LE increase during the decade was due to reduction of mortality from traumas and poisonings as well as neoplasms and other diseases apart from 7 main classes of death causes. Mortality from neoplasms declined in all ages providing LE increase in working ages by 0.24 and 0.10 years in males and females correspondingly. Mortality from traumas and poisonings declined in males in all ages, in females – in all ages excluding age group 20-34 years. This provided LE growth in working ages by 0.82 and 0.22 years. Mortality from other causes was increasing.

Positive trends in external causes and neoplasms invite some questions but for various reasons. As to external causes, driving forces of mortality reduction have no explanation. There were no evident improvements in living standards of poor population groups, there were no implemented measures concerning anti-alcohol policy, there was no special attention to healthy life styles. This is supported by negative dynamics of mortality from causes being markers of these problems – infections and digestive diseases.
In is necessary to make a special reference to neoplasms. Developed positive dynamics of mortality from these causes both in males and females at the whole working age range fell to the most difficult period for Russian health care – protracted reforms at the background of chronic under-financing. It’s difficult to find valid explanations for this phenomenon.

The specific feature of that period was the age structure of mortality growth from chronic non-infectious diseases and first of all from cardiovascular diseases. Both in males and females growth of mortality from cardiovascular pathology was concentrated in young and middle ages. At the same time there was no growth in males over 45 years and there was even slight reduction in females. It is nonsense from the medical-biological viewpoint. Nevertheless, such age profile for mortality was characteristic to respiratory and digestive diseases as well.

Despite of substantial differences of 1994-2005 decade from the previous period it is possible to notice similarity of both decades characteristic to long-running stage of transformation in socio-economic sphere in Russia. This refers to continuous growth of mortality from infections and from “Symptoms, signs and ill-defined conditions” covering whole range of working ages.
Since the middle 2000s we observe the new period of overcoming of negative mortality trends and life expectancy growth in Russia. LE increase was provided by all age groups with maximal input of ages over 45 years: 1.35 of 3.03 years in men and 0.67 of 1.16 years in women.

LE growth was provided by all main death causes excluding infections. In males, infections maintained their negative trends in age group 30-34 years but the overall result in working ages was positive. In females, negative trends of infections covered wider age interval 25-39 years and the overall result for working ages occurred to be negative.

According to younger age profile of life expectancy increase in males the main input into LE growth was made by external causes - 1.5 years; twice less by cardio-vascular diseases – 0.86 years. In females, significance of these causes was equal – by 0.44 years. The role of all other causes occurred to be less evident.

It is necessary to note that age distribution of external causes input into life expectancy increase in working ages was sufficiently proportional. At the same time, the maximal gain from reduction of mortality from chronic non-infectious diseases, first of all from cardio-vascular diseases was registered in ages over 40 years both in females and males. Thus, cumulated since the middle of 1990s rejuvenation of mortality from cardio-vascular diseases was not affected yet.

Conclusions
Summarizing 50-years of history of Russian mortality it is necessary to note several important things:
- up to the end of the first decade of XXI century Russia returned to 45-years old life expectancy levels; this visible reversion constitutes of women’s gain of 1.3 years and the equal disadvantage in men;
- the main source of both disadvantages and gains during all stages of nearly half century dynamics is the population of working ages; in males cumulative losses due to population of 15-59 years following the results of 1965-2010 occurred to be 2.5 years of life expectancy, in females – 0.5 years.
- resulting losses during working-life period are combined from 2 age groups: 30-44 year-olds and 45-59 year-olds – in males proportion of those 2
groups was 40% and 60%; in females reverse – 60% and 40%. Input of younger age groups both in males and females occurred to be near zero point i.e. levels of mortality in men and women 15-29 years old in 2010 and in 1965 nearly didn’t differ.

- in the ages where resulting mortality during analyzed period didn’t change in general, and in ages where it increased the structure of death causes visibly changed:

- as to structure of mortality in young ages, input of respiratory and digestive diseases as well as ill-defined conditions and plus infections and cardio-vascular diseases in males the resulting mortality remained the same as in 1965 only due to reduction of traumas and poisonings and neoplasms;
- as to structure of mortality in working ages over 30 years, input of main somatic diseases (except neoplasms) and external causes increased which determined growth of summarized mortality in these ages.

Altogether, conducted analysis rises following questions:
- reliability of reduction of mortality from external causes that started before improvement of socio-economic situation in the country and implementation of any measures in that sphere (during 1994-2005);
- reliability of growth of cardio-vascular mortality in middle and especially young ages (15-29 years);
- reliability of diagnostics of somatic death causes in general including neoplasms.

References

1. Gavrilova NS, Semyonova VG, Dubrovin EV, Evdokushkina GN, Ivanova AE, Gavrilov LA. Russian Mortality Crisis and the Quality of Vital Statistics. Population Research and Policy Review. 2008 (27): 551-574

2. Ivanova AE, Semyonova VG. New phenomena of Russian mortality [Novye yavleniya Rossiskoy smertnosti]. Population [Narodonaselenie]. 2004; (3): 85-93. (in Russian)

3. Pogorelova EI. On improvement of records on mortality statistics [O sovershenstvovanii uchetnyh dokumentov po statistike smertnosti]. Social aspects of public health. [Sotsialnye aspekty zdorovya naseleniya] (Electronic publication), 2007; 2 (2): Available from: URL: http://vestnik.mednet.ru/content/view/27/30/lang,ru/ (in Russian)

4. Semyonova VG, Antonova OL. The reliability of mortality statistics (for example deaths from injury and poisoning in Moscow) [Dostovernost statistiki smertnosti (na primere smertnosti ot travm i otravleny v Moskve)]. Social aspects of public health. [Sotsialnye aspekty zdorovya naseleniya] (Electronic publication), 2007; 2 (2): Available from: URL: http://vestnik.mednet.ru/content/view/28/29/ (in Russian)

5. Semyonova VG., Evdokushkina GN. "Ill-defined" epidemiological crisis ["Netochno oboznachenny" epidemiologichesky krizis]. In: Health of the Russian population in the social context of the 90 years: problems and prospects [Zdorove naseleniya Rossii v sotsialnom kontekste 90-x godov: problemy i perspektivy] (Edt. Starodubov VI, Mikhailov YV, Ivanova AE). Moscow: Medicine, 2003: 85-94. (in Russian)

6. Semyonova VG., Dubrovin EV, Gavrilova NS, Evdokushkina GN, Gavrilov LA. On the problems of trauma mortality in Russia: Kirov region case.
Shapiro J. The Russian mortality crisis and its causes. In: Economic reform at risk (Ed. A Aslund) London, 1995: 149-178.

Shkolnikov V., Mesle F., Vallin J. Health crisis in Russia. Narodonaselenie (Population), 1996; 8: 123-190.

Shkolnikov VM., Chervyakov VV. (Edt.) Policies for the Control of the Transition’s Mortality Crisis in Russia. Moscow, 2000. 191 (in Russian).

Wiseman DA, Dubrovina EV, Redko AN. Information support of research on mortality in Russia [Informatsionnoe obespechenie issledovany po problemam smertnosti v Rossi]. Public health and disease prevention [Obschestvennoe zdrove i profilaktika zabolovaniy], 2006; (6): 31-38. (in Russian)
Estimating evaluation period of a fund manager.

Evgeny A. Ivin, Alexey N. Kurbatskiy, and Alexander V. Slovesnov

Moscow School of Economics, 1, Building 61, Leninskie Gory, M.V. Lomonosov
MSU, Moscow, Russia, 119991. (E-mail: evg.ivin@gmail.com,
akurbatskiy@gmail.com, alexslovesnov@yandex.ru)

Аннотация We compare yields of the leading Russian managed fund with the
MICEX index for estimating the time necessary to make a conclusion about manager’s
performance using martingale approach and we also show that short-term periods are
not optimal a priori and it’s not worth to invest in Russian stocks for only a few
months. It is shown that the MICEX index can’t be considered as a benchmark to
compare different funds and at the end of the paper such universal index is suggested.

Keywords: financial time series, yield, volatility, fund manager, benchmark, martingale.

1 Introduction

In this paper we deal with a problem of estimating the performance of management
of a unit investment trust. This is a crucial problem both for top management of
a managing company which examines investment trust manager’s performance
from the employer’s side and for investor’s who analyze earnings yields or yield-
risk factors of different funds to find the best one in the terms of these rates.
If we speak about the Russian equity market the estimation is usually based
on comparison of fund’s rates with appropriate characteristics of the Russian
benchmark that is the MICEX index. It’s impossible to predict such indicators
of a fund as yield and volatility nevertheless the bounds of their variation quite
often can be forecasted. At the same time for both company’s top management
and investors the decision-making is limited to taking into consideration only
short-term periods because they all aim at instant profit. Every manager fills
his fund with investment ideas but he is able to estimate only approximately
the probability with which they will be realized. It may turn out that some
of his ideas won’t be realized at all or they will be affected by unpredictable
scenarios. Despite the popularity of the risk estimation problem [1,2,3], it so
happened that the problem of valuing manager’s performance has not received
due attention in Russia.

2 Martingale approach

We try to estimate evaluation period using prevalent martingale approach
assuming that values of fund’s net asset value per share (NAVPS) and the
MICEX index are random variables. Let \( x_t \) denotes index value at time \( t \) and
\( x_{t-1} \) is its value at the previous moment. We use weekly data of indices, i.e.
we consider the quote of index at last day of a week, and analyze two time
periods: from 2005 to 2006 and from 2005 to 2012. The first is of interest because there was stability and pretty good growth during the period and the second is necessary for understanding overall picture of what’s going on at the equity market. Let’s choose funds for our research. The entry criteria are as follows: Russia dedicated equity fund (benchmark is the MICEX index); exists from 2005 or earlier; assets under management (on average for the last five years) not less than 300 billion rubles; fund’s strategy hasn’t considerably changed from 2005.

The approach is based on an example from the book [4] (chapter II, §5). Considering its lacunary style we venture more detailed description with some elaborations.

Consider a model in which relative change of the index value \( X_n, \ n = 0, 1, \ldots \), is defined by the recurrence relation

\[
X_{n+1} = \theta \cdot X_n + (1 - \theta) \cdot \xi_{n+1}, \ n \geq 0, \tag{1}
\]

here \( \theta \in R \) is unknown parameter. Random variables \( X_0, \xi_1, \xi_2, \ldots \), are defined in some probability space \((\Omega, \mathcal{F}, P)\) and are considered to be nondegenerate, independent, square integrable and expectation of each innovation \( \xi_n \) is zero. It’s worth to note that the multiplier \((1 - \theta)\) from the right part of (1) is redundant and take part in the equation to make it similar with EWMA model which is widely spread in practical realizations (description of the model see, for example, in [5]).

Consider each value \( X_0, X_1, \ldots \) as observations of a random variable and define the estimator \( \theta_n \) of the parameter \( \theta \) as following:

\[
\theta_n = \frac{\sum_{k=0}^{n-1} X_k X_{k+1}}{\sum_{k=0}^{n-1} D_{k+1}} = D(\xi_k), \ k \in N, \tag{2}
\]

Here we set \( \theta_n \) equal to zero if the denominator vanishes. We will assume further that the condition \( P[\omega : X_0(\omega) = 0] = 0 \) holds. This implies inequality \( \sum_{k=0}^{n-1} D_{k+1} \neq 0 \) with probability 1.

By virtue of the equality (1) we have (with probability 1)

\[
\theta_n = \theta + (1 - \theta) \frac{M_n}{A_n}, \quad M_n = \sum_{k=0}^{n-1} X_k \xi_{k+1}, \quad A_n = \sum_{k=0}^{n-1} D_{k+1}.
\]

A stochastic sequence \( \{M_n\}_{n \geq 1} \) is a martingale with respect to a filtration \( \mathcal{F}_n = \sigma(X_0, \xi_1, \ldots, \xi_n) \) as

\[
E(M_n | \mathcal{F}_{n-1}) = \left( \sum_{k=0}^{n-2} \frac{X_k \xi_{k+1}}{D_{k+1}} + \frac{X_{n-1} \xi_n}{D_n} \right) \mathcal{F}_{n-1}.
\]

\(^1\) All computations are based on the data for NAVPS of funds and the MICEX index from the sources www.investfunds.ru and www.rts.micex.ru.
\[
\sum_{k=0}^{n-2} \frac{X_k \xi_{k+1}}{D_{k+1}} + \frac{X_{n-1}}{D_n} E(\xi_n | F_{n-1}) = M_{n-1}.
\]

The latter equality holds due to \(X_0, \xi_1, \ldots, \xi_n\) mutual independence and the hypothesis \(E \xi_n = 0\). Redefine \(F_0 = \{\emptyset, \Omega\}, M_0 = 0\) and consider the Doob decomposition of a submartingale (see [6,7]) \((M_n^2, F_n)_{n \geq 1}\):

\[
m_n = \sum_{k=1}^n \left( M_k^2 - E(M_k^2 | F_{k-1}) \right), \quad A_n = \sum_{k=1}^n \left( E(M_k^2 | F_{k-1}) - M_k^2 \right).
\]

The predictable process \(A_n\) we transform as follows:

\[
A_n = \sum_{k=1}^n E(M_k^2 - M_{k-1}^2 | F_{k-1}) = \sum_{k=1}^n E(M_k - M_{k-1})^2 | F_{k-1}) =
\]

\[
= \sum_{k=1}^n E \left( \left( \frac{X_{k-1} \xi_k}{D_k} \right)^2 \bigg| F_{k-1} \right) = \sum_{k=1}^n \frac{X_{k-1}^2}{D_k^2} E(\xi_k^2 | F_{k-1}) = \sum_{k=1}^\infty \frac{X_{k-1}^2}{D_k^2}.
\]

Thus the estimator \(\theta_n\) can be decomposed as the sum \(\theta_n = \theta + \frac{M_n}{A_n}\).

Since \(A_1 = E(M_1^2 | F_0) = E(X_0^2 \xi_1^2) = E(X_0^2) \cdot E(\xi_1^2) \neq 0\) then (see [4]) the following conditions are sufficient for the convergence (with probability 1)

\[
\theta_n \rightarrow \theta, n \rightarrow \infty^2,
\]

\[
\sup_n \frac{D_{n+1}}{D_n} < \infty, \quad \sum_{n=1}^\infty E \left[ \min \left( \frac{\xi_n}{D_n}, 1 \right) \right] = \infty.
\]

Returning to the problem of estimating fund’s manager efficiency, let us return to the model (1) for NAV’s relative changes in which random variables \(\xi_1, \xi_2, \ldots\) are identically distributed. In this case the conditions (3) are satisfied and the estimator \(\theta_n = \left( \frac{\sum_{k=0}^{n-1} X_k \xi_{k+1}}{\sum_{k=0}^{n-1} X_k} \right) / \left( \frac{\sum_{k=0}^{n-1} X_k^2}{\sum_{k=0}^{n-1} X_k} \right)\) converges with probability 1 to unknown parameter \(\theta\). To check the applicability of the model (1) for the analyzed fund see the plots of the changes of the estimator \(\theta_n\) during the periods 2005-2006 and 2005-2012.

According to the given plots \(\theta_n\) converges. Moreover at the end of 2008 it is observed the typical splash caused by the market instability which is graded afterwards. This effect can be interpreted as a stability of the model and together with the convergence of \(\theta_n\) justified its application.

The estimated parameter \(\theta\) determines the dependence of future return \(X_{n+1}\) from previous observation \(X_n\) and innovation \(\xi_{n+1}\) and that’s why it might be considered as a valid risk measure. Hence if \(\theta_n\) is close to a true value of a parameter it can be used for valuing manager’s efficiency. Necessary time for reliable conclusions about his quality will be determined by the convergence rate of \(\theta_n\).

\(^2\)To prove this proposition thoroughly one has to strengthen a condition \(A_1 \geq 1\) of the Theorem 4 up to \(A_1 > 0\) (with probability 1) and a condition \(a_n > 0, n \in N\) of the Problem 6 up to \(a_1 > 0; a_n > 0, n \geq 2\).
To analyze the convergence rate turn to the period 2005-2007. Take the last value of the estimator $\theta_n$ which is equal to 0.0889 as a true value of $\theta$ and for each of 12 quarters $K_j$ of this period compute the magnitude

$$\Delta_j = \max_{n \in K_j} |\theta_n - \theta|, \quad j = 1, \ldots, 12.$$  

Corresponding tables are given below:

### Таблица 1. The magnitude $\Delta$ for the Fund 1 from 01.2005 to 06.2006

| Quarter | $K_1$ | $K_2$ | $K_3$ | $K_4$ | $K_5$ | $K_6$ |
|---------|-------|-------|-------|-------|-------|-------|
| Months  | 3     | 6     | 9     | 12    | 15    | 18    |
| $\Delta$ | 0.3570 | 0.0570 | 0.1929 | 0.1730 | 0.1828 | 0.1447 |

| Quarter | $K_7$ | $K_8$ | $K_9$ | $K_{10}$ | $K_{11}$ | $K_{12}$ |
|---------|-------|-------|-------|----------|----------|----------|
| Months  | 21    | 24    | 27    | 30       | 33       | 36       |
| $\Delta$ | 0.6405 | 0.0169 | 0.0159 | 0.0147   | 0.0164   | 0.0037   |

### Таблица 2. The magnitude $\Delta$ for the Fund 1 from 07.2006 to 12.2007

We see that acceptable accuracy of the estimator $\theta_n$ (in terms of convergence to $\theta$) emerged only after 18-21 months of observation and that comply with the previous results. Also notice that the data of other funds lead to the similar
estimators and thus support the necessity of at least 18 months of observation (see a corresponding graph in the next section).

The model (1) makes possible for the investors to compare different funds according to the estimators \( \theta_n \). It’s interesting to note that the estimator of the parameter \( \theta \) computed for the MICEX index can’t be used as a benchmark. Indeed if we look at the graph of the sequence \( \theta_n \) which corresponds to the MICEX index the situation is following:

![Graph](image)

**Pic. 3.** The MICEX index: 2005 - 2007.

According to the plot assessed value \( \theta_n \) becomes negative when \( n \) goes up and consequently it can’t be treated as a valid risk measure. Below we provide a method to build such benchmark.

3 Constructing a universal index

As we mention above estimation of manager’s activity in most cases is based on comparison of fund’s characteristics (mean, variance, Sharpe ratio) with the appropriate characteristics of universal index. Up to here the MICEX index was considered to be the universal one. However in many cases it is more correct to construct such universal benchmark which is for example based on NAV of the largest managed index funds.

The most commonly used way to construct the universal benchmark is still the calculation of the simple mean of the several chosen values. That can be explained by the simplicity of this method and by its possession of many special properties. Below we present the improved method in which the simple mean is replaced by the weighted sum. The weights are selected in such a way that the concluding benchmark has the maximal correlation with the MICEX index. The advantages of this method are as follows. Not all managed funds have the same structure as that of the index (the most often it occurs with the moderate funds). The big funds have another feature: because of their popularity they often meet with active input/output of cash assets which leads to the necessity of carrying out regular operations and consequently effects the yield badly. The influence of such funds on our weighted index will be minimal.

To formulate the problem in mathematical terms we need the following notations. Let \( \vec{v} \) and \( \vec{w} \) be vectors in the euclidian space \( R^n \) and let \( \Sigma \) be a strictly positive definite matrix. By the expressions \( \| \vec{v} \| \) and \( \| \vec{v} \| \) we mean the
norms of the $R^n$ generated by the scalar products

$$(\bar{v}, \bar{w}) = \bar{v}^T \cdot \bar{w} = \sum_{k=1}^{n} v_k w_k \quad \text{and} \quad (\bar{v}, \bar{w})_\Sigma = \bar{v}^T \Sigma \bar{w} = \sum_{i,j=1}^{n} v_i \Sigma_{ij} w_j.$$ 

Let $\xi_1, \xi_2, \ldots, \xi_n, \eta$ be nontrivial stochastic variables which correspond to indices of the largest selected funds and the MICEX respectively. Then the problem stated above is equivalent to finding the vector $\tilde{\alpha}^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_n^*)$ such that

$$\tilde{\alpha}^* = \arg \max_{\alpha \in R^n \setminus \{0\}} \left[ \text{cor}(\alpha_1 \xi_1 + \alpha_2 \xi_2 + \ldots + \alpha_n \xi_n, \eta) \right].$$

Properties of the correlation coefficient imply the equality

$$\text{cor}(\alpha_1 \xi_1 + \alpha_2 \xi_2 + \ldots + \alpha_n \xi_n, \eta) = \frac{\text{cov}(\alpha_1 \xi_1 + \alpha_2 \xi_2 + \ldots + \alpha_n \xi_n, \eta)}{\sqrt{D\eta} \sqrt{D(\alpha_1 \xi_1 + \alpha_2 \xi_2 + \ldots + \alpha_n \xi_n)}} =$$

$$= \frac{\alpha_1 \text{cov}(\xi_1, \eta) + \alpha_2 \text{cov}(\xi_2, \eta) + \ldots + \alpha_n \text{cov}(\xi_n, \eta)}{\sqrt{D\eta} \| \tilde{\alpha} \|_\Sigma},$$

where $\Sigma$ is the covariance matrix of the variables $\xi_1, \xi_2, \ldots, \xi_n$. We assume this matrix to be strictly positive definite and note that the restriction is not a loss of generality. Actually in case of the matrix $\Sigma$ singularity we can state the same problem for a basis of $\{\xi_1, \xi_2, \ldots, \xi_n\}$ which dispersion matrix is strictly positive.

Introducing the notion $\tilde{v} = \left( \text{cov}(\xi_1, \eta), \text{cov}(\xi_2, \eta), \ldots, \text{cov}(\xi_n, \eta) \right)$ we reduce the problem to the maximization of the function

$$f_n(\tilde{\alpha}) = \frac{\alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_n v_n}{\| \tilde{\alpha} \|_\Sigma} = \frac{(\tilde{\alpha}, \tilde{v})}{\| \tilde{\alpha} \|_\Sigma}.$$ 

It differs from the correlation coefficient by the constant $\sqrt{D\eta}$. This function is $\mathcal{O}$-homogeneous as

$$|f_n(\tilde{\alpha})| = |f_n(\lambda \tilde{\alpha})| \quad \forall \lambda \in R \setminus \{0\}.$$ 

The Cauchy-Bunyakovsky inequality implies

$$f_n(\tilde{\alpha}) = \frac{\tilde{\alpha}^T \cdot \tilde{v}}{\| \tilde{\alpha} \|_\Sigma} = \frac{\tilde{\alpha}^T \cdot \Sigma \cdot (\Sigma^{-1} \tilde{v})}{\| \tilde{\alpha} \|_\Sigma} = \frac{(\tilde{\alpha}, \Sigma^{-1} \tilde{v})}{\| \tilde{\alpha} \|_\Sigma} \leq \| \Sigma^{-1} \tilde{v} \|_\Sigma =$$

$$= \left( (\Sigma^{-1} \tilde{v})^T \cdot \Sigma \cdot \Sigma^{-1} \tilde{v} = \tilde{v}^T \cdot \Sigma^{-1} \cdot \tilde{v} = \| \tilde{v} \|_{\Sigma^{-1}} \right).$$

So the vectors $\lambda \cdot \Sigma^{-1} \tilde{v}$, $\lambda \in R \setminus \{0\}$ form the critical set of the function $f_n$. Moreover $f_n$ reaches the maximum $\| \tilde{v} \|_{\Sigma^{-1}}$ at the point $\Sigma^{-1} \tilde{v}$.

Thus we have shown that the problem has explicit solution. However the weights found above could be negative that is the solution could mismatch the
The notion of weighted sum. To find the positive weights one need consider the maximization problem of the function $f_n$ on the set
\[ M = \left\{ (\alpha_1, \ldots, \alpha_n) : \alpha_i \geq 0 \ \forall i = 1, \ldots, n; \sum_{i=1}^{n} \alpha_i = 1 \right\}. \] (4)

The latter can be reduced to the unconditional problem by the special partition of the set $M$.

Let $\delta \subset \{1, 2, \ldots, n\}$ be a nonempty subset of indices. Define the set $M_{\delta}$ in the following way

\[ M_{\delta} = M \cap \{(\alpha_1, \alpha_2, \ldots, \alpha_n) : \alpha_i > 0 \ \forall i \in \delta; \ \alpha_i = 0 \ \forall i \notin \delta\}. \]

Sets $M_{\delta_1}$ and $M_{\delta_2}$ have empty intersection if they correspond to different subsets $\delta_1$ and $\delta_2$. So the set $M$ can be represented as the union $\cup M_{\delta}$ of $2^n - 1$ disjoint sets. Thus the solution $\alpha^* = \arg \max f_n|_M$ exists due to compactness of the set $M$ and lies in some $M_{\delta}$.

1. The set $M_{\delta}$ corresponds to the one-element subset $\delta = \{k\}$. In this case the $M_{\delta}$ contains the single point so $\alpha^* = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$ ($1$ is the $k$-th term of this sequence).

2. The set $M_{\delta}$ corresponds to the subset $\delta$ which cardinality exceeds $1$. In this case note that

\[ f_n(\alpha) \bigg|_{M_{\delta}} = \left(\frac{\bar{\alpha}_{\delta}, \bar{v}_{\delta}}{\|\bar{\alpha}_{\delta}\|\Sigma_{\delta}}\right) = f\bar{\delta}(\bar{\alpha}_{\delta}), \]

where $\Sigma_{\delta}$ is a $\delta \times \delta$ matrix formed by removing from the $\Sigma$ its rows and columns with index out of $\delta$ and $\bar{\alpha}_{\delta}$, $\bar{v}_{\delta}$ are $\delta$-dimension vectors formed from components of $\bar{\alpha}$, $\bar{v}$ with index in the subset $\delta$. Thus the restriction of the function $f_n$ to the set $M_{\delta}$ is of the same type as original one but has the smaller dimension.

As shown above the point $\bar{\alpha}_{\delta}^*$ is critical for $f\bar{\delta}$ and is collinear with the vector $\Sigma_{\delta}^{-1}\bar{v}_{\delta}$.

So the problem of the $f_n|_M$ maximization can be solved algorithmically in the following way. Let $\Delta$ be the set of nonempty subsets $\delta \subset \{1, 2, \ldots, n\}$ such that all components of the vector $\Sigma_{\delta}^{-1}\bar{v}_{\delta}$ are nonnegative and at least one is positive. Define $\delta^*$ as

\[ \delta^* = \arg \max_{\delta \in \Delta} f\bar{\delta}(\Sigma_{\delta}^{-1}\bar{v}_{\delta}). \]

The maximum of the $f_n$ on the set $M$ equals to $f\bar{\delta}(\Sigma_{\delta^*}^{-1}\bar{v}_{\delta^*})$ and is reached at the point $\bar{\alpha}_{\delta}^*$ formed by supplementing the vector $\Sigma_{\delta^*}^{-1}\bar{v}_{\delta^*}$ with $n - |\delta^*|$ zeros.

Now use the universal index as a benchmark for the model (1) to examine one of its the possible application. As an example we take two invest funds and build a sequence $\theta_{n}$ for the period from 2005 to 2007. In previous part we considered a plot and a convergence rate Table of $\theta_{n}$ for the first fund, for the second fund they are given below (the last value on the plot is $\theta = 0, 1308$):

In this case weighting coefficients for the universal index are $\alpha_1 = 0, 2369$, $\alpha_2 = 0, 7631$, and the graph of the sequence $\theta_{n}$ for it is as follows:

The last value $\theta_{13}$ equals to 0, 1129 that is in some way a mean of corresponding values of the funds which are involved in the construction of the index.

489
Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013) International Conference, Mataró (Barcelona), Spain 25 - 28 June 2013

Рис.4. Fund 2: 2005 - 2007.

| Fund 2 01.2005 — 06.2006 |
|---------------------------|
| Quarter | \( K_1 \) | \( K_2 \) | \( K_3 \) | \( K_4 \) | \( K_5 \) | \( K_6 \) |
| Months | 3 | 6 | 9 | 12 | 15 | 18 |
| \( \Delta \) | 0.0718 | 0.1745 | 0.1839 | 0.2035 | 0.1467 | 0.0832 |

Таблица3. The magnitude \( \Delta \) for the Fund 2 from 01.2005 to 06.2006

| Fund 2 07.2006 — 12.2007 |
|---------------------------|
| Quarter | \( K_7 \) | \( K_8 \) | \( K_9 \) | \( K_{10} \) | \( K_{11} \) | \( K_{12} \) |
| Months | 21 | 24 | 27 | 30 | 33 | 36 |
| \( \Delta \) | 0.0219 | 0.0211 | 0.0196 | 0.0269 | 0.0270 | 0.0087 |

Таблица4. The magnitude \( \Delta \) for the Fund 2 from 07.2006 to 12.2007

Рис.5. Universal index: 2005 - 2007.

Список литературы

1. Engle, R.F. (2004). Risk and Volatility: Econometric Models and Financial Practice, American Economic Review, American Economic Association, vol. 94(3).
2. Peters E.E. (1994). Fractal Market Analysis: Applying Chaos Theory to Investment and Economics. New York: Wiley.
3. Peters E.E. (1991). Chaos and Order in the Capital Markets: A New View of Cycles, Prices, and Market Volatility. New York: Wiley.
4. Shiryaev A. N. (2004). Probability v. 1, 2 — 3rd ed. - Moscow: MCCME. (in Russian)
5. Shiryaev A. N., (1999). Essentials of stochastic finance. Facts, models, theory, Advanced Series on Statistical Science and Applied Probability, 3, World Scientific Publishing Co. Inc., River Edge, NJ.
6. Shiryaev A.N. (1995). Probability. 2nd ed. New York: Springer-Verlag.
7. Doob, Joseph Leo (1933). Stochastic Processes. New York, John Wiley and Sons.
Calculation of semi-Markov flow characteristics

Ivnitskiy V.A.

Russia, Moscow State University of Railway Engineering

ivnit@vniizht.ru

Article is devoted to finding arbitrary order initial moments and a correlation functions of the process $\xi(t)$, meaning a time interval to the next time arrival, and of the process $\nu(t)$, meaning a number of arrivals before a time $t$, for semi-Markov flow under general initial conditions.

**Key words and phrases:** semi-Markov flow, semi-Markov process, arbitrary order initial moments, correlation functions, Laplace transformations, general initial conditions.

1. **Introduction**

In article W.L. Smith [1] is pointed out that most of stated… theorems (on renewal theory) were obtained within last 10 years; all they are characterized extreme simplicity conditions and wide community of applications. It is necessary to notify that we will have affair only with limit-theorems by simple reason, that it is had very little results relating to renewal processes and possessing sufficient community, that could not carry character of limit-theorems.

In Discussion on Dr. Smith’s paper Prof. M.S. Bartlett said: “…I could put in a plea for the user of renewal theory, for whom limit-theorems and “finite” results… are equally important [1, p. 258]. The present paper is devoted of receipt results taking place for finite time $t$ to semi-Markov flows.

Semi-Markov processes and flows attracted attention investigators on queuing systems already more 50 years back. Their sufficiently detailed description is had to [2]. Queuing systems with such flows are investigated to [3]. Queuing systems with Semi-Markov service are investigated to [4–5]. In paper they are obtained arbitrary order initial moments and a correlation functions of the process $\xi(t)$, meaning a time interval to the next time arrival, and of the process $\nu(t)$, meaning a number of arrivals before a time $t$, for semi-Markov flow under general initial conditions.
For obtaining final results they are used systematically one- and two-measure Laplace transforms.

Let $X$ - final or accounting multitude with elements $i_1, i_2, \ldots$. It is had homogeneous Markov chain $\{v_n, n \geq 1\}$ with values to $X$ and transition matrix $((p_{ij}))$. Semi - Markov process is random process $v_i(t), t \geq 0$, with properties: - 1) to half-interval $[t_i, t_{i+1})$ $v_i(t) = v_{i+1}, i = 0, 1, \ldots, t_0 = 0$, 2) under fixed realization Markov chain $v_n = i_k, n \geq 1$, sojourn values $t_1, t_2 - t_1, \ldots$ process $v_i(t)$ to states $i_1, i_2, \ldots$. positive and independent, sojourn value $z_{v_n, v_{n+1}}$ depend on state $v_n$, in which process $v_1(t)$ is, and on next state $v_{n+1}$; it are had distribution functions $P(\{z_{v_n, v_{n+1}} < x/v_n = i, v_{n+1} = j\} = F_{ij}(x), n \geq 1$. Times $t_0, t_1, t_2, \ldots$ make semi-Markov flow.

They are possible three type dependence semi-Markov flow from $v(0)$, which has distribution $P(\nu(t) = i) = p_i, \sum_{i=0}^{\infty} p_i = 1$. This distribution don’t dependent from semi-Markov flow. First type dependence - customers of flow are added to $v(0)$.

Cross to second type dependence semi-Markov flow from $v(0)$. To time $t_0 = 0$ it is set initial state Markov chain $k_0$, with probabilities they are played $p_{k_0, k}, k \in X$, it is next state Markov chain $k_1$, it is determined random value $z_{k_0, k_1}$ with distribution function $F_{k_0, k_1}(x)$, time $t_1$ is $t_1 = t_0 + z_{k_0, k_1}$, on this first step ended. Further second step go et cetera, but customers flow don’t act. To time $t = 0$ it is played realization $v(0)$. If significance $i$ falls out, then first arrival of flow enters on $i + 1$ step to time $z_{k_i, k_{i+1}}$, having distribution function $F_{k_i, k_{i+1}}(x)$ and so on. Arrivals are added to $v(0)$.

Third type dependence semi-Markov flow from unites both first type and besides depends on $v(0)$ following manner. Distribution functions lengths of intervals between successive times arrivals will depend on $v(0)$ that is they will have form $F_{k_i, k_{i+1}}(x, v(0)), i = 0, 1, 2, \ldots$. 492
Let random process $\xi(t)$ is residual interval time to the next arrival time of semi-Markov flow and $\nu(t)$ is the number of arrivals before the time $t$ of semi-Markov flow.

It is necessary to obtain the non-stationary arbitrary order moments of the $\xi(t)$, and $\nu(t)$, also the correlation functions of these processes for above-mentioned three types dependence semi-Markov flow from $\nu(0)$.

2. Calculation non-stationary state probability distribution of random process $\{k(t), \xi(t)\}$

Notice that $\xi(t)$ is residual time from $t$ to next arrival of flow. We will use the method of stochastic difference equations (MSDE) [6–9] to obtain the initial moments. One can summarise its essence as follows. It is suggested to construct the difference equations at once for the initial moments of Markov process of desired order first, then to use the operator of the mathematical expectation, then to take the limit and finally obtain a system of differential equations for the moments. This method turns out to be more effective for obtaining the arbitrary order non-stationary moments, but area of it applications is more narrow.

Show its applying on proof next theorem. Let $P\{X_{dk} = 1\} = p_{dk}$ and $P\{X_{dk} = 0\} = 1 - p_{dk}$. Then $M I(X_{dk} = 1) = p_{dk}$. Following theorem holds.

**Theorem 1.** The following system of linear differential equation with private derivatives for $\varphi_k(t,x)$ holds
\[
\frac{\partial}{\partial t} \phi_k(t, x) - \frac{\partial}{\partial x} \phi_k(t, x) = - \frac{\partial}{\partial x} \phi_k(t, 0) + \sum_{d \in X} \frac{\partial}{\partial x} \phi_d(t, 0) p_{dk} F_{dk}(x), n \geq 1, \ k = 1, 2, \ldots, n,
\]

(2.1)

with initial conditions \( \phi_l(0, x) = \varphi_l^{(0)}(x) \geq 0, \sum_{k \in X} \varphi_l^{(0)}(x) = 1. \)

Proof. Use MSDE. We have stochastic difference equation for \( I(\xi_k(t) < x) \)

\[
I(\xi_k(t + \Delta t) < x - \Delta t) = I(\xi_k(t) < x) - I(\xi_k(t) < \Delta t) + \sum_{d \in X} I(\xi_d(t) < \Delta t) I(\chi_{dk} = 1) I(z_{dk} < x) + o(\Delta t) \Rightarrow \]

\[
- I(\xi_k(t) < \Delta t) + \sum_{d \in X} I(\xi_d(t) < \Delta t) I(\chi_{dk} = 1) I(z_{dk} < x) + o(\Delta t). \quad (2.2)
\]

Here \( o(\Delta t) \) mean that expectation next members expansion is \( o(\Delta t) \).

Applying to (2.2) expectation, divide (2.2) by \( \Delta t \) and let \( \Delta t \to 0 \) obtain (2.1). Theorem 1 is proved.

Имеем \( M \xi(t) = \sum_{k \in X} M I(\xi_k(t) < x) = \sum_{k \in X} \phi_k(t, x) \). We will find a system differential equations decision (2.1).

Put \( \tilde{\phi}_k(s, u) = \int_0^\infty e^{-at-st} d_s \phi_k(t, x), \frac{\partial}{\partial s} \tilde{\phi}_k(u, 0) = \int_0^\infty e^{-at}d \frac{\partial}{\partial x} \phi_k(t, 0), \tilde{\phi}_k^{(0)}(s) = \)

\[
\int_0^\infty e^{-at}d_s \phi_k^{(0)}(x), \tilde{f}_{dk}(s) = \int_0^\infty e^{-at}d_x F_{dk}(x), \ d, k \in X .
\]

**Theorem 2.** The following formula holds for \( \tilde{\phi}_k(s, u) \)

\[
\tilde{\phi}_k(s, u) (u - s) = - \frac{\partial}{\partial x} \tilde{\phi}_k(u, 0) + \sum_{d \in X} \frac{\partial}{\partial x} \tilde{\phi}_d(u, 0) p_{dk} \tilde{f}_{dk}(s) + \tilde{\phi}_k^{(0)}(s), \ k \in X , \quad (2.3)
\]

where \( \frac{\partial}{\partial x} \tilde{\phi}_k(u, 0), \ k \in X \), is obtained by the following system of linear algebraic equations

\[
- \frac{\partial}{\partial x} \tilde{\phi}_k(u, 0) + \sum_{d \in X} \frac{\partial}{\partial x} \tilde{\phi}_d(u, 0) p_{dk} \tilde{f}_{dk}(u) + \tilde{\phi}_k^{(0)}(u) = 0, \ k \in X . \quad (2.4)
\]

Proof. Applying Laplace-Stieltjes transform to (2.1) we get (2.3) and unknown variables \( \frac{\partial}{\partial x} \tilde{\phi}_k(u, 0), \ k \in X \). The left-hand part of (2.3) vanishes at
\( s = u \) because \( \tilde{\varphi}_k(s,u) \) is analytic in the domain \( \text{Re}\{s \geq 0\}, \text{Re}\{u > 0\} \). Hence, the right-hand part of (2.3) should equal zero at \( s = u \) as well. This gives us the system linear algebraic equations (2.4) for calculation unknown variables

\[
\frac{\partial}{\partial x} \tilde{\varphi}_k(u,0) \quad k \in X \ , \quad (2.4) \text{is solved known methods. Theorem 2 is proved.}
\]

### 3. Residual time to the next arrival

Use MSDE to \( x_k(t) = \tilde{x}(t)I(k(t) = k) \) for obtaining \( \tilde{a}_k^{(i)}(u) = \int_0^\infty e^{-ut} M \tilde{x}_k(t)^i \, dt \).

MSDE is possible apply by quantity \( x_k(t) = \tilde{x}(t)I(k(t) = k) \) for calculation non-stationary initial moment of residual time arbitrary order.

Put \( \tilde{a}_k^{(i)}(u) = \int_0^\infty e^{-ut} M \tilde{x}_k(t)^i \, dt \).

**Theorem 3.** Let moments \( M\tilde{x}_k(0)^i \) and \( Mz_{dk}^i \), \( d,k \in X \), exist. Then the next formula holds for the Laplace transform \( \tilde{a}_k^{(i)}(u) \), where \( i \) - the arbitrary finite whole number

\[
\tilde{a}_k^{(i)}(u) = \sum_{j=2}^i (-1)^{i-j} \frac{j!}{u^{i-j+1}} (Mz_{dk}^j) + \sum_{d \in X} \frac{\partial}{\partial x} \tilde{\varphi}_d(u,0) p_{dk} Mz_{dk}^j + (-1)^{i-j} \frac{j!}{u^{i-j} \tau_0}
\]

\[
\quad + \sum_{d \in X} \frac{\partial}{\partial x} \tilde{\varphi}_d(u,0) p_{dk} Mz_{dk}^j - u^{-j}, \quad k \in X \ , \quad (3.1)
\]

where quantities \( \frac{\partial}{\partial x} \tilde{\varphi}_d(u,0), \quad d \in X \), are obtained from the system equations (2.4) p. 2.

**Proof.** For \( i = 1 \) have

\[
\tilde{x}_k(t + \Delta t) = (\tilde{x}_k(t) + \Delta t) I(\tilde{x}_k(t) > \Delta t) + \sum_{d \in X} I(\tilde{x}_d(t) < \Delta t) I(\chi_{dk} = 1)z_{dk} + o(\Delta t) \Rightarrow
\]

\[
\tilde{x}_k(t + \Delta t) = (\tilde{x}_k(t) + \Delta t)(1 - I(\tilde{x}_k(t) < \Delta t)) + \sum_{d \in X} I(\tilde{x}_d(t) < \Delta t) I(\chi_{dk} = 1)z_{dk} + o(\Delta t) \Rightarrow
\]

\[
\tilde{x}_k(t + \Delta t) - \tilde{x}_k(t) = -\Delta t - (\tilde{x}_k(t) - \Delta t) I(\tilde{x}_k(t) < \Delta t) + \sum_{d \in X} I(\tilde{x}_d(t) < \Delta t) I(\chi_{dk} = 1)z_{dk}
\]

\[ + o(\Delta t) \Rightarrow \tilde{x}_k(t + \Delta t) - \tilde{x}_k(t) = -\Delta t + \sum_{d \in X} I(\tilde{x}_d(t) < \Delta t) I(\chi_{dk} = 1)z_{dk} + o(\Delta t) \ . \quad (3.2)
\]

Applying the mathematical expectation operator to (3.2) and dividing the result by \( \Delta t \) we get (3.3) as \( \Delta t \to 0 \)
\[
\frac{\partial}{\partial t} M \xi_k(t) = \sum_{d \in X} \frac{\partial}{\partial X} \phi_d(t,0) p_{dk} M z_{dk} - 1, \quad k \in X,
\] (3.3)

with initial condition \( M \xi_k(0) = \tau_0 = \sum_{d \in X} \phi_d^{(0)}(\infty) \tau_d^{(0)}, \quad \tau_d^{(0)} = \int_0^\infty xd \phi_d^{(0)}(x). \)

Applying to (3.3) Laplace transform, obtain
\[
\tilde{\xi}_k^{(i)}(u) = u^{-1} \left[ (\tau_0 + \sum_{d \in X} \frac{\partial}{\partial X} \tilde{\phi}_d(u,0) p_{dk} M z_{dk} - u^{-1}) \right], \quad k \in X.
\] (3.4)

Cross to calculation non-stationary initial moment of residual time arbitrary order from \( \xi_k(t) \). From (3.2) we have
\[
\xi_k(t + \Delta t) = (\xi_k(t) - \Delta t) I(\xi_k(t) > \Delta t) + \sum_{d \in X} I(\xi_d(t) < \Delta t) I(\chi_{dk} = 1)z_{dk} + o(\Delta t).
\]

Raising both sides of (3.2) to power \( i \), where \( i \) - arbitrary finite whole number, we get
\[
\xi_k(t + \Delta t)^i = (\xi_k(t) - \Delta t)^i + \left( \sum_{d \in X} I(\xi_d(t) < \Delta t) I(\chi_{dk} = 1)z_{dk} \right)^i + o(\Delta t) \Rightarrow \xi_k(t + \Delta t)^i = (\xi_k(t) - \Delta t)^i - I(\xi_k(t) < \Delta t) (\xi_k(t) - \Delta t)^i + \sum_{d \in X} I(\xi_d(t) < \Delta t) I(\chi_{dk} = 1)z_{dk}^i + o(\Delta t) \Rightarrow
\]
\[
\xi_k(t + \Delta t)^i - \xi_k(t)^i = -i \xi_k(t)^i \Delta t + \sum_{d \in X} I(\xi_d(t) < \Delta t) I(\chi_{dk} = 1)z_{dk}^i + o(\Delta t).
\] (3.5)

Applying to (3.5) the mathematical expectation operator, dividing the result by \( \Delta t \) we get (3.6) as \( \Delta t \to 0 \)
\[
\frac{\partial}{\partial t} M \xi_k(t)^i = \sum_{d \in X} \frac{\partial}{\partial X} \phi_d(t,0) p_{dk} z_{dk}^i - i M \xi_k(t)^i - 1, \quad k \in X.
\] (3.6)

under initial condition \( M \xi_k(0)^i \). Applying Laplace transform obtain
\[
u \tilde{a}_k^{(i)}(u) = M \xi_k(0)^i - i \tilde{a}_k^{i-1}(u) + \sum_{d \in X} \frac{\partial}{\partial X} \tilde{\phi}_d(u,0) p_{dk} M z_{dk}^i.
\] (3.7)

Formula (3.7) is a recurrence relation w.r.t. \( k \). Uncovering it with use (3.4) obtain (3.1). Theorem 3 is proved.

We will find the correlation function random process \( \{k(t), \xi(t)\} \) - residual time to next arrival under Markov chain state \( k(t) \). For this it is necessary to find a covariance matrix \( (M \xi_k(t) \xi_j(t') - M \xi_k(t) M \xi_j(t')) \). The invers transform of (3.4) gives \( M \xi_k(t) \). It is necessary to obtain \( M \xi_k(t) \xi_j(t') \). Put \( t' > t \) .

496
Find the correlation matrix \( (M \xi_k(t) \xi_j(t')) - M \xi_k(t) M \xi_j(t') \)). Turning the Laplace transformation, obtaining from (3.8) under \( i = 1 \), find \( M \xi_k(t) \). Now it is necessary to obtain \( M \xi_k(t) \xi_j(t') \). Let \( t' > t \).

**Theorem 4.** The following formula holds for \( M \xi_k(t) \xi_j(t') \)

\[
M \xi_k(t) \xi_j(t') = \int_{t'-t}^{\infty} x^2 d_x \varphi_k(t,x) - (t' - t) \int_{t'-t}^{\infty} xd_x \varphi_k(t,x) + \sum_{m=0}^{\infty} \left( \int_{t'-t}^{\infty} \int_{t'-t-y}^{\infty} x^2 d_x \varphi_k(t,x) dF_{m+1}(u) d\Psi_m(y) - \int_{t'-t-y}^{\infty} \int_{t'-t-y}^{\infty} xd_x \varphi_k(t,x) dF_{m+1}(u) d\Psi_m(y) \right),
\]

(3.9)

where \( F_l(y) = \sum_{d_l=1}^{n} p_{d_l,ld_l} F_{d_l,ld_l}(y) \), \( l = 1,\ldots,m+1 \), \( \Psi_m(y) \) - composition of distributions \( F_l(y),\ldots,F_m(y) \).

**Proof.** There are two possibilities.

1) Assume: \( \xi_k(t) > t' - t \). In this case we have \( \xi_k(t) - \xi_k(t') = t' - t \), whence \( \xi_k(t') = \xi_k(t) - (t' - t) \). Then \( M \xi_k(t) \xi_k(t') I( \xi_k(t) > t' - t ) = M \xi_k(t) \times (\xi_k(t) - (t' - t)) I( \xi_k(t) > t' - t ) = M \xi_k(t)^2 I( \xi_k(t) > t' - t ) - (t' - t) M \xi_k(t) \times I( \xi_k(t) > t' - t ) \).

Turn out double Laplace-Stieltjes transform obtained in theorem 3.1 find non-stationary distribution function of residual time to next arrival \( \varphi_k(t,x) \) and condition that to time \( t \) state Markov chain is \( k \). Then

\[
M \xi_k(t)^2 I( \xi_k(t) > t' - t ) = \int_{t'-t}^{\infty} x^2 d_x \varphi_k(t,x), \quad M \xi_k(t) I( \xi_k(t) > t' - t ) = \int_{t'-t}^{\infty} xd_x \varphi_k(t,x).
\]

(3.9)

Therefore,

\[
M \xi_k(t) \xi_k(t') I( \xi_k(t) > t' - t ) = \int_{t'-t}^{\infty} x^2 d_x \varphi_k(t,x) - (t' - t) \int_{t'-t}^{\infty} xd_x \varphi_k(t,x). \quad (3.10)
\]
2. If, on contrary, $\xi_k(t) < t' - t$, then we have a denumerable set of possibilities

Determined by the following inequalities:

$$t' - t - \sum_{l=1}^{m+1} \sum_{d_i \in X} I(\mathcal{X}_{d_i, \eta_l}) = 1)z_{d_i, \eta_l} < \xi_k(t) < t' - t - \sum_{l=1}^{m} \sum_{d_i \in X} I(\mathcal{X}_{d_i, \eta_l}) = 1)z_{d_i, \eta_l}, \quad m = 0, 1, 2, \ldots$$

(3.11)

In this case we have

$$\sum_{l=1}^{m+1} \sum_{d_i \in X} I(\mathcal{X}_{d_i, \eta_l}) = 1)z_{d_i, \eta_l} + \xi_k(t) - \xi_j(t') = t' - t, \quad d_0 = k, d_{m+1} = j,$$

so $\xi_j(t') = \xi_k(t) - (t' - t) + \sum_{l=1}^{m+1} \sum_{d_i \in X} I(\mathcal{X}_{d_i, \eta_l}) = 1)z_{d_i, \eta_l}$. Then

$$MI(t' - t - \sum_{l=1}^{m+1} \sum_{d_i \in X} I(\mathcal{X}_{d_i, \eta_l}) = 1)z_{d_i, \eta_l} < \xi_k(t) < t' - t - \sum_{l=1}^{m} \sum_{d_i \in X} I(\mathcal{X}_{d_i, \eta_l}) = 1)z_{d_i, \eta_l}) \xi_k(t) \times$$

$$= \int_{t' - t} \int_{y} \int_{t' - t} x(x + y + u - (t' - t))d_{\eta_j} \varphi_j(t, x)dF_{m+1}(u)d\Psi_m(y), \quad (3.12)$$

Where $F_i(y) = \sum_{d_i \in X} p_{d_i, \eta_l} G_{d_i, \eta_l}(y), \quad l = 1, \ldots, m+1$, $\Psi_m(y)$ is the $m$-fold convolution of the probability distribution functions $F_1(y), \ldots, F_m(y)$.

Equality (3.12) can be written as

$$MI(t' - t - \sum_{l=1}^{m+1} \sum_{d_i \in X} I(\mathcal{X}_{d_i, \eta_l}) = 1)z_{d_i, \eta_l} < \xi_k(t) < t' - t - \sum_{l=1}^{m} \sum_{d_i \in X} I(\mathcal{X}_{d_i, \eta_l}) = 1)z_{d_i, \eta_l}) \xi_k(t) \times$$

$$= \int_{t' - t} \int_{y} \int_{t' - t} x^2 d_{\eta_j} \varphi_j(t, x)dF_{m+1}(u)d\Psi_m(y) -$$

$$+ \int_{t' - t} \int_{y} \int_{t' - t} (t' - t) \int_{0} \int_{t' - t - y} x^2 d_{\eta_j} \varphi_j(t, x)dF_{m+1}(u)d\Psi_m(y) + \int_{t' - t} \int_{y} \int_{t' - t - y} x^2 d_{\eta_j} \varphi_j(t, x)dF_{m+1}(u)d\Psi_m(y). \quad (3.13)$$

Hence,
\[ M \xi_k(t) \xi_j(t') = M \xi_k(t) \xi_j(t') I(\xi_k(t) > t' - t) + \sum_{m=0}^{\infty} MI(t' - t - \sum_{l=1}^{m} \sum_{d_i=1}^{d} I(\chi_{d_i,x} = 1)z_{d_i,x} < \xi_k(t) < t' - t - \sum_{l=1}^{m} \sum_{d_i=1}^{d} I(\chi_{d_i,x} = 1)z_{d_i,x} \xi_k(t) \xi_j(t'). \] (3.14)

Substituting formulae (3.10) and (3.13) into (3.14) we get (3.6). Theorem 4 is proved.

The following formula holds for the correlation function \( R_{\xi(t)}(t,t') \) of \( \xi(t) \)

\[ R_{\xi(t)}(t,t') = \sum_{k \in X} \sum_{j \in X} M \xi_k(t) \xi_j(t') - \sum_{k \in X} M \xi_k(t) \sum_{j \in X} M \xi_j(t'). \] (3.15)

4. Calculation mathematical expectation of arrivals

semi-Markov flow

Let us consider the process \( \nu(t) \) which is not markovian. One can extend the process \( \nu(t) \) by two supplementary variables: above-mentioned \( \xi(t) \) and \( \overline{\nu}(t) \) - state Markov chain at time \( t \). For markovization of process \( \nu(t) \) we shall use supplementary variables without increase of dimension \( \nu(t) \). We call this method Markovization without increase of dimension (MWID for short).

Essence of this method in the present case consists in the multiplication of the initial process \( \nu(t) \) on the indicators of event \( I(\overline{\nu}(t) = k) \) and the event \( I(\xi_k(t) < x) \), that is we use the Markov processes \( \nu_k(t,x) = \nu(t)I(\xi_k(t) < x) I(\overline{\nu}(t) = k) \), \( k \in X \), and apply to it the MSDE. Put \( n_k(t,x) = M \nu_k(t,x) \), \( k \in X \).

**Theorem 5.** The following system of linear differential equations with private derivatives for \( n_k(t,x) \), \( k \in X \), holds

\[
\frac{\partial}{\partial t} n_k(t,x) - \frac{\partial}{\partial x} n_k(t,x) = - \frac{\partial}{\partial x} n_k(t,0) + \sum_{d \in X} \frac{\partial}{\partial x} n_d(t,0) p_{dk} F_{dk}(x) + \\
\sum_{d \in X} \frac{\partial}{\partial x} \varphi_d(t,0) p_{dk} F_{dk}(x) \quad k \in X, \]

(4.1)
with initial conditions \( n_m(0, x) = n_m^{(0)}(x), n_m(0, \infty) = n_m^{(0)}, m \in X \), where \( \frac{\partial}{\partial x} \varphi_j(t, 0), \)
d \( \in X \), is obtained by turning their Laplace transformations, which is obtained by
system of linear algebraic equations (2.4) theorem 2, \( \varphi_k^{(0)}(x) = n_k^{(0)}(x) / n_k^{(0)}, k \in X \).

**Proof.** Vector, consisting from quantities \( v_k(t, x) = v(t)I(\xi(t) < x) \times I(\nu(t) = k), k \in X \), is Markov random process and for it can usual way make up stochastic difference equation, that is have

\[
v_k(t + \Delta t, x - \Delta t) = v_k(t, x)I(\Delta t < \xi(t) < x) + \sum_{d \in X} (v_d(t) + 1)I(\xi(t) < \Delta t)I(\chi_{dk} = 1)I(z_{dk} < x) + o(\Delta t) =
\]

\[
I(z_{dk} < x) + \sum_{d \in X} I(\xi(t) < \Delta t)I(\chi_{dk} = 1)I(z_{dk} < x) + o(\Delta t),
\]

(4.2)

where \( \chi_{dk}, d, k \in X \), - totality on \( d, k \) events, отражающих cross Markov chain from some state to other forming full group and independent both from \( \xi(t) \)
and from \( v_d(t) \) and \( v_k(t), \chi_{dk} \) can have 2 significances: 0 or 1, \( M I(\chi_{dk} = 1) = p_{dk}, \sum_{k \in X} p_{dk} = 1 \) under fixed \( d \in X \), \( v_d(t) = v(t)I(\nu(t) = d). \)

**Remark 4.1.** We have identities \( v_k(t, x)I(\xi(t) < x) = v_k(t, x) \) and \( v_k(t, x) \times I(\xi(t) < \Delta t) = v_k(t)I(\xi(t) < \Delta t). \)

Moving \( v_k(t, x) \) into left-hand side of (4.2), applying the mathematical expectation operator to (4.2) and dividing the result by \( \Delta t \) we get (4.1) as \( \Delta t \to 0 \)
where \( \sum_{d \in X} \frac{\partial}{\partial x} \varphi_d(t, 0) p_{dk} F_{dk}(x) \) is result of application the mathematical expectation
operator to \( \sum_{d \in X} I(\xi_d(t) < \Delta t)I(\chi_{dk} = 1)I(z_{dk} < x). \) Theorem 5 is proved.

Now let us solve the partial differential equation (4.1). Set

\[
\tilde{n}_k(u, s) = \int_0^\infty \int_0^\infty e^{-ut-sx} d_x n_k(t, x)dt , \frac{\partial}{\partial x} \tilde{n}_k(u, 0) = \int_0^\infty e^{-ut} \frac{\partial}{\partial x} n_k(t, 0)dt , \tilde{n}_k^{(l)}(u) = \int_0^\infty e^{-ut} \times
\]

\[
Mv(t)^l dt , \tilde{n}_k^{(0)}(s) = \int_0^\infty e^{-sx} d_x n_k^{(0)}(x), k \in X \}, \tilde{f}_{dk}(s) = \int_0^\infty e^{-sx} d_x F_{dk}(x), l = 1, 2, ..., k, d \in X.
\]
Theorem 6. The Laplace transformation $\tilde{n}_k(u,s)$ is obtained by the formula

$$
\tilde{n}_k(u,s)(u-s) = -\frac{\partial}{\partial x}\tilde{n}_k(t,0) + \sum_{d \in X} \frac{\partial}{\partial x}\tilde{n}_d(u,0) \ p_{dk} \tilde{f}_{dk}(s) + \\
\sum_{d \in X} \frac{\partial}{\partial x}\tilde{\varphi}_d(u,0) \ p_{dk} \tilde{f}_{dk}(s) + \tilde{n}_k^{(0)}(s), \ k \in X,
$$

where $\frac{\partial}{\partial x}\tilde{n}_k(u,0)$, $k \in X$, is obtained by system of linear algebraic equations

$$
-\frac{\partial}{\partial x}\tilde{n}_k(u,0) + \sum_{d \in X} \frac{\partial}{\partial x}\tilde{n}_d(u,0) \ p_{dk} \tilde{f}_{dk}(u) + \sum_{d \in X} \frac{\partial}{\partial x}\tilde{\varphi}_d(u,0) \ p_{dk} \tilde{f}_{dk}(u) + \tilde{n}_k^{(0)}(u), \ k \in X,
$$

(4.3)

$$
\frac{\partial}{\partial x}\tilde{\varphi}_d(u,0), \ d \in X, \ is \ obtained \ by \ system \ of \ linear \ algebraic \ equations \ (4).
$$

Proof. Applying Laplace transform and the Laplace-Stieltjes transform to (4.1) we get (4.3). The left-hand part of (4.3) vanishes at $s=u$ because $\tilde{n}_k(u,s)$ is analytic in the domain $\text{Re}\{s \geq 0\}, \text{Re}\{u > 0\}$. Hence, the right-hand part of (4.3) should equal zero at $s=u$ as well. This gives us system of linear algebraic equations (4.4). Deciding it we get $\frac{\partial}{\partial x}\tilde{n}_k(u,0)$, $k \in X$.

Theorem 6 is proved.

Consequence 1. The following formula holds for $\tilde{n}(u)$

$$
\tilde{n}(u) = \sum_{k=1}^{\infty} \tilde{n}_k(u),
$$

(4.5)

where $\tilde{n}_k(u) = u^{-1}(-\frac{\partial}{\partial x}\tilde{n}_k(u,0) + \sum_{d=1}^{\infty} \frac{\partial}{\partial x}\tilde{n}_d(u,0) \ p_{dk} + \sum_{d \in X} \frac{\partial}{\partial x}\tilde{\varphi}_d(u,0) \ p_{dk} + n_k^{(0)})$, $k \in X$, $\frac{\partial}{\partial x}\tilde{n}_d(u,0)$ and $\frac{\partial}{\partial x}\tilde{\varphi}_d(u,0)$, $d \in X$, are obtained by theorem 6.

5. Calculation second initial moment of arrivals

semi-Markov flow

Cross to calculation second initial moment of arrivals semi-Markov flow $v(t)$. Put $n_k^{(l)}(t,x) = M \ v_k^{(l)}(t,x)$, $n_k^{(l)}(t) = M \ v_k^{(l)}(t)$, $n^{(l)}(t) = M \ v^{(l)}(t)$, $n^{(l)}(t) = \\
\sum_{k \in X} n_k^{(l)}(t) \frac{\partial}{\partial x} n_k^{(l)}(0) = \frac{\partial}{\partial x} n_k(t,0), \ l = 2,3, \ldots, \ k \in X$. 

501
\textbf{Theorem 7.} The following system of linear differential equations with private derivatives for $n_k^{(2)}(t,x)$, $k \in X$, holds

$$\frac{\partial}{\partial t} n_k^{(2)}(t,x) - \frac{\partial}{\partial x} n_k^{(2)}(t,x) = - \frac{\partial}{\partial x} n_k^{(2)}(t,0) + \sum_{d \in X} \left( \frac{\partial}{\partial x} n_d^{(2)}(t,0) + 2 \frac{\partial}{\partial x} n_d(t,0) + \frac{\partial}{\partial x} \varphi_d(t,0) \right) p_{dk} F_{dh}(x), \quad k \in X, \tag{5.1}$$

with initial conditions $n_m^{(2)}(0,x) = n_m^{(0,2)}(x)$, $n_m^{(0,2)}(0,\infty) = n_m^{(0,2)}$, $m \in X$, where $\frac{\partial}{\partial x} n_d(t,0)$ and $\frac{\partial}{\partial x} \varphi_d(t,0)$, $d \in X$, are obtained by theorem 6.

\textbf{Proof.} Raising both sides of (4.2) p. 4 to power 2 with registration remark 4.1 p. 4 we get

$$v_k(t + \Delta t, x - \Delta t)^2 = (v_k(t, x)I(\xi_k(t) < x) - v_k(t, x)I(\xi_k(t) < \Delta t) + \sum_{d \in X} (v_d(t) + 1)I(\xi_d(t) < \Delta t)

I(\chi_{dk} = 1) I \left( z_{dk} < x \right) + o(\Delta t))^2 = v_k(t, x)^2 - 2 v_k(t, x) \times (v_k(t, x)I(\xi_k(t) < \Delta t) - \sum_{d \in X} (v_d(t) + 1)I(\xi_d(t) < \Delta t)I(\chi_{dk} = 1) I \left( z_{dk} < x \right) + (v_k(t, x)I(\xi_k(t) < \Delta t) - \sum_{d \in X} (v_d(t) + 1)I(\xi_d(t) < \Delta t)I(\chi_{dk} = 1)

I(\xi_k(t) < \Delta t) + 2 v_k(t, x) \sum_{d \in X} (v_d(t) + 1)I(\xi_d(t) < \Delta t)I(\chi_{dk} = 1) I \left( z_{dk} < x \right) + v_k(t, x)^2 \times

I(\xi_k(t) < \Delta t) - 2 v_k(t, x)I(\xi_k(t) < \Delta t) \sum_{d \in X} I \left( z_{dk} < x \right) (v_d(t) + 1)I(\xi_d(t) < \Delta t)I(\chi_{dk} = 1)

+ \sum_{d \in X} (v_d(t) + 1)I(\xi_d(t) < \Delta t)I(\chi_{dk} = 1) I \left( z_{dk} < x \right) + o(\Delta t) = v_k(t, x)^2 - I(\xi_k(t) < \Delta t) v_k(t)^2

+ 2 v_k(t, x) v_k(t) + 1)I(\xi_k(t) < \Delta t) I(\chi_{kk} = 1) I \left( z_{kk} < x \right) - 2 v_k(t, v_k(t) + 1)I(\xi_k(t) < \Delta t) \times

I(\chi_{kk} = 1) I \left( z_{kk} < x \right) + \sum_{d \in X} (v_d(t) + 1)I(\xi_d(t) < \Delta t)I(\chi_{dk} = 1) I \left( z_{dk} < x \right) + \sum_{d \in X} \sum_{I \in I, l \in d} (v_d(t) + 1)(v_l(t) + 1)I(\xi_d(t) < \Delta t)I(\xi_l(t) < \Delta t)I(\chi_{dk} = 1)I(\chi_{lk} = 1) \times

I \left( z_{dk} < x \right) I(\xi_k(t) < \Delta t) + o(\Delta t) = v_k(t, x)^2 - v_k(t)^2 I(\xi_k(t) < \Delta t) + \sum_{d \in X} (v_d(t) + 1)^2 I(\xi_d(t) < \Delta t)I(\chi_{dk} = 1) I \left( z_{dk} < x \right) + o(\Delta t). \tag{5.2}
To these transforms was take into consideration the following identities:
\[ I(\xi_k(t) < \Delta t) = I(\xi_k(t) < \Delta t) I(\bar{v}(t) = k) \) and \( I(\xi_d(t) < \Delta t) = I(\xi(t) < \Delta t) I(\bar{v}(t) = d) \),
that is \( I(\xi_k(t) < \Delta t) I(\xi_d(t) < \Delta t) = 0 \), as events \( I(\bar{v}(t) = k) \) and \( I(\bar{v}(t) = d) \) are incompatibles. Then
\[
2v_k(t, x) \sum_{d \in X} \left( v_d(t) + 1 \right) I(\xi_d(t) < \Delta t) I(\chi_{dk} = 1) I(z_{dk} < x) = 2v_k(t, x) I(\xi_k(t) < \Delta t) \times
\]
\[
\sum_{d \in X} \left( v_d(t) + 1 \right) I(\xi_d(t) < \Delta t) I(\chi_{dk} = 1) I(z_{dk} < x) = 2v_k(t, x) I(\chi_{ik} = 1) I(z_{ik} < x) \times
\]
\[
(v_k(t) + 1) I(\xi_k(t) < \Delta t).
\]
On the same considerations \( \sum_{d=1}^n \sum_{l \in X, l \neq d} I(z_{dk} < x) I(z_{lk} < x) \times
\]
\[
(v_d(t) + 1)(v_l(t) + 1) I(\xi_d(t) < \Delta t) I(\xi_l(t) < \Delta t) I(\chi_{dk} = 1) I(\chi_{lk} = 1) = 0.
\]
Thus we have equality
\[
v_k(t + \Delta t, x - \Delta t)^2 = v_k(t, x)^2 - v_k(t)^2 I(\xi_k(t) < \Delta t) + \sum_{d \in X} I(z_{dk} < x) \times
\]
\[
(v_d(t) + 1)^2 I(\xi_d(t) < \Delta t) I(\chi_{dk} = 1) + o(\Delta t).
\] (5.3)

Moving \( v_k(t, x)^2 \) into left-hand side of (5.3) and dividing the result by \( \Delta t \) we get (5.1) as \( \Delta t \to 0 \)

Theorem 7 is proved.

Put \( \tilde{n}_k^{(1)}(u, s) = \int_0^\infty \int_0^\infty e^{-ut-sx} n_k^{(1)}(t, x) dt, \)
\[
\frac{\partial}{\partial x} \tilde{n}_k^{(1)}(u, 0) = \int_0^\infty e^{-ut} \frac{\partial}{\partial x} n_k^{(1)}(t, 0) dt,
\]
\[
n_k^{(0,l)}(0, x) = n_k^{(0,l)}(x), \ n_k^{(0,l)}(0, \infty) = n_k^{(0,l)}, \ \tilde{n}_k^{(0,l)}(s) = \int_0^\infty e^{-sx} d n_k^{(0,l)}(x), \ k \in X, \ l = 2, 3, ..., \]

Theorem 8. The Laplace transformation \( \tilde{n}_k^{(2)}(u, s) \) is obtained by formula
\[
\tilde{n}_k^{(2)}(u, s) = -\frac{\partial}{\partial x} \tilde{n}_k^{(2)}(u, 0) + \sum_{d \in X} \left( \frac{\partial}{\partial x} \tilde{n}_d^{(2)}(u, 0) + 2 \frac{\partial}{\partial x} \tilde{n}_d(u, 0) \right) p_{dk} \tilde{f}_d(s) + \frac{\partial}{\partial x} \tilde{\phi}_d(u, 0) \right) p_{dk} \tilde{f}_d(s) + \tilde{n}_k^{(0,l)}(s), \ k \in X, \]
where \( \frac{\partial}{\partial x} \tilde{n}_k^{(2)}(u, 0), \ k \in X, \) is obtained by system of linear algebraic equations

503
\[
\frac{\partial}{\partial x} \tilde{n}^{(2)}_k(u,0) = \sum_{d \in X} \left( \frac{\partial}{\partial x} \tilde{n}^{(2)}_d(u,0) + 2 \frac{\partial}{\partial x} \tilde{\phi}_d(u,0) \right) + \frac{\partial}{\partial x} \tilde{f}_d(u) + \frac{\partial}{\partial x} \tilde{\phi}_d(u,0) \right) p_{dk} \tilde{f}_k(u) + \tilde{n}^{(0,2)}_k(u),
\]

(5.5)

\[k \in X, \frac{\partial}{\partial x} \tilde{n}_d(u,0) \text{ and } \frac{\partial}{\partial x} \tilde{\phi}_d(u,0), d \in X, \text{ are obtained by theorem 6.}\]

Proof. Applying Laplace transform and the Laplace-Stieltjes transform to (5.1) we get (5.4). The left-hand part of (5.4) vanishes at \( s = u \) because \( \tilde{n}^{(2)}_k(u,s) \) is analytic in the domain \( \text{Re}\{s \geq 0\}, \text{Re}\{u > 0\} \). Hence, the right-hand part of (5.4) should equal zero at \( s = u \) as well. This gives us system of linear algebraic equations (5.5) relatively \( \frac{\partial}{\partial x} \tilde{n}^{(2)}_k(u,0) \). Deciding it we get

\[\frac{\partial}{\partial x} \tilde{n}^{(2)}_k(u,0), k \in X. \text{ Theorem 8 is proved.}\]

Consequence 2. The following formula holds for \( \tilde{n}^{(2)}(u) \)

\[\tilde{n}^{(2)}(u) = \sum_{k=1}^{\infty} \tilde{n}^{(2)}_k(u), \quad (5.6)\]

where \( \tilde{n}^{(2)}_k(u) = u^{-1} \left( - \frac{\partial}{\partial x} \tilde{n}^{(2)}_k(u,0) + \sum_{d \in X} \left( \frac{\partial}{\partial x} \tilde{n}^{(2)}_d(u,0) + 2 \frac{\partial}{\partial x} \tilde{\phi}_d(u,0) \right) x p_{dk} + n^{(0,2)}_k \right), k \in X, \frac{\partial}{\partial x} \tilde{n}^{(2)}_k(u,0), k \in X, \text{ are obtained by theorem 8, } \frac{\partial}{\partial x} \tilde{n}_d(u,0), \frac{\partial}{\partial x} \tilde{\phi}_d(u,0) \text{ are obtained by theorem 6.}\]

Theorems 2, 6 and 8 form recurrent correlations for \( \frac{\partial}{\partial x} \tilde{n}^{(2)}_k(u,0), \frac{\partial}{\partial x} \tilde{n}_d(u,0) \) and \( \frac{\partial}{\partial x} \tilde{\phi}_d(u,0) \).

Turning \( \tilde{n}(u) \) and \( \tilde{n}^{(2)}(u) \), \( n(t) \) and \( n^{(2)}(t) \) are obtained. Then

\[D\nu(t) = n^{(2)}(t) - (n(t))^2.\]

6. Calculation initial moment of order \( m \) of arrivals

semi-Markov flow

Cross to calculation initial moment of order \( m \) of process \( \nu(t) \), where \( m \) - arbitrary finite whole positive number.
Theorem 9. The following system of linear differential equations with private derivatives for \( n^{(m)}_k(t,x) \), \( k \in X \), \( m = 1,2,3,... \), holds

\[
\frac{\partial}{\partial t} n^{(m)}_k(t,x) - \frac{\partial}{\partial x} n^{(m)}_k(t,x) = - \frac{\partial}{\partial x} n^{(m)}_k(t,0) + \sum_{d=1}^{n} \left( \sum_{l=1}^{m} C^l_m \frac{\partial}{\partial x} n^{(l)}_d(t,0) n^{(l)}_d(t,0) + \frac{\partial}{\partial x} \phi_d(t,0) \right) p_{dk} F_{dk}(x), \quad k \in X, \tag{6.1}
\]

with initial conditions \( n^{m}_k(0,x) = n^{(0,m)}_k(x), \quad n^{(0,m)}_k(0,\infty) = n^{(0,m)}_k, \quad k \in X, \quad m = 1,2,3,... \),

where \( \frac{\partial}{\partial x} n^{(m)}_k(u,0), \quad \frac{\partial}{\partial x} \tilde{n}^{(2)}_k(u,0), \quad d \in X \), are obtained by theorem 6 and 8 and

\[
\frac{\partial}{\partial x} \tilde{\phi}_d(u,0), \quad d \in X, \text{ are obtained by theorem 2,} \quad \frac{\partial}{\partial x} n^{(l)}_d(u,0), \quad l > 2, \text{ are obtained by theorem 10.}
\]

Theorem 10. The Laplace transformation \( \tilde{n}^{(m)}_k(u,s) \) is obtained by formula

\[
\tilde{n}^{(m)}_k(u,s) (u-s) = - \frac{\partial}{\partial x} \tilde{n}^{(m)}_k(u,0) + \sum_{d=1}^{n} \left( \sum_{l=1}^{m} C^l_m \frac{\partial}{\partial x} \tilde{n}^{(l)}_d(u,0) + \frac{\partial}{\partial x} \tilde{\phi}_d(u,0) \right) p_{dk} \tilde{f}_{dk}(s) + \tilde{n}^{(0,m)}_k(u),
\]

where \( \frac{\partial}{\partial x} \tilde{n}^{(m)}_k(u,0), \quad k \in X, \) are obtained by following system of linear algebraic equations

\[
- \frac{\partial}{\partial x} \tilde{n}^{(m)}_k(u,0) + \sum_{d=1}^{n} \frac{\partial}{\partial x} \tilde{n}^{(m)}_d(u,0) p_{dk} \tilde{g}_{dk}(u) + \sum_{d=1}^{n} \left( \sum_{l=1}^{m-1} C^l_m \frac{\partial}{\partial x} \tilde{n}^{(l)}_d(u,0) + \frac{\partial}{\partial x} \tilde{\phi}_d(u,0) \right) \times p_{dk} \tilde{g}_{dk}(u) + \tilde{n}^{(0,m)}_k(u) = 0, \quad k = 1,2,...,n, \tag{6.2}
\]

which present recurrent correlations, expressing \( \frac{\partial}{\partial x} \tilde{n}^{(l)}_k(u,0) \) across \( \frac{\partial}{\partial x} \tilde{n}^{(l)}_k(u,0) \),

\( l > l_1, \quad \frac{\partial}{\partial x} \tilde{n}^{(l)}_d(u,0), \quad d \in X, \text{ are obtained by theorem 6,} \quad \frac{\partial}{\partial x} \tilde{\phi}_d(u,0) - \text{theorem 2.}
\]

Consequence 3. The following formula holds for \( \tilde{n}^{(m)}(u) \)

\[
\tilde{n}^{(m)}(u) = \sum_{k=1}^{\infty} \tilde{n}^{(m)}_k(u), \tag{6.3}
\]

where \( n^{(m)}_k(u) = u^{-1} (- \frac{\partial}{\partial x} \tilde{n}^{(m)}_k(u,0) \frac{\partial}{\partial x} \tilde{n}^{(m)}_k(u,0) + \sum_{d=1}^{n} \left( \sum_{l=1}^{m} C^l_m \frac{\partial}{\partial x} \tilde{n}^{(l)}_d(u,0) + p_{dk} \tilde{f}_{dk}(s) \times \frac{\partial}{\partial x} \tilde{\phi}_d(u,0) \right) + \tilde{n}^{(0,m)}_k), \quad l > 0, \text{ and } d \in X, \text{ are obtained by theorem 10.}
\]
Find the correlation matrix \((Mv_k(t)v_j(t'))\). Turning the Laplace transformation, obtaining from (10), find \(Mv_k(t)\). Now it is necessary to find \(Mv_k(t)v_j(t')\). Let \(t' > t\).

**Theorem 11.** The following formula holds for \(Mv_k(t)v_j(t')\)

\[
Mv_k(t)v_j(t') = n_k^{(2)}(t' - t) - n_k^{(2)}(t'),
\]

\[
Mv_k(t)v_j(t') = \sum_{i=1}^{\infty} \sum_{m=0}^{\infty} \sum_{d_1 \in X}^{\ldots} \sum_{d_m \in X} I(v(t) = i)I(v(t') = i + m + 1) \times
\]

\[
\int_0^{\infty} \int_0^{\infty} (n_k^{(2)}(t, t' - t - y_1)) - n_k^{(2)}(t, t' - t - y_2)) dF_{d_1, \ldots, d_m}(y_1) dG_{d_1, \ldots, d_m}(y_2) +
\]

\[
(m + 1) \int_0^{\infty} \int_0^{\infty} (n_k^{(1)}(t, t' - t - y_1)) -
\]

\[
n_k^{(1)}(t, t' - t - y_1 - y_2)) dF_{d_1, \ldots, d_m}(y_1) dG_{d_1, \ldots, d_m}(y_2),
\]

where \(G_{d_1, d_l}(y) = p_{d_1, d_l}F_{d_1, d_l}(y), d_1 = k, d_m+1 = j, l = 2, \ldots, m+1, \quad F_{d_1, \ldots, d_m}(y)\) - composition of distributions \(G_{d_1, d_2}(y), \ldots, G_{d_m, d_m}(y)\).

The following formula holds for the correlation function \(R_{v(t), v(t')}\) of \(v(t)\)

\[
R_{v(t), v(t')} = \sum_{k \in X} \sum_{j \in X} Mv_k(t)v_j(t') - \sum_{k \in X} Mv_k(t) \sum_{j \in X} Mv_j(t').
\]

**7. Registration of dependence on initial state**

Make registration of dependence on initial state. First type dependence – customers of flow are added to \(v(0)\). This make simple.

Second type dependence on \(v(0)\) in conformity with its description in beginning of paper also make simple.

Third type dependence semi-Markov flow from unites both first type and besides depends on \(v(0)\) following manner. Distribution functions lengths of intervals between successive times arrivals will depend on \(v(0)\) that is they will have form \(F_{k, k_{i+1}}(x, v(0)), i = 0, 1, 2, \ldots\) Its realization is a few unwieldy and it is to take down to introduction realization \(i\) random value \(v(0)\) in distribution functions \(F_{k, k_{i+1}}(x, i), i = 0, 1, 2, \ldots\).
References

1. Smith W.L. “Renewal theory and its ramifications”, Journal of the Royal Statistical Society, B, 20, (1958).

2. Gnedenko B.V., Kovalenko I.N. Introduction to queuing theory. -M.: Nauka, Fizmatlit, 1987. P. 336.

3. Dudin A.N., Klimenok V.I., Kim C.S., Lee M.H. The SM/PH/N queueing system with broadcasting service // Proceedings of the 13th International conference on analytical and stochastic modeling techniques and applications. – Bonn, Germany, 2006. – P. 8–13

4. Cinlar E. Time dependence of queues with semi-Markovian services // Appl. Probab. – 1967. V. 4. – P. 356–364.

5. Sengupta B. The semi-markovian queue: theory and applications / Stochastic Models. – 1990. – V. 6. – № 3. – P. 383–413.

6. Ivnitskiy V.A. Non-stationary Markov networks theory. V. 1 Closed queuing networks/ Preface Kovalenko I.N. M.:URSS, 2010. –P. 400. (in Russian).

7. Ivnitskiy V.A. Non-stationary Markov networks theory. V. 2 Open queuing networks/ Preface Kovalenko I.N. M.:URSS, 2010. –P. 408. (in Russian).

8. Ivnitskiy V.A. General input flow theory with applications to non-stationary queuing systems and networks / Preface Kovalenko I.N. Palmarium Academic Publishing, Saarbrucken, Germany, - 2012. – P. 641 . (in Russian).

9. Ivnitskiy V.A. Recurrent modeling diskrete-continuous Markov processes/ Preface Kovalenko I.N. Palmarium Academic Publishing, Saarbrucken, Germany, - 2013. – P. 624 . (in Russian).
