The use of MDS and HCA enabled pharmacists to reveal their roles which reflect on country of practice and cultural differences when improving patients' adherence to asthma medication

Andrea Manfrin[1], Cecilio Mar Molinero[2], Luca Cacciolatti[2]

[1] Medway School of Pharmacy University of Kent, Central Avenue, ME4 4TB Chatham Maritime, Kent UK.
(Email: a.manfrin@kent.ac.uk)
[2] Kent Business School, University of Kent, CT2 7PE Canterbury, UK
(Email: cm235@kent.ac.uk), (Email: l.cacciolatti@kent.ac.uk)

Abstract: Pharmacists are health care professionals who have a lot more to offer to the public rather than just selling pills. Pharmacists live a dichotomy role: health care consultants versus shop keepers. The aim of this research was to assess how pharmacists intervenes improving patient adherence to asthma medication and to see how pharmacists perceived themselves. The research was conducted by means of a multiple choice questionnaire that was completed by 551 pharmacists. The replies were analysed using Ordinal Multidimensional Scaling and Cluster Analysis, with the routine PROXSCAL. Pharmacists revealed their roles in terms of four dimensions: a community health orientation, theoretical-applied drug orientation, the degree of involvement with the patient and the role of the pharmacist as a consultant. These dimensions have different salience in each one of the countries.
Keywords: pharmacy, country, asthma, adherence multidimensional scaling.

1 Introduction

In Europe, over 30 million people have asthma. The total annual cost of asthma was estimated in 2003 to be €17.7 billion (£15 billion) in Europe (Loddenkemper et al 2003)[1]. The pharmacist is usually the last health professional in contact with such patients giving advice, and will often be the first point of contact in the detection of adverse drug reactions (ADRs). One of the key tasks set for community pharmacists is to improve patients' knowledge of their disease, and to give advice on the use of medicines in order to increase the effectiveness of therapy. However, although there is a common role to all pharmacists, there appears not to be a cross-country study of cultural differences in pharmacy management. We report here on a study aimed at filling this void. The aim of the study was to investigate the activities of community pharmacist in relation to asthma medication. The research instrument was an on-line survey of community pharmacists in Italy, Switzerland, and the UK. In total, 551 pharmacists replied to the survey. Answers to survey questions were coded in the form of keywords, and analysed by means of multivariate statistical techniques, in particular Multidimensional Scaling (MDS), and Cluster Analysis. MDS was chosen because it allows the visualisation of the results, while keeping a solid mathematical basis. After this section we report on the way in which the questionnaire was designed. This is followed by a technical section in which, after giving summary statistics, we explain how the MDS model was implemented and interpreted. The paper ends with a discussion of the results.

2 Terminologies

There are two technical terms relating to pharmacy practice that need to be first defined: adherence and compliance. Adherence is the extent to which the patient's behaviour matches agreed recommendation from the prescriber. Compliance is the extent to which a patient's behaviour matches the prescriber's recommendations (NCCSDO, 2005)[2]. The difference between the two terms relies on the fact that compliance denotes a relationship where the doctor decide on the appropriate treatment (authoritarian role) and the patient follows the doctor's order (passive role), while adherence emphasizes the concept of an agreement between doctor and patient.

3 Aims

The aim of the research was to explore community pharmacist's role (activity) in relation to encouraging adherence to prescribed medication for patients with asthma, and how these roles are culture and country dependent.

4 The research instrument (questionnaire)

The questionnaire had 11 questions, the first asked about the country in which the practice was situated, as one of the aims of the research was to explore international differences in pharmacy practice. Question 2 was aimed at evaluating what pharmacists understood as adherence, and to exploring whether they understood the difference between compliance and adherence. Questions 5 and 9 tried to establish how pharmacists identified patients that were non-adherent with their asthma medications and the most common reasons why patients did not take their asthma medications properly. Question 4 was set to establish whether pharmacists gave advice on the correct inhaler technique only to new patients, only to established patients, both or neither. Question 6 was aimed at providing information on what tools pharmacists have at their disposal to improve and encourage patients' correct use of their asthma medication. Questions 7 and 10 explored pharmacists' perceived level of involvement in improving patients' use of medicines. Questions 3 and 8 looked at pharmacists' level of confidence in giving advice on inhaler techniques and at their perceptions on the result of such advice. The last question, Question 11, was aimed to understanding, which were for pharmacists, the three most positive outcomes of improving patient's adherence. The questionnaire was distributed in November 2009, and responses were collected until the end of March 2010. It was designed with three language options: English, French and Italian. It was sent through Qualtrics® system to UK Swiss and Italian pharmacists. The MUR (Medicine Use Review) database held by the Medway School of Pharmacy (UK) was used to distribute the questionnaire in the UK. In Italy and Switzerland the survey was e-mailed through the pharmaceutical associations of Pordenone (Friuli Venezia Giulia Region, Italy) and Canton Ticino in Switzerland. One reminder was sent to UK pharmacists but no reminders were sent to Swiss and Italian chemists. Of the 6611 emails that were sent to UK pharmacists, 370 replied giving a response rate of 5.6%. In Switzerland, 38 pharmacists answered the survey giving a response rate of 10% and in Italy 103 replied giving a response rate of 27%. Country of origin could not be established in 40 questionnaires.

5 Data analysis

Responses to questions were coded into the package PASW, version 19 (previously SPSS). Each respondent was treated as a case. A set of keywords was derived from the replies to the questions. All keywords were coded with the values zero or one; one, if the respondent had replied positively to the question, and zero otherwise. When a question had several possible answers, a keyword was allocated to each option. The data set was a matrix of 551 questionnaires by 58 keywords. All entries in this matrix were either zero or one. Missing country values were coded as zero.

5.1 Multidimensional scaling

The data set of pharmacies by keywords was analysed by means of Ordinal Multidimensional Scaling (MDS) with the routine PROXSCAL, although standard menus could not be used, and a routine was coded using the syntax facility. Our interest was not in comparing pharmacists, but in finding out if there were patterns in the way the questions were answered by pharmacists. We were particularly interested in finding out which questions were relevant to particular countries and not to others. MDS is a multivariate statistical technique that starts by calculating a measure of similarity between the entities of interest, the keywords. The measure of similarity used was proposed by Russell and Rao (Yin and Yasuda, 2005)[3]. The Russell and Rao measure assesses the "proximity" between two keywords by counting the number of times that respondents had replied "yes" to both keywords. This procedure is similar to Mar Molinero and Xie (2007)[4]. A table of similarities (proximities) between pairs of keywords was obtained in this way. The next step was to plot the keywords in the space in such a way that when the measure of similarity between two keywords is high, the keywords are located next to each other in the space. When the similarity between two keywords is low, they are located far apart in the space. The space in which the keywords are located may have more than two dimensions, and it is necessary to work with projections on sets of two dimensions. When working with more than two dimensions, it is necessary to complement MDS with other statistical techniques in order to interpret the results. The quality of the statistical representation varies with the number of dimensions of the space in which the keywords are located. The higher the number of dimensions the better the quality of fit. There are various ways of assessing goodness of fit. Stress1 was chosen as the measure of quality of fit. Stress1 is based on the proportion of unexplained variance. We used the ordinal approach to build the configuration, untying tied observations, (Coxon, 1982)[5]. The value of Stress1 was calculated for six dimensions: D6 (Stress 1= 0.109), D5 (Stress 1 = 0.124), D4 (Stress 1 = 0.139), D3 (Stress 1 = 0.173), D2 (Stress 1 = 0.224), and D1 (Stress 1= 0.331). As the number of dimensions increases, the value of Stress1 decreases. The values of Stress 1 in D5 and D6 did not improve massively and for this reason rather than estimate the model again in four dimensions, we used the first four dimensions of the six dimensional representations. The projection of

the six dimensional map in D1 and 2 can be seen in Figure 1, and the projection of the six dimensional map in D3 and 4 is reproduced in Figure 2.

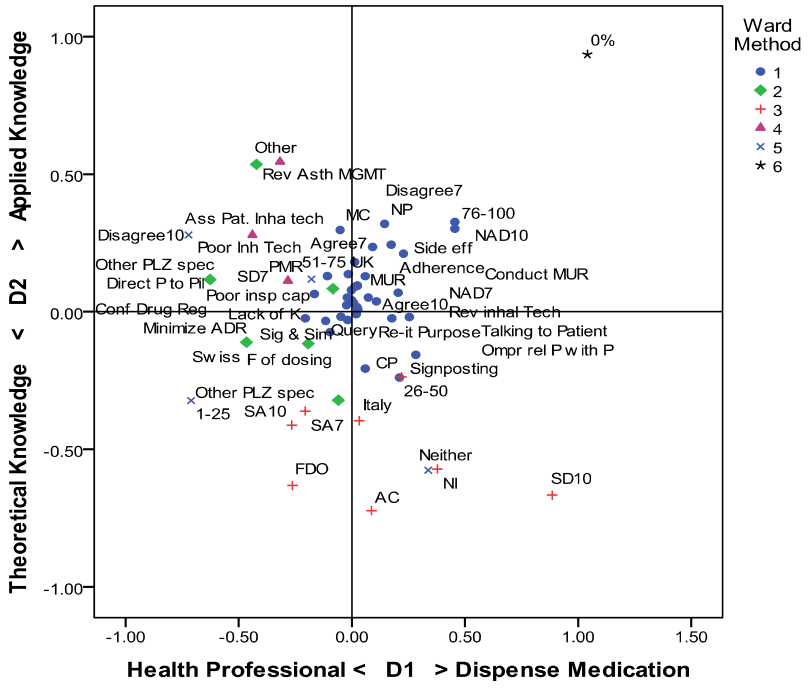


Figure 1. Representation in Dimensions 1 and 2 with Ward cluster membership

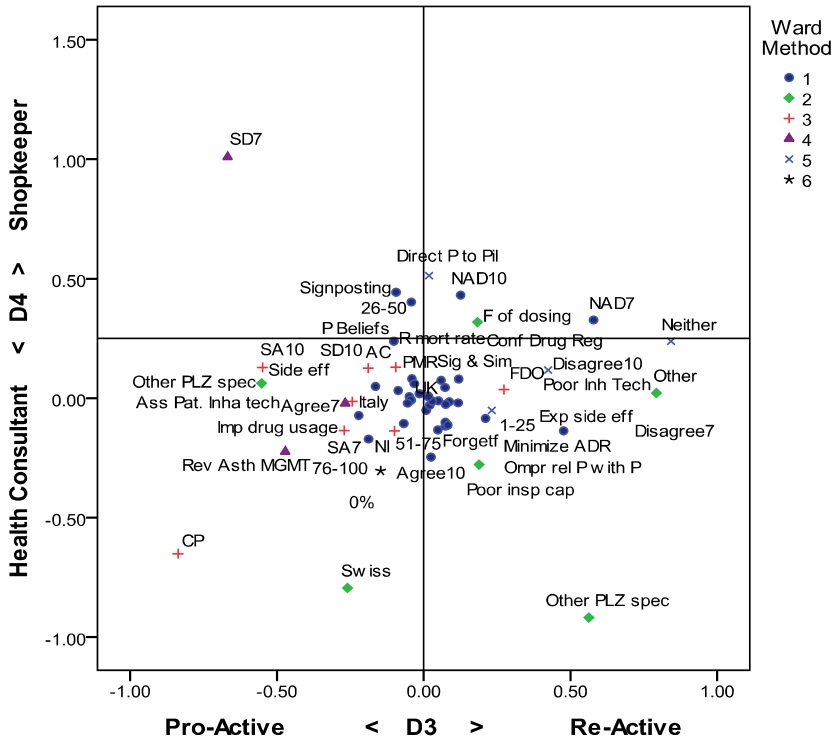


Figure 2. Representation in Dimensions 3 and 4 with Ward cluster membership

It is now possible to attach meaning to the dimensions in the plot. We will start with D1. On the far left hand side of Figure 1 we find that pharmacists find out if patients are not using the medicine properly

reviewing asthma management plan, checking PMR (patient medication record), looking for "Sign and Symptoms"; they intervene assessing patient inspiration capability and minimizing adverse drug reaction. From the centre to the left side of the D1 axis we found pharmacists that strongly agree and agree to the fact that is the responsibility of the pharmacist to ensure that the patient takes the medicine properly; and that the pharmacist thinks that a positive outcome of improving patient adherence is to "minimise adverse drug reaction" as well. At the right hand side of Figure 1 we find a disagreement with the view that the pharmacist has to consider patient's adherence more often; and, for example, that when the pharmacist discovers that the patient is not taking the medicine correctly, takes the action of "signposting to another health care professional"; that pharmacists neither agree nor disagree regarding to the fact that it is their "responsibility to ensure patients use their medication properly" and neither agree nor disagree that they "need to consider patient adherence more often". This dimension appears to be discriminating between pharmacists that see their role as community health professionals, and pharmacists that see themselves as just in charge of dispensing medicines as prescribed by the medical doctor. Using similar considerations we interpreted D2 as pharmacists mainly looking at drug from the theoretical point of view (bottom) which means a general knowledge of drug such as frequency of dosing, to a pharmacists looking at drug under the application point of view (patients' implication) identifying specific issues such as assessing patient inhalation technique, poor inspiration capability. D2 is associated with a higher drug orientation form theoretical knowledge (negative side of the axis) to applied knowledge (positive side of the axis) about drugs. It appears also very clear the link between MUR and UK pharmacists which reinforce the concept of the applied knowledge as well. The presence of the keyword "PMR" (patient medication history) is reinforcing the concept of the knowledge which is applied in a real situation. UK and Swiss pharmacists have PMR in place, but Italian pharmacists are not allowed to keep PMR due data protection regulation. Dimension 3 was found to be related to the degree of pharmacists' involvement in the management of the disease: on the negative side of this axis, the pharmacists are actively involved (Pro-Active), assessing and reviewing which means taking the initiative; on the right hand side, the pharmacists tend to state why poor adherence can occur (Re-Active). Dimension 4 appears to be related to the extent to the role of the pharmacist as a health consultant, the lower end of this dimension being associated with a higher involvement in health advice: health consultant. The opposite end is showing pharmacists who do not tend to get engaged with patients (shopkeeper behaviour). The MDS analysis has revealed that pharmacists see their role in four dimensions: as community health professionals – prescription dispenser; as people with theoretical-applied knowledge; in terms of their relationship with the patient; and as health consultants. We can see in Figures 1 and 3 that the keyword "Swiss" is located towards the left hand side of D1, and just below the middle of D2, towards the left of D3, and towards the negative side of D4. This would indicate that Swiss pharmacists can be described as community health oriented (D1), with a slightly heavier component of theoretical knowledge, with pro-active attitude involved in the management of the patient (D3), perceiving themselves health consultant. The keyword "UK" is in the middle of D1, towards the positive side of D2, it is also in the middle of D3 and just at the beginning of the negative value of Dimension. This could be explained by the fact that UK pharmacists perceived themselves across two systems: being involved in the management of the patient and dispensing prescriptions where the weight of the two components seems to be balanced. The presence of UK in the positive area o D2 underpins the focus of UK pharmacists on the applied knowledge. They still are in the negative part of D4, but they perceived themselves less health care consultant compared to their Swiss colleagues. With regards to the Italian pharmacists they appeared to be salient in D1 and very close to the middle of this dimension but slightly skewed to the right. The location of Italian pharmacists in D2 is clear: the theoretical knowledge is predominant versus the applied knowledge. D3 shows the same level of involvement in patient management such as the Swiss; D4 placed them slightly below the UK but far from the Swiss meaning that they did not perceive themselves having a consultant role such as their Swiss colleagues.

5.2 Cluster Analysis

Having attached meaning to the dimensions, and having used these meanings in order to explore the health management culture of pharmacists in the three countries from which information was obtained, it is also relevant to establish if there are questions, or groups of questions, that are answered in the same way by pharmacists. We are particularly interested in knowing which positive answers to questions serve to differentiate pharmacists in the three countries. To assess the proximity of the points in the space we used Cluster Analysis. Distances between keywords in the six dimensional spaces were calculated from the coordinates in the MDS configuration using the Euclidean metric. We used Hierarchical Cluster Analysis with the method suggested by Ward, since it maximises the homogeneity within clusters and the heterogeneity between clusters. The output of the analysis is a dendrogram and it is a matter of

judgement to decide how many clusters are relevant. One is looking for "long branches" in the diagram, since these reveal the heterogeneity of the clusters. As the branches become shorter the clusters increase in homogeneity, but also in number. In the end we settled for six clusters. Cluster 6 and Cluster 5 stand out as discordant from the rest. They both relate to single variables, and summarise responses from very few pharmacists. There will not be any further discussion on these two clusters. Cluster 1 group together very many answers to questions related to the pharmacist as a health care provider. The questions included are: "How do you find out if a patient is not using their asthma medication properly" which reflect the patient's attitude towards poor adherence. Very close to this are the answers to the question "How would you intervene when you know a patient is not using their medication properly" which represent the way in which the pharmacists is acting to improve patient adherence to medication. Conducting Medicine Use Review (MUR) appears in this cluster despite the fact that MUR was a formalised service only in UK at the time of this study. In this cluster there is a strong representation of another question: "What do you think are the reasons patient may not take their medication properly". This question is particularly relevant because it shows that pharmacists believe that non-intentional non adherence is very common among patients and it is represented by forgetfulness, misunderstanding, and lack of knowledge. The presence of the last question "What do you think are the 3 major positive outcomes of improving patient adherence" was answered positively by pharmacists who selected improving quality of life, reduce mortality rate, and minimize adverse drug reaction (ADR). Pharmacists located in the UK are represented in this cluster. Cluster 2 relates to the pharmacists' attitude with respect to the patients' problems. This cluster includes statement which shows what pharmacists think are the reasons patients may not take their medication properly, which are frequency of dosing and poor inspiration capability. In this cluster we find Swiss pharmacists. Cluster 3 group pharmacists who misunderstood the meaning of adherence and confused with compliance. They tend to give advice to continuing patient and they confirmed that they need to improve their confidence when giving advice about inhaler technique. Other variable found in this cluster where SA10 and SA7 which are indicating the level of agreement with the fact that pharmacists needs to get involved in medicine management. This confirms the positive pharmacists' attitude towards patient issues. Here we find Italian pharmacists. Finally, the questions that come together under Custer 4 are related to the way in which pharmacists assess if patients are not using their medication properly, and the way in which the pharmacists' act when they find out. Pharmacists who belong to this cluster tend to review asthma management plan and assess patient inspiration capability.

5.3 Classification of pharmacists

Figures 1 and 2 bring together the MDS representation of keywords and their classification using Cluster Analysis. It is to be noticed that there is little overlap between the clusters, something that facilitates interpretation. We can see in figure 1 that Cluster 1 is mainly located at the centre of the axis and UK pharmacists are represented in it. They seem to have good applied knowledge and rely in between the health professional and the pharmacists involved in the dispensing procedures as well. In figure 2 Cluster 1 is mainly concentrated in the lower part of D4 (health consultant) but it is also between the pro-active and re-active dimension. Probably UK pharmacists are acting according to patients' requirement. In figure 1, cluster 2 is located on the negative axis of D1 but it seems to be present in either part of D2. This revels that Swiss pharmacists see themselves as health professionals where the knowledge seems to be spread across theoretical and applied. Figure 3 does not revel very much about cluster although it is mainly present in the negative part of D4 (health consultant). Cluster 3 appears evident in figure 2 and represents the Italian pharmacists; the theoretical knowledge is the dominant element in this cluster and they perceive themselves between health professional and dispensing activity. The dispensing component seems slightly heavier compared to the UK ones. This cluster is also present in figure 3, it is well represented in the pro-active part of D3 (negative axis) and in the health consultant (negative axis) of D4.

6 Conclusions

The visual interpretation has been able to demonstrate the way in which pharmacists reveal themselves through the questionnaire, and the cultural differences that exist in these three countries. There is much debate on the role of pharmacists as a keystone in the provision of heath. The WHO (World Health Organization) together with FIP (International Pharmaceutical Federation) discussed the role that pharmacists should aim for, and observed that pharmacists have been always linked with manufacturing and supplying of medications, but due to the increasing demand for health services, they are facing new and challenging roles. Such future roles were discussed in a conference on the 23rd November 2006 in Geneva (Switzerland) where it was observed that the complexity of new medications coupled with poor

adherence to prescribed medications is contributing to a shift of the pharmacy model from drug-based to patient-based; Wiedenmayer et al, 2006. The above suggests that there are various ways in which pharmacy can be practiced. It is therefore appropriate to ask up to what point differences in the way in which the practice of pharmacy takes place can be discerned by looking at different cultural settings. This issue has been explored by means of a questionnaire completed by pharmacists in the UK, Italy, and Switzerland. The analysis of questionnaire answers revealed that there are four dimensions to pharmacy: a community health orientation (health professional-dispense medication); theoretical-applied knowledge; the degree of involvement with the patient (pro-active – re-active); and the role of the pharmacist as a consultant-shopkeeper. It was also identified the positive answers to questions that are relevant to each type of pharmacist. It was observed that Swiss pharmacists can be described as health professionals, community oriented, where the knowledge seems to be spread across theoretical (slightly heavier) and applied and they tend to act as consultant. UK pharmacists have shown good applied knowledge and were place between the health professional and pharmacists involved in the dispensing procedures. They did not show a clear shift towards pro-active or re-active role, but they perceive themselves as health consultant too. Italian pharmacists appear to be the more shifted towards the theoretical knowledge but also to dispensing. They perceive themselves involved in patient management and so pro-active acting as health consultant. The health consultant shift is really evident in the Swiss pharmacists and less evident in the Italian and UK pharmacists respectively. This study was limited to a (relatively) small sample of pharmacists in only three countries. Further research needs to be conducted involving a larger number of pharmacists from more countries in order to fully understand and share best practice in order to achieve a patient centred pharmacy model.

References

1.Loddenkemper R, Gibson GJ, Sibille Y. (2003) European Respiratory Society/European Lung Foundation European Lung White Book. Sheffield, ERSJ., pp. 16-25.

2.National Co-ordinating Centre for NHS Service Delivery and Organisation R & D (NCCSDO) 2005. Concordance, adherence and compliance in medicine taking [Online] Available at http://www.medslearning.leeds.ac.uk/pages/documents/useful_docs/76-final-report%5B1%5D.pdf [Accessed 6th January 2010]

3.Yin, Y. and Yasuda, K. (2005) Similarity coefficient methods applied to the cell formation problem: a comparative investigation. Computers and Industrial Engineering 48: 471-489.

4.Mar Molinero C. and Xie A. (2007). What do UK employers want from OR/MS. Journal of the Operational Research Society, 58, 1543-1553.

5.Coxon, A.P.M. (1982) The Users Guide to multidimensional scaling, London: Heinemann.

6. Wiedenmayer, K; Summers, RS; Mackie, CA; Gous, AGS; Everard, M; Tromp, D (2006) Developing pharmacy practice: a focus on patient care. World Health Organisation, Geneva, Switzerland.

Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013)
International Conference, Mataró (Barcelona), Spain 25 - 28 June 2013

Optimal Chattering Regimes in Nonhomogeneous Bar Model

Larisa Manita

National Research University Higher School of Economics,
Moscow Institute of Electronics and Mathematics,
Bolshoy Trehsviatitelskiy Per. 3,
109028 Moscow, Russia.
(E-mail: lmanita@hse.ru)

Abstract. We consider an optimal control problem of longitudinal vibrations of a nonhomogeneous bar with clamped ends. We assume that the density of external forces is a control function. Using the Fourier method we prove that the optimal control is the chattering control, i.e., it has an infinite number of switchings in a finite time interval.
Keywords: Vibrations, Optimal control, Singular solutions, Nonhomogeneous bar.

1 Introduction

Consider small longitudinal vibrations of a nonhomogeneous bar of length l. The longitudinal displacement at a typical point x is denoted $y(t,x)$ where t is the time. Let $g(x,t)$ be a density of external longitudinal force at the instant of time t at the point x. Suppose that $g(t,x) = u(t)f(x)$ where the force profile function $f(x)$ is assumed to be given, $u\,(t)$ is the control function. Assume that

$$-1 \leq u\,(t) \leq 1 \tag{1}$$

The equation of longitudinal vibrations of the bar can be written as

$$p\,(x)\,y_{tt}(t,x) - \left(k\,(x)\,y_x(t,x)\right)_x = u(t)f(x) \tag{2}$$

Here $p\,(x) = \rho\,(x)\,S\,(x)\,,\ k\,(x) = E\,(x)\,S\,(x)\,,\ \rho\,(x)$ is the density of bar, $S\,(x)$ is the cross-sectional area, $E\,(x)$ is the Young's modulus at x, see, for example, [1–3]. We assume that the ends of the bar are clamped:

$$y|_{x=0} = y|_{x=l} = 0, \quad t > 0 \tag{3}$$

and the initial position and velocity are fixed:

$$y|_{t=0} = y_0(x), \quad y_t|_{t=0} = y_1(x), \qquad x \in [0,l] \tag{4}$$

We suppose that the coefficient functions k, p are smooth enough and

$$k\,(x) \geq k_0 > 0, \quad p\,(x) \geq p_0 > 0 \quad \forall x \in [0,l] \tag{5}$$

We consider an optimal control problem: to find such a control function $u\,(t)$ that minimize the following mean square functional

$$\int_0^\infty \int_0^l p\,(x)\,y^2\,(t,x)\,dxdt \to \inf \tag{6}$$

under (1)–(5).

The problems of longitudinal vibrations of a bar were considered in [1–3]. In [1,2] the dynamics of the longitudinal vibrations of a bar subjected to viscous boundary conditions was studied. The optimal boundary control problem for longitudinal vibrations of a bar was considered in [3]. By using a maximum principle the optimal control was expressed in terms of an adjoint variable.

In this paper for the problem of controlling the longitudinal vibrations of a bar (1)–(6) we construct a solution $y(t,x)$ in the form

$$y(t,x) = \sum_{j=1}^{\infty} s_j(t) h_j(x) \tag{7}$$

where $\{h_j(x)\}_{j=1}^{\infty}$ are eigenfunctions of the Sturm-Liouville problem, $\{s_j(t)\}_{j=1}^{\infty}$ are corresponding Fourier coefficients. To find Fourier coefficients we consider an optimal control problem in the space l^2. For the control problem in l^2 we show that the optimal solutions contain singular trajectories and chattering trajectories. A trajectory is called a chattering trajectory if it has an infinite number of a control switchings on a finite time interval. By similar method we studied the optimal control problem for a rotating uniform Timoshenko beam [4,5]. But for the Timoshenko beam similar results hold only for a dense set of initial conditions in the space l^2.

2 Optimal control problem in l_2

Define an operator L in $C^2([0,l])$ by $Lh = (kh_x)_x$. Consider the following Sturm-Liouville eigenvalue problem with Dirichlet boundary conditions:

$$Lh + \lambda p(x) h = 0, \quad x \in (0,l), \qquad h(0) = 0, \quad h(l) = 0 \tag{8}$$

Here the functions $k(x)$ and $p(x)$ satisfy (5). It is known (see [6]) that the problem (8) has an infinite sequence of eigenvalues $\{\lambda_j\}_{j=1}^{\infty}$, which are simple and positive: $0 < \lambda_1 < \lambda_2 < \ldots, \qquad \lambda_j \to \infty, \quad j \to \infty$. To each eigenvalue λ_j corresponds a single eigenfunction h_j, and the sequence of eigenfunctions $\{h_j(x)\}_{j=1}^{\infty}$ forms an ortonormal basis of $L_2((0,l);p)$ with the inner product $(z,w)_p = \int_0^l p(x) z(x) w(x) dx$. If $k(x), p(x)$ are smooth enough (for example, $k', p \in C^1([0,l])$) and the condition (5) holds, then the eigenvalues λ_j admit the asymptotic form [7–9]:

$$\frac{\lambda_j}{j^2} \sim \pi^2 \left(\int_0^l \sqrt{p(x)/k(x)} dx \right)^{-2}, \quad j \to \infty \tag{9}$$

Assume that $y(t,\cdot) \in L_2((0,l);p)$. For any $t > 0$ we expand the solution $y(t,x)$ of (2) in the basis $\{h_j(x)\}_{j=1}^{\infty}$:

$$y(t,x) = \sum_{j=1}^{\infty} s_j(t) h_j(x) \tag{10}$$

$$s_j(t) = \int_0^l p(x) y(t,x) h_j(x) dx = (y, h_j)_p$$

Using (4) we obtain $s_j(0) = (y_0, h_j)_p$ and $\dot{s}_j(0) = (y_1, h_j)_p$. Multiplying the equation (2) by h_j and integrating it in x we get that the function $s_j(t)$ satisfies the following equation:

$$\ddot{s}_j(t) + \lambda_j s_j(t) = c_j u(t), \quad j = 1, 2, \ldots$$

where

$$c_j = (f, h_j) = \int_0^l f(x) h_j(x) dx \tag{11}$$

Substituting (10) into (6) and using Parseval's equality we get:

$$\int_0^\infty \int_0^l p(x) y^2(t, x) dxdt = \int_0^\infty \sum_{j=1}^\infty s_j^2(t) dt \tag{12}$$

We reduce the control problem for partial differential equation (2)–(6) to a control problem for a countable system of ordinary differential equations:

$$\int_0^\infty \sum s_j^2(t) dt \to \inf \tag{13}$$

$$\ddot{s}_j(t) + \lambda_j s_j(t) = C_j u(t), \quad j = 1, 2, \ldots \tag{14}$$

$$s_j(0) = a_j, \quad \dot{s}_j(0) = b_j \tag{15}$$

$$-1 \le u(t) \le 1 \tag{16}$$

Here $a_j = (y_0, h_j)_p$, $\quad b_j = (y_1, h_j)_p$.

We shall assume everywhere below that

$$c_j \neq 0 \quad \text{for all } j = 1, 2, \ldots \tag{17}$$

Remark. Assumption (17) is very essential for the problem (13)–(16). Indeed, let $c_{j_0} = 0$ for some j_0. Then j_0-th equation in (14) takes the form

$$\ddot{s}_{j_0}(t) + \lambda_{j_0} s_{j_0}(t) = 0$$

Hence, if $|a_{j_0}| + |b_{j_0}| \neq 0$ then the corresponding solution $s_{j_0}(t)$ does not vanish as $t \to \infty$. Therefore the integral (13) is equal to $+\infty$ and the optimization problem (13)–(16) has not any sense.

Assume that

$$y_0, y_1 \in L_2((0, l); p), \quad f \in L_2(0, l) \tag{18}$$

Put $\omega_j = \sqrt{\lambda_j}$, $\quad \tau_j(t) = \dot{s}_j(t)$. Then the problem (13)–(16) takes the form

$$\int_0^\infty \sum s_j^2(t) dt \to \inf \tag{19}$$

$$\dot{s}_j = \tau_j, \quad \dot{\tau}_j = -\omega_j^2 s_j + c_j u \tag{20}$$

$$s_j(0) = a_j, \quad \tau_j(0) = b_j, \quad j = 1, 2, \ldots \tag{21}$$

$$-1 \leq u(t) \leq 1 \tag{22}$$

Denote

$$s(t) = (s_1(t), s_2(t), \ldots), \quad \tau(t) = (\tau_1(t), \tau_2(t), \ldots)$$

$$a = (a_1, a_2, \ldots), \quad b = (b_1, b_2, \ldots), \quad c = (c_1, c_2, \ldots)$$

Consider the standart Hilbert space l_2:

$$l_2 = \left\{ w = (w_1, w_2, \ldots) : \ w_n \in \mathbf{R}, \ \sum_{n=1}^{\infty} w_n^2 < \infty \right\}$$

with inner product $(v, w) = \sum_{n=1}^{\infty} v_n w_n$. Using assumption (18) we get that $a, \ b, \ c \in l_2$.

The existence and uniqueness of a solution $(s(t), \tau(t))$ to problem (19)–(22) in the space $l_2 \times l_2$ were proved in [5] for any initial data from an open neighborhood of the origin $(s = 0, \tau = 0)$. We apply a formal generalization of the Pontryagin maximum principle to the problem (19)–(22). Denote by $\psi_i = (\psi_{i1}, \psi_{i2}, \ldots)$ $(i = 1, 2)$ adjoint variables. Define the Pontryagin function

$$\begin{aligned} H(\psi_1, \psi_2, s, \tau, u) &= \sum_{j=1}^{\infty} \left(\psi_{1j}\tau_j - \psi_{2j}\omega_j^2 s_j + \psi_{2j} c_j u - s_j^2/2 \right) = \\ &= H_0(\psi_1, \psi_2, s, \tau) + u H_1(\psi_1, \psi_2, s, \tau) \end{aligned}$$

where

$$H_0(\psi_1, \psi_2, s, \tau) = \sum_{j=1}^{\infty} \left(\psi_{1j}\tau_j - \psi_{2j}\omega_j^2 s_j - \frac{1}{2} s_j^2 \right), \quad H_1(\psi_1, \psi_2, s, \tau) = \sum_{j=1}^{\infty} \psi_{2j} c_j$$

For breavity we denote $z = (\psi_1, \psi_2, s, \tau)$. In the space $l_2 \times l_2 \times l_2 \times l_2$ let us consider the Hamiltonian system

$$\begin{array}{ll} \dot{\psi}_{1j} = \psi_{2j}\omega_j^2 + s_j, & \dot{s}_j = \tau_j \\ \dot{\psi}_{2j} = -\psi_{1j}, & \dot{\tau}_j = -\omega_j^2 s_j + c_j u^*(t) \end{array} \qquad j = 1, 2, \ldots \tag{23}$$

where $u^*(t)$ satisfies the following maximum condition:

$$H(z(t), u^*(t)) = \max_{u \in [-1,1]} H(z(t), u) = \max_{u \in [-1,1]} (u H_1(z(t))) \tag{24}$$

It was proved [5] that the Pontryagin maximum principle is the necessary and sufficient condition of optimality for the problem (19)–(22).

If $H_1(z(t)) \neq 0$ along the trajectory the optimal control is uniquely determined as a function of time from the maximum condition (24):

$$u^*(t) = \operatorname{sign}(H_1(z(t))) = \operatorname{sign}\left(\sum_{j=1}^{\infty} \psi_{2j}(t) c_j \right)$$

Suppose that there exists an interval (t_1, t_2) such that

$$H_1(z(t)) \equiv 0, \quad \forall t \in (t_1, t_2)$$

To find an optimal control $u(t)$ in this case we will differentiate the identity $H_1(z(t)) \equiv 0$ by virtue of the Pontryagin maximum principle system (23) until the control u occurs in the resulting expression with a non-zero coefficient.

$$\begin{aligned}
&\frac{d}{dt}\bigg|_{(23)} H_1(z) = \frac{d}{dt}\bigg|_{(23)} \left(\sum_{j=1}^{\infty} \psi_{2j} c_j\right) = \left(-\sum_{j=1}^{\infty} c_j \psi_{1j}\right) \\
&\frac{d^2}{dt^2}\bigg|_{(23)} H_1(z) = \frac{d}{dt}\bigg|_{(23)} \left(-\sum_{j=1}^{\infty} c_j \psi_{1j}\right) = -\sum_{j=1}^{\infty} c_j \left(\psi_{2j}\omega_j^2 + s_j\right) \\
&\frac{d^3}{dt^3}\bigg|_{(23)} H_1(z) = -\frac{d}{dt}\bigg|_{(23)} \sum_{j=1}^{\infty} c_j \left(\psi_{2j}\omega_j^2 + s_j\right) = -\sum_{j=1}^{\infty} c_j \left(-\omega_j^2 \psi_{1j} + \tau_j\right) \\
&\frac{d^4}{dt^4}\bigg|_{(23)} H_1(z) = \sum_{j=1}^{\infty} c_j \omega_j^2 \left(\psi_{2j}\omega_j^2 + 2s_j\right) - u \sum_{j=1}^{\infty} c_j^2 \qquad (25)
\end{aligned}$$

In our case the control appears in the forth derivative. Assume that all series in (25) are convergent in l^2. Denote the subsequent derivatives by H_2, H_3, H_4:

$$H_2(z) = -\sum_{j=1}^{\infty} c_j \psi_{1j}, \quad H_3(z) = -\sum_{j=1}^{\infty} c_j \left(\psi_{2j}\omega_j^2 + s_j\right)$$

$$H_4(z) = -\sum_{j=1}^{\infty} c_j \left(-\psi_{1j}\omega_j^2 + \tau_j\right)$$

From (25) it follows that

$$H_1(z(t)) = H_2(z(t)) = H_3(z(t)) = H_4(z(t)) = 0, \quad t \in (t_1, t_2)$$

We say a solution of the (23)–(24) is singular if it belongs to the surface

$$\Sigma = \{ z : \ H_1(z) = H_2(z) = H_3(z) = H_4(z) = 0 \} \qquad (26)$$

A singular control $u^0(t)$ is determited from the equation $\frac{d^4}{dt^4}\Big|_{(23)} H_1(z) = 0$. Using (25) we obtain

$$u^0(t) = \sum_{j=1}^{\infty} c_j \omega_j^2 \left(\psi_{2j}\omega_j^2 + 2s_j\right) / \sum_{j=1}^{\infty} c_j^2 \qquad (27)$$

Note that the origin $(\psi_1 = 0, \ \psi_2 = 0, \ s = 0, \ \tau = 0)$ is the singular trajectory and the corresponding singular control $u^0(t)$ equals 0.

It was proved [5] that in a certain neighborhood of the origin the structure of the optimal solutions is the following one: the optimal nonsingular trajectories attain the singular surface Σ in finite time with countable many control switchings, after that the optimal trajectories remain on the singular surface. Namely, the following theorem holds.

Theorem 1. *[5]. Let* $\left(c_1\omega_1^3, c_2\omega_2^3, c_3\omega_3^3, \ldots\right) \in l_2$ *and* $c_j \neq 0 \quad \forall j$. *Assume that there exist positive constants* δ *and* K *such that*

$$|\omega_{j+1}| - |\omega_j| \geq \delta, \quad |\omega_j| \leq K \cdot j, \quad j = 1, 2, \ldots$$

Then there exists an open neighborhood of the origin in the space (s, τ) *such that the following statements hold for all initial data* (a, b) *from this neighborhood.*

(i) The problem (19)–(22) has a unique optimal solution.

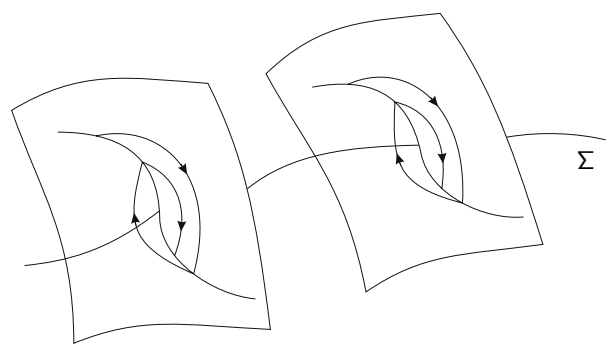
(ii) In the space $z = (\psi_1, \psi_2, s, \tau)$ *there exists the singular surface* Σ *of codimension 4 given by the equations*

$$\sum_{j=1}^{\infty} \psi_{2j} c_j = 0, \qquad \sum_{j=1}^{\infty} c_j \psi_{1j} = 0$$

$$\sum_{j=1}^{\infty} c_j \left(\psi_{2j}\omega_j^2 + s_j\right) = 0, \qquad \sum_{j=1}^{\infty} c_j \left(-\psi_{1j}\omega_j^2 + \tau_j\right) = 0$$

which is filled in by the singular extremals of the problem (19)–(22). The control on singular extremals are defined by (27).

(iii) For all initial data not belonging to the projection of the singular surface Σ *on the space* (s, τ), *the optimal trajectories arrive at* Σ *in finite time with countable many control switchings, i.e., the optimal trajectories are chattering trajectories.*



The space $l_2 \times l_2 \times l_2 \times l_2$. The singular surface Σ.

3 Optimal solution for controlling vibrations

Denote $Q_T = (0, l) \times (0, T)$, where $T > 0$. Consider the Sobolev space $H^k(Q_T) = W_2^k(Q_T)$, $k \geq 0$. The space $H^k(Q_T)$ consists of all functions $v \in L_2(Q_T)$ whose generalized derivatives up to order k exist and belong to $L_2(Q_T)$. The space $H_0^k(Q_T)$ can be defined as a completion of $C_0^\infty(Q_T)$ with respect to the norm of the space $H^k(Q_T)$.

Let $g = uf \in L_2(Q_T)$, $\quad y_1 \in L_2(0, l)$.

Definition 1. [10] We say a function $y \in H^1(Q_T)$ is a *weak solution* of the (2)–(4) if

(i) $y|_{x=0} = y|_{x=l} = 0, \quad t > 0, \qquad y|_{t=0} = y_0(x), \quad x \in [0, l];$

(ii)

$$\int_{Q_T} (k y_x v_x - p y_t v_t)\, dxdt = \int_{Q_T} gv\, dxdt + \int_0^l y_1(x) v(0,x)\, dx$$

for each $v \in H^1(Q_T)$: $v|_{x=0} = v|_{x=l} = 0$, $v|_{t=T} = 0$.

Definition 2. [10] We say $y \in H^2(Q_T)$ is a *almost everywhere solution* of the problem (2)–(4) provided y satisfies equation (2) in Q_T for almost all $(t, x) \in Q_T$ and y satisfies (3)–(4).

The following theorem is the main result for the problem (1)–(6).

Theorem 2. *Let $y_0 \in H_0^2(0, l)$, $y_1 \in H_0^1(0, l)$, $k(x), p(x)$ are smooth enough (for example, $k, p \in C^4([0, l])$), $p(x) \geq p_0 > 0$, $k(x) \geq k_0 > 0$. Assume that $f \in C^4[0, l]$,*

$$f(0) = f(l) = 0, \quad f^{(i)}(0) = f^{(i)}(l) = 0, \ i = 1, 2, 3 \tag{28}$$

and condition (17) holds. Then there exist positive constants q_1 and q_2 such that if

$$\|y_0\|_{L_2((0,l);p)} < q_1, \quad \|y_1\|_{L_2((0,l);p)} < q_2$$

then

(i) the problem (2)–(6) has a unique optimal solution $(y^*(t, x), u^*(t))$;
(ii) $y^* \in H^2(Q_T)$ for all $T > 0$;
(iii) the optimal control $u^*(t)$ has an infinite number of switchings in a finite time interval.

Proof. Here we use notations introduced in Section 2. Since the functions p, k, f satisfy conditions of Theorem 2 it follows [8] that $c_j \sim j^{-4}$ as $j \to \infty$, where

$$c_j = (f, h_j) = \int_0^l f(x)\, h_j(x)\, dx$$

Then we have

$$\sum_{j=1}^{\infty} \left(c_j \omega_j^3\right)^2 = \sum_{j=1}^{\infty} \left(c_j^2 \omega_j^6\right) = \sum_{j=1}^{\infty} c_j^2 \lambda_j^3 < \infty \quad \Longrightarrow \quad \left(c_1\omega_1^3, c_2\omega_2^3, c_3\omega_3^3, \ldots\right) \in l_2$$

The property (9) of the eigenvalues $\{\lambda_j\}_{j=1}^{\infty}$ of the problem (8) imply that there exist positive constants δ and B such that

$$|\omega_{j+1}| - |\omega_j| \geq \delta, \quad |\omega_j| \leq Bj$$

Now we may apply Theorem 1 to the problem (19)–(22). Then there exists an optimal solution $(s^*(t), u^*(t))$ of the problem (19)–(22) and the optimal control $u^*(t)$ has an infinite number of switchings in the finite time interval.

Consider

$$y^*(t,x) = \sum_{j=1}^{\infty} s_j^*(t)\, h_j(x) \tag{29}$$

where $\{h_j(x)\}_{j=1}^{\infty}$ are eigenfunctions of the Sturm-Liouville problem (8).

Series (29) formally satisfies equation (2), boundary conditions (3) and initial conditions (4). Since the functions f, y_0, y_1 satisfy the conditions of Theorem 2 it follows (see [10,11]) that the function $y^*(t,x)$ defined by (29) is a unique weak solution of the problem (2)–(4) and $y^* \in H^2(Q_T)$ for any $T > 0$. Hence [10] $y^*(t,x)$ is the solution almost everywhere of the problem (2)–(4). Thus the function $y^*(t,x)$ satisfies (2) for almost all $(t,x) \in (0,l) \times (0,+\infty)$, boundary and initial conditions (3)–(4).

Since the function $s^*(t) = (s_1^*(t), s_2^*(t), \ldots)$ minimizes the functional (13) and the identity (12) holds then the function $y^*(t,x)$ minimizes the functional (6). Thus $(y^*(t,x), u^*(t))$ is a solution of the problem (1)–(6).

4 Conclusion

We considered the optimal control problem of longitudinal vibrations of a nonhomogeneous bar with clamped ends. We proved that the optimal trajectories contain singular part and nonsingular one with accumulation of control swithings.

Acknowledgements

This research was supported by the Russian Foundation of Basic Research (projects No. 11-01-00986 and 13-08-00665).

References

1.V. Jovanovic, A Fourier series solution for the longitudinal vibrations of a bar with viscous boundary conditions at each end, Journal of Engineering Mathematics (2012).

2.F.E. Udwadia, On the longitudinal vibrations of a bar with viscous boundaries: Super-stability, super-instability, and loss of damping, International Journal of Engineering Science 50 (2012) 79-100.

3.I.S. Sadek, J.M. Sloss, S. Adali, J.C. Bruch, Optimal Boundary Control of the Longitudinal Vibrations of a Rod Using a Maximum Principle, Journal of Vibrations and control 3 (1997) 235-254.

4.M.I. Zelikin, L.A. Manita, Optimal control for a Timoshenko beam, C.R. Mécanique 334, Issue 5 (2006) 292-297

5.V. Borisov, M. Zelikin, L. Manita, Optimal Synthesis in an Infinite-Dimensional Space, Proceedings of the Steklov Institute of Mathematics 271 (2010) 34-52.

6.M.A. Al-Gwaiz, Sturm-Liouville Theory and its Applications, Springer — Verlag, London, 2008.
7.A. Zettl, Sturm-Liouville Problems, in D. Hinton, P.W. Schaefer (Eds), Spectral Theory and Computational Methods of Sturm-Liouville Problems, Lecture Notes in pure and applied mathematics, v.191, Marcel Dekker, Inc., N.Y., 1997.
8.I.G. Petrovsky, Lectures on Partial Differential Equations, Dover Publications Inc., N.Y., 1992.
9.L. Collatz, Eigenwertaufgaben mit technischen Anwendungen, Akademische Verlagsgesellschaft Geest & Portig, Leipzig, 1963
10.V.P. Mikhailov, Partial Differential Equations, Mir Publishers, Moscow, 1978.
11.O.A. Ladyzhenskaya, The Boundary Value Problems of Mathematical Physics, Springer — Verlag, N.Y. Inc., 1985

Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013)
International Conference, Mataró (Barcelona), Spain 25 - 28 June 2013

Intrinsic Space Scales for Multidimensional Stochastic Synchronization Models

Anatoly Manita

Department of Probability, Faculty of Mathematics and Mechanics,
Lomonosov Moscow State University, 119992, Moscow, Russia
(E-mail: manita@mech.math.msu.su)

Abstract. We consider the interacting particle system $x(t) = (x_1(t), \ldots, x_N(t))$, $t \in \mathbb{R}_+$, consisting of N identical Brownian particles with stochastic synchronization-like interaction. We provide the answer to the question: what is a typical distance between particles in the synchronized system (when $t = +\infty$) and in the system approaching to "stochastic synchrony" (for large finite t). It appears that in some sense a distance between any pair of particles x_i and x_j is of order N. We obtain some asymptotical results about exponential moments of the distances at time t and show that distributions of $(x_i - x_j)/N$ have exponential tails uniformly in N. For the synchronized N-particle system we prove that all $(x_i - x_j)/N$ have a symmetric bilateral exponential distribution with asymptotically constant parameter: $c_{0,N} \to c_0$ as $N \to \infty$.
Keywords: Stochastic synchronization, Brownian particle system, Markov process, renewal process.

1 Introduction

We consider here a stochastic process $x(t) = (x_1(t), \ldots, x_N(t)) \in \mathbb{R}^N$, $t \in \mathbb{R}_+$, previously studied in [8] and [9]. Components of this process evolve as independent Brownian motions with constant drift coefficient $\sigma > 0$ except random time moments τ_1, τ_2, ... of synchronization jumps when states of a randomly chosen pair of components become equal. We give the rigorous definition of the model in Sect. 2.

Mathematical study of stochastic systems with synchronization ([3–7]) is motivated by many applications in computer science (asynchronous parallel and distributed algorithms [1]), wireless network theory (clock synchronization problems [11]) etc.

In the papers [8] and [9] it was assumed that intervals between synchronizations $\tau_n - \tau_{n-1}$ are independent and have common distribution not depending on N. Due to the random nature of the dynamics the so called perfect synchronization ($x_1 = \cdots = x_N$) is not possible for the present system. But as it was proved in [8] the model reaches synchronization in a stochastic sense: in relative coordinates of a moving observer the system has limit in distribution when $t \to \infty$. Unfornunately, the limiting distribution cannot be found explicitely and the study of its properties for large N is a challenging problem. The papers [8] and [9] were devoted mainly to a prestationary evolution of $x(t)$, namely, it was proved that on different time scales ($t = t_N \to \infty$ as $N \to \infty$) the particle

system has completely different qualitative behaviour. It was shown that the system passes three different phases before it reaches the final synchronization.

A very important question is: what is a natural space scale for the system of N synchronized particles? In other words: what is a typical distance between particles in the synchronized system? Some information on the distribution of $x_i - x_j$ can be derived from [8] and [9]: assuming that $\mathsf{E}\,(x_i(0) - x_j(0))^2 < \infty$ for all i and j we conclude that this second moment remains finite for all $t > 0$, moreover, it is bounded in t and $\mathsf{E}\,(x_i(\infty) - x_j(\infty))^2$ is of order N^2. As it will be seen from Theorem 1 of the present paper a similar statement is not true for all exponential moments. We prove (Theorem 1) that $\mathsf{E}\exp\left(c\,|x_i(t) - x_j(t)|\,/N\right)$ is bounded in t only for sufficiently small $c > 0$. Fortunatelly, this result gives us possibility to get sharp probabilistic estimates for distribution of the rescaled distance $|x_i(t) - x_j(t)|\,/N$ for large t. Moreover, for the synchronized system we show that distribution of $|x_i - x_j|\,/\sqrt{(N-1)N}$ does not depend on N and find this distribution in explicit form (Theorem 2). So we come to the conclusion that the natural space scale for our model is of order N.

This work is supported by the Russian Foundation of Basic Research (grant 12-01-00897).

2 Model

We remind here the definition of the model $x(t) = (x_1(t), \ldots, x_N(t)) \in \mathbb{R}^N$ called Brownian particles with synchronization ([8,9]). There is a sequence

$$0 = \tau_0 < \tau_1 < \cdots < \tau_n < \cdots$$

of random variables. On intervals (τ_{n-1}, τ_n) the particles $x_1, \ldots, x_N$ move according to the free dynamics: $dx_i(t) = \sigma dB_t^i$, $i = 1, \ldots, N$, where $\boldsymbol{B}_t = \left(B_t^1, \ldots, B_t^N\right)$, $t \geq 0$, is a N-dimensional Brownian motion independent of $\{\tau_n\}_{n=1}^{\infty}$, the diffusion coefficient $\sigma > 0$ is assumed to be constant. At epoch τ_n a pair of particles (i_n, j_n) is chosen at random[1] and the particle j_n jumps to the particle i_n:

$$x_{j_n}(\tau_n + 0) = x_{i_n}(\tau_n), \tag{1}$$

$$x_k(\tau_n + 0) = x_k(\tau_n), \quad k \in \{1, \ldots, N\} \setminus \{j_n\}\,. \tag{2}$$

In the framework of particle systems such perturbation of the free dynamics can be regarded as a synchronization-like interaction between particles. It is convenient to consider the jump (1)–(2) as a map $S_{(i,j)} : \mathbb{R}^N \to \mathbb{R}^N$:

$$S_{(i,j)} : \quad (x_1, \ldots, x_i, \ldots, x_j, \ldots, x_N) \mapsto (x_1, \ldots, x_i, \ldots, \underset{j}{x_i}, \ldots, x_N) \tag{3}$$

We need some assumptions on the sequence of $\{\tau_n\}_{n=1}^{\infty}$ concerning intervals between synchronization jumps $\Delta_n = \tau_n - \tau_{n-1}$, $n = 1, \ldots$.

[1] I.e., any ordered pair of particles (i_n, j_n) is chosen with probability $\dfrac{1}{(N-1)N}$.

A1. $\Delta_1, \ldots, \Delta_n, \ldots$ are independent.
A2. $\Delta_1, \ldots, \Delta_n, \ldots$ are identically distributed with some continuous probability distribution function F.
A3. Δ_n has exponential distribution with mean δ^{-1}: $\mathsf{P}\{\Delta_n > s\} = \exp(-\delta s)$, $s > 0$.

Under Assumption A1–A3 the stochastic process $(x(t),\, t \geq 0)$ is Markovian. Such model was considered in [8]. In [9] a more general particle system was studied under assumptions A1–A2.

For any random vector $z = (z_1, \ldots, z_N)$ with values in $\mathbb{R}^N$ we denote by $\mathcal{P}_z$ the distribution law of z. Hence $\mathcal{P}_z$ is some probability measure on $\mathbb{R}^N$.

We need also some *assumptions* on the *initial* particle configuration $x(0)$ which is assumed to be random.

AI. The initial particle configuration $x(0)$ is independent of $\{\tau_n\}_{n=1}^{\infty}$ and $(\boldsymbol{B}_t,\, t \geq 0)$.

Assumption AI is supposed to be held throughout all this paper. Sometimes we will make more restrictive assumptions AIS and AI0.

AIS. Assume that Assumption AI holds and that the initial distribution $\mathcal{P}_{x(0)}$ is invariant with respect to permutations of indices, i.e.,

$$\mathcal{P}_{\pi \star x(0)} = \mathcal{P}_{x(0)}$$

where $\pi : (1, \ldots, N) \to (i_1, \ldots, i_N)$ is an arbitrary permutation and

$$\pi \star (z_1, \ldots, z_N) = (z_{i_1}, \ldots, z_{i_N})\,.$$

AI0. $x_i(0) = 0$ for all $i = 1, \ldots, N$.

Evidently, AI0 is a subcase of AIS.

3 Long-time synchronization

It is intuitively clear from definition of the model that collective behavior of a particle system with synchronization is a superposition of two opposite tendencies: with the course of time the free dynamics increases the "spread" of the particle system while the synchronizing interaction tries to decrease it. The *first question* is about a long-time behavior of the N-particle system $x(t) = (x_1(t), \ldots, x_N(t))$. We fix N and let $t \to \infty$.

It can be shown [8] that the distribution $\mathcal{P}_{x(t)}$ <u>has no limit</u> in law as $t \to \infty$. Let us consider an improved process $x^\circ(t) \in \mathbb{R}^N$

$$x_i^\circ(t) = x_i(t) - M(x(t)), \qquad M(x) := \frac{1}{N}\sum_{m=1}^{N} x_m\,,$$

One can say that $x^\circ(t) = (x_1^\circ(t), \ldots, x_N^\circ(t))$ are coordinates of particles viewed by an moving observer placed in the center of mass $M(x)$. Note that

$\sum_{i=1}^N x_i^\circ(t) \equiv 0$ for any $t \geq 0$. It appears that $x^\circ(t)$ is an ergodic Markov process (see [8]). In particular,

$$\mathcal{P}_{x^\circ(t)} \to \mu^N \qquad (t \to \infty)$$

where μ^N is some probability measure on $\mathbb{R}^N$ which is the unique stationary distribution of the Markov process $(x^\circ(t),\ t \geq 0)$. This fact holds under the general assumption AI. Of course, the limiting measure μ^N depends on σ and δ.

Hence the particle system of Section 2 reaches synchronization in a stochastic sense. Properties of the synchronized system can be obtained by studying the probability measure μ^N. This measure is supported on the $(N-1)$-dimensional subspace

$$\left\{x^\circ = (x_1^\circ, \ldots, x_N^\circ) \,:\, \textstyle\sum_{i=1}^N x_i^\circ = 0\right\} \subset \mathbb{R}^N$$

and seems to be very complicated, especially for large N. So we come to the *next* very natural *question*: what is a typical "size" of the particle configuration $(x_1, \ldots, x_N)$ under the distribution μ^N ?

To formalize a notion of "size" or "spread" of the configuration $x = (x_1, \ldots, x_N)$ we define some functions on the configuration space $\mathbb{R}^N$. Consider functions

$$d_{ij}(x) = x_i - x_j, \quad d_{ij} : \mathbb{R}^N \to \mathbb{R}\,.$$

Clearly, $|d_{ij}(x)|$ is a distance between particles i and j in the configuration $x \in \mathbb{R}^N$. Note that $d_{ij}(x) = d_{ij}(x^\circ) = x_i^\circ - x_j^\circ$. Define also functions $D : \mathbb{R}^N \to \mathbb{R}$ and $\mathbf{D}_N : \mathbb{R}_+ \to \mathbb{R}_+$, as follows

$$D(x) := \frac{1}{C_N^2} \sum_{i<j} (x_i - x_j)^2 \;=\; \frac{2}{N(N-1)} \sum_{i<j} d_{ij}^2(x)\,,$$

$$\mathbf{D}_N(t) = \mathsf{E}\; D(x(t)), \qquad t \in \mathbb{R}_+\,.$$

Evidently, $\sqrt{D(x)}$ and $\sqrt{\mathbf{D}_N(t)}$ characterize desynchronization of the total particle system. It was proved in [8] and [9] that for any sequence $\{t_N\}$ such that $t_N/N^2 \to \infty$ as $N \to \infty$ we have

$$\mathbf{D}_N(t_N) \sim \delta^{-1}\sigma^2 N^2\,. \tag{4}$$

Dynamics of the Markov process $x(t)$ is symmetric with respect to different particles. Therefore the unique stationary distribution μ_N is invariant with respect to permutation of indices. Denoting by E_N the expectation taken with respect to the probability distribution μ^N one can derive from (4) that

$$\mathsf{E}_N\, d_{ij}^2(x^\circ(\infty)) \sim \text{const} \cdot N^2\,.$$

Hence we come to the next suggestion: typical distances between particles in the synchronized system are of order N.

Main goal of the present paper is to give more detailed information on the distribution of "rescaled distances"

$$N^{-1} d_{ij}(x(t)) = \frac{x_i(t) - x_j(t)}{N}$$

for large t (Theorem 1) or even for $t = +\infty$ (Theorem 2).

4 Main Results

4.1 Finite-time analysis

Define the function $G_{N,c}: \mathbb{R}^N \to \mathbb{R}$,

$$G_{N,c}(x) := \frac{1}{(N-1)N} \sum_{j_1<j_2} \exp\left(\frac{c}{N}\,|x_{j_2} - x_{j_1}|\right), \qquad c > 0.$$

It is useful to keep in mind physical dimensions of all parameters of the model. Assume that x_i is measured in meters (m). Then dimension of σ is $m/sec^{1/2}$, dimension of δ is sec^{-1}, values of the function $G_{N,c}$ are dimensionless, dimension of c is m^{-1}.

Theorem 1 *Let assumptions A1–A3 and AI0 hold.*

Then there exists $c_0 = c_0(\sigma, \delta) > 0$ such that the following properties hold:

i) *for $c < c_0$*

$$\limsup_{t\to+\infty} \mathsf{E}\, G_{N,c}(x(t)) \le B(c) < +\infty \tag{5}$$

where $B(c) > 0$ does not depend on N,

ii) *for $c = c_0$*

$$N - \frac{1}{2} \le \liminf_{t\to+\infty} \mathsf{E}\, G_{N,c_0}(x(t)) \le \limsup_{t\to+\infty} \mathsf{E}\, G_{N,c_0}(x(t)) < N,$$

iii) *for any $c > c_0$ there is $N_0 = N_0(c)$ such that*

$$\lim_{t\to+\infty} \mathsf{E}\, G_{N,c}(x(t)) = +\infty$$

for all $N \ge N_0$.

We assume here that all $x_i(0) = 0$ only for brevity of explanations. As it will be seen from the proof (Section 5) Assumption AI0 is not necessary and can be replaced by Assumption AI.

Let us discuss some useful corollaries of Theorem 1. Note that if the model satisfies to Assumption AIS then for any $t > 0$ the distribution $\mathcal{P}_{x(t)}$ is invariant with respect to the action of the permutation group. Hence all $d_{ij}(x(t)) = x_i(t) - x_j(t)$ are equally distributed and

$$\mathsf{E}\, G_{N,c}(x(t)) = \frac{1}{2}\, \mathsf{E}\, \exp\left(\frac{c}{N}\,|x_1(t) - x_2(t)|\right).$$

It follows from (5) that distance between particles $|x_i(t) - x_j(t)|$ has exponential moments

$$\mathsf{E}\, \exp\left(\frac{c}{N}\,|x_1(t) - x_2(t)|\right) < +\infty \quad \text{for} \quad 0 < c < c_0\,.$$

By the Chebyshev's inequality this gives an exponential bound for the distance between particles. Namely, *under Assumption AI0* it follows from Section 5 that for any $0 < c < c_0$

$$\mathsf{P}\left(|x_1(t) - x_2(t)| > rN\right) \le 2B(c) \exp\left(-cr\right), \qquad r > 0$$

uniformly in $t \ge 0$. The meaning of this result: a typical distance between particles in the particle system is of order N. Under Assumption AIS a similar results holds for sufficiently large $t \ge t_0(c, N)$.

4.2 Distances between particles of the synchronized system

Let us rescale the functions d_{ij} as follows

$$d_{ij}^{(N)}(x) = (x_i - x_j)/\sqrt{(N-1)\,N}\,.$$

It seems very surprising that for the synchronised system of Brownian particles distribution of the functions $d_{ij}^{(N)}(x)$ can be given in explicit form.

Theorem 2 *Let assumptions A1–A3 and AI hold. Consider the particle system $x = (x_1, \ldots, x_N)$ of Section 2 being in its stationary state μ^N. All random variables $d_{ij}^{(N)}(x)$, $(i, j = 1, \ldots, N)$ have a common distribution not depending on N which appears to be the bilateral exponential distribution with density*

$$p(y) = \frac{1}{2} c_0 \exp\left(-c_0\,|y|\right), \quad y \in \mathbb{R}^1,$$

where $c_0 = \sigma^{-1}\sqrt{2\delta}$.

In particular, in the synchronised system

$$\mathsf{E}_N\ |x_i - x_j| = \frac{\sigma}{\sqrt{2\delta}}\sqrt{(N-1)\,N} \sim \frac{\sigma}{\sqrt{2\delta}}\ N\,.$$

5 Proof of Theorem 1

Consider the function

$$V_{c_N}(x) = \frac{1}{(N-1)N} \sum_{j_1 < j_2} \cosh\left(c_N\,|x_{j_2} - x_{j_1}|\right).$$

Note that

$$\frac{e^{|x|}}{2} < \cosh x = \frac{e^x + e^{-x}}{2} \le \frac{e^{|x|} + 1}{2}\,.$$

So it is sufficient to prove main result for the function V_{c_N} with $c_N = c/N$:

$$2V_{c/N}(x) - \frac{1}{2} \le G_{N,c}(x) < 2V_{c/N}(x). \tag{6}$$

From this point and till the end of the below Lemma 3 we fix some *arbitrary* sequence $\{c_N\}$. To have shorter notation we write $V(x)$ instead of $V_{c_N}(x)$. We introduce also an auxiliary function

$$V_0(x) = \frac{1}{(N-1)N} \sum_{j_1 < j_2} g_N\left(x_{j_1}, x_{j_2}\right), \qquad g_N(a,b) = \cosh\left(c_N(a-b)\right) - 1\,.$$

It is easy to check that $V(x) = V_0(x) + \frac{1}{2}$. Note that

$$g_N(a,b) = g_N(b,a) \quad \text{and} \quad g_N(a,a) = 0\,. \tag{7}$$

Moreover, $g_N(a,b) > 0$ for $a \neq b$, hence $V_0(x) \geq 0$ and

$$V_0(x) = 0 \quad \Leftrightarrow \quad x_1 = \cdots = x_N \,.$$

Let $f = f(x)$ be some function on the configuration space $\mathbb{R}^N$. Denote

$$f^{(n)} = \mathsf{E}\left(f(x(\tau_n)) \mid \{\tau_j\}_{j=1}^{\infty}\right), \qquad n = 1, 2, \ldots \,. \tag{8}$$

Hence $f^{(n)}$ is a random variable functionally depending on the sequence $\{\tau_j\}_{j=1}^{\infty}$. Consider now $\left\{V^{(n)}\right\}$ and $\left\{V_0^{(n)}\right\}$. Evidently, $V^{(n)} = V_0^{(n)} + \frac{1}{2}$.

Define a map-valued random variable $\mathcal{S}$ such that

$$\mathsf{P}\left\{\mathcal{S} = S_{(i,j)}\right\} = \frac{1}{(N-1)N}, \qquad i \neq j.$$

To continue the proof we need next two lemmas.

Lemma 1 *There exists $\varkappa > 0$ such that for any $x \in \mathbb{R}^N$*

$$\mathsf{E}\, V_0(\mathcal{S}x) = k_N V_0(x),$$

where $k_N = 1 - \varkappa / \left((N-1)N\right)$.

Lemma 2 *For $s > 0$, $x \in \mathbb{R}^N$*

$$\mathsf{E}\, V(x + \sigma \boldsymbol{B}_s) = V(x) e^{2qs} \tag{9}$$

where $q = (c_N \sigma)^2 / 2$.

Proof of Lemma 1 can be obtained using properties (7) and ideas of Lemma 2 in [7]. The constant $\varkappa$ is dimensionless. Moreover, it follows from [7] that$\varkappa = 2$ for the case of pair interaction (1)–(2). We omit details. The proof of Lemma 2 is a straigtforward calculation based on the identity

$$\mathsf{E}\, \exp\left(bB_s^i\right) = \exp\left(b^2 s/2\right).$$

It follows from Lemma 1 that

$$\mathsf{E}\, V(\mathcal{S}x) = k_N V(x) + \frac{1 - k_N}{2}\,. \tag{10}$$

Note that the stochastic process $x(t)$ has left continuous paths. Since on the intervals $[\tau_{n-1}, \tau_n)$ the process has only one synchronization jump at time τ_n, combining (9)–(10) with Assumption A1 we get

$$V^{(n)} = k_N V^{(n-1)} e^{2q\Delta_n} + l_N$$

where $l_N = \frac{1}{2}(1 - k_N)$. Hence $V^{(n)} = k_N^2 V^{(n-2)} e^{2q(\Delta_{n-1} + \Delta_n)} + k_N e^{2q\Delta_n} l_N + l_N$. Finally, we come to the following representation

$$\begin{aligned} V^{(n)} = {} & k_N^n V^{(0)} e^{2q(\Delta_1 + \cdots + \Delta_n)} + k_N^{n-1} e^{2q(\Delta_2 + \cdots + \Delta_n)} l_N + \cdots + \\ & + k_N e^{2q\Delta_n} l_N + l_N\,. \end{aligned}$$

Fix some $t > 0$. The number of synchronization jumps on the time interval $[0, t]$ is random. To proceed with the proof we recall some facts and notation from renewal theory [2]. Let $\Pi_t = \max\{m : \tau_m \le t\}$ and hence $\tau_{\Pi_t} = \max\{\tau_i : \tau_i \le t\}$. From Assumption A2 it follows that (with probability 1) on the open time interval (τ_{Π_t}, t) there is no synchronization jump. Hence

$$\begin{aligned}\mathsf{E}\left(V(x(t) \mid \{\tau_j\}_{j=1}^{\infty}\right) &= V^{(\Pi_t)} e^{2q(t-\tau_{\Pi_t})} = \\ &= k_N^{\Pi_t} V^{(0)} e^{2qt} + k_N^{\Pi_t - 1} e^{2q(t-\tau_1)} l_N + \cdots \\ &\quad + k_N e^{2q(t-\tau_{\Pi_t - 1})} l_N + e^{2q(t-\tau_{\Pi_t})} l_N \,.\end{aligned}$$

Now we are going to consider a conditional expectation $\mathsf{E}\,(\cdot \mid \Pi_t)$ of this expression:

$$\mathsf{E}\,(V(x(t) \mid \Pi_t = n) = k_N^n V^{(0)} e^{2qt} + l_N \sum_{i=1}^{n} k_N^{n-i} \mathsf{E}\left(e^{2q(t-\tau_i)} \mid \Pi_t = n\right) . \quad (11)$$

Under Assumption A3 (Π_t, $t \ge 0$) is a Poisson process of intensity δ. It is easy to check (see, for example, the technical lemma 4.2 on page 62 of [10]) that conditional distributions of $t - \tau_m$ belong to the class of Beta-distributions:

$$\begin{aligned}\mathsf{P}\,\{t - \tau_m \in (y, y + dy) \mid \Pi_t = n\} = n\, C_{n-1}^{m-1} \frac{y^{n-m}(t-y)^{m-1}}{t^n}\, dy, \\ y \in (0, t),\ m = \overline{1, n}.\end{aligned}$$

Considering the second summand in (11) and applying the Binon formula

$$\sum_{i=1}^{n} k_N^{n-i} \int_0^t e^{2qy}\, n\, C_{n-1}^{i-1} \frac{y^{n-i}(t-y)^{i-1}}{t^n}\, dy = t^{-1} \int_0^t e^{2qy}\, n \left(\frac{k_N y + (t-y)}{t}\right)^{n-1} dx$$

we get

$$\mathsf{E}\,(V(x(t) \mid \Pi_t = n) = k_N^n V^{(0)} e^{2qt} + l_N t^{-1} \int_0^t e^{2qy}\, n \left(\frac{k_N x + (t-x)}{t}\right)^{n-1} dy\,.$$

Averaging in n and using notation $\varphi(y) = \mathsf{E}\, y^{\Pi_t}$ for the generating function of Π_t we come to the following representation.

$$\mathsf{E}\, V(x(t) = \varphi(k_N)\, V^{(0)} e^{2qt} + l_N\, t^{-1} \int_0^t e^{2qy}\, \varphi'\left(1 - \frac{(1-k_N)\, y}{t}\right) dy$$

Recall that under Assumption A3 the random variable Π_t is Poissonian with the mean δt. Since $\varphi(y) = \exp\,(\,\delta t\,(y-1)\,)$ and $\varphi'(y) = \delta t\, \exp\,(\,\delta t\,(y-1)\,)$, we

have

$$t^{-1}\int_0^t e^{2qy}\,\varphi'\left(1-\tfrac{(1-k_N)\,y}{t}\right)\,dy \;=t^{-1}\int_0^t e^{2qy}\,\delta t\,\exp\left(-\delta t\,\cdot\,\tfrac{(1-k_N)\,y}{t}\right)\,dy\;=$$

$$=\delta\int_0^t \exp\left(2qy-\delta\,(1-k_N)\,y\,\right)\,dy\;=$$

$$=\begin{cases}\dfrac{\delta}{a_N}\,\left(1-e^{-a_N t}\right)\,, & \text{if } a_N\neq 0,\\[2ex] \delta t, & \text{if } a_N=0,\end{cases}$$

where $a_N=\frac{\varkappa\delta}{N(N-1)}-2q$. Moreover,

$$\varphi(k_N)\,\exp(2qt)=\exp(\,2qt-\delta t\,(1-k_N)\,)=e^{-a_N t}\,.$$

So we have got explicit formulae for $\mathsf{E}\,V(x(t))$. Let us summarize this result in the next lemma.

Lemma 3 *In the case $a_N\neq 0$*

$$\mathsf{E}\,V(x(t))=e^{-a_N t}\,\mathsf{E}\,V(x(0))\,+\,\frac{l_N\delta}{a_N}\,\left(1-e^{-a_N t}\right) \tag{12}$$

and in the case $a_N=0$

$$\begin{aligned}\mathsf{E}\,V(x(t))&=\mathsf{E}\,V(x(0))\,+l_N\delta t\;=\\ &=\mathsf{E}\,V(x(0))\,+(c_N\sigma)^2\,t\end{aligned} \tag{13}$$

where

$$l_N=\frac{\varkappa}{2N(N-1)}\,,\qquad a_N=\frac{\varkappa\delta}{N(N-1)}-(c_N\sigma)^2\;.$$

In other words, the function $f(t)=\mathsf{E}\,V(x(t)$ solves the following equation

$$f'=-a_N f+l_N\delta\,.$$

With this lemma in mind we can finish the proof of Theorem 1. Indeed, we see from (12) and (13) that $\mathsf{E}\,V(x(t))$ is bounded in t if and only if $a_N>0$. Putting $c_N=c/N$ we conclude

$$a_N>0\quad\Leftrightarrow\quad c^2<\frac{\varkappa\delta}{N(N-1)}\cdot\frac{N^2}{\sigma^2}\quad\Leftrightarrow\quad c<\frac{\sqrt{\varkappa\delta}}{\sigma}\cdot\sqrt{\frac{N}{N-1}}\,. \tag{14}$$

From this observation the statement of Theorem 1 follows with $c_0=\sigma^{-1}\sqrt{\varkappa\delta}$. Indeed, the assumption on initial configuration ($x_i(0)=0$ for all $i=1,\dots,N$) implies that $V(x(0))=\frac{1}{2}$. If $c\le c_0$ then $a_N>0$. Hence $\mathsf{E}\,V(x(t))\uparrow w_N(c)$ as $t\to+\infty$ where

$$w_N(c)=\frac{l_N\delta}{a_N}=\frac{1/2}{1-\dfrac{c^2\sigma^2}{\varkappa\delta}\cdot\dfrac{N-1}{N}}\,.$$

Taking into account (6) in the case $c < c_0$ we put

$$B(c) = \frac{1}{1-(c/c_0)^2}$$

and get the item i) of the theorem. In the case $c = c_0$ we have $w_N(c_0) = N/2$. Now the item ii) of Theorem 1 immediately follows from (6).

Assume that $c > c_0$. It is easy to see from (14) that there exists N_0 such that $a_N < 0$ for all $N \geq N_0$. For any such N the r.h.s. of (12) tends to infinity as $t \to +\infty$.

The proof of Theorem 1 is over.

Evidently, this proof is valid also for the general initial assumption AI. The only difference is that for $c \leq c_0$ we still have convergence $\mathsf{E}\, V(x(t)) \to w_N(c)$ as $t \to +\infty$ but we cannot ensure that this convergence is monotonic.

6 Laplace Transform: Proof of Theorem 2

We proved in Section 5 that if $c \leq c_0$ then

$$\mathsf{E}\, V_{c/N}(x(t)) \;\to\; w_N(c) \qquad (t \to \infty),$$

where $c_0 = \sigma^{-1}\sqrt{\varkappa\delta}$ and

$$w_N(c) = \frac{1/2}{1-\frac{c^2}{c_0^2}\,\frac{N-1}{N}}\ .$$

Let ξ be some random variable. Denote by $\mathcal{L}(\xi; s)$ is the Laplace transform of distribution of ξ:

$$\mathcal{L}(\xi; s) := \mathsf{E}\, \exp\left(-s\xi\right).$$

If the distribution of ξ is symmetric, i.e. $\mathcal{P}_\xi = \mathcal{P}_{-\xi}$, then $\mathcal{L}(\xi; s) = \mathsf{E}\, \cosh\left(s\xi\right)$.

It follows from Section 3 the limiting stationary distribution μ^N is invariant with respect to permutations of indices. Hence under the distribution μ^N all random variables

$$d_{ij}(x) = x_i - x_j = x_i^\circ - x_j^\circ$$

are equally distributed and symmetric. Hence

$$\mathsf{E}_N V_{c/N}(x^\circ) = \frac{1}{2}\,\mathsf{E}_N \cosh(c d_{12}(x^\circ)/N) = \frac{1}{2}\,\mathcal{L}(d_{12}(x^\circ)/N\,; c).$$

From all these arguments and the following observation

$$\mathsf{E}_N V_{c/N}(x^\circ) = w_N(c) = \frac{1/2}{1-\frac{c^2}{c_0^2}\,\frac{N-1}{N}}\ .$$

we conclude that the rescaled differences $(x_i - x_j)/N$ have the Laplace transform

$$\mathcal{L}(c) = \frac{1}{1-\frac{c^2}{c_0^2}\,\frac{N-1}{N}} \tag{15}$$

and hence they have the common bilateral exponential distribution with probability density $p(y) = \frac{1}{2} c_{0,N} \exp\left(-c_{0,N}\,|y|\right)$, $y \in \mathbb{R}$, where

$$c_{0,N} = \frac{c_0}{1 - \frac{1}{N}}\,. \tag{16}$$

The presence of N in the formulae (15) and (16) can be eliminated if we improve the scale factor as follows:

$$d_{ij}^{(N)}(x) = (x_i - x_j)/\sqrt{(N-1)\,N}\,.$$

For this case *in the synchronized system* any $d_{ij}^{(N)}$ has the bilateral exponential distribution with parameter $c_0 = \sigma^{-1}\sqrt{\varkappa\delta}$:

$$\mathcal{L}(c) = \frac{1}{1 - \frac{c^2}{c_0^2}}\,, \qquad p(y) = \frac{1}{2} c_0 \exp\left(-c_0\,|y|\right).$$

In particular, $\mathsf{E}_N \left|d_{ij}^{(N)}\right| = c_0^{-1} = \dfrac{\sigma}{\sqrt{\varkappa\delta}}\,, \qquad \mathsf{Var}_N\left(d_{ij}^{(N)}\right) = 2c_0^{-2} = \dfrac{2\sigma^2}{\varkappa\delta}$ and

$$\mathsf{E}_N\,|x_i - x_j| = \frac{\sigma}{\sqrt{\varkappa\delta}}\sqrt{(N-1)\,N} \sim \frac{\sigma}{\sqrt{\varkappa\delta}}\,N\,.$$

We recall remark made after Lemma 1 that $\varkappa = 2$ in the case of pair interaction (see [7]).

It should be noted that the nonstationary "finite-time" distributions $\mathcal{P}_{d_{12}(x(t))}$, $t > 0$, do not have such simple and explicit form as $\mathcal{P}_{d_{12}(x)}$ in the synchronized system.

7 Concluding remarks

The main result of this paper remains true in more wide situations than assumptions of Sect. 2 and Sect. 4. Instead of pair interaction (1)–(2) we can consider the class of interactions called k-particle symmetrical synchronizations [7,8]. The only difference is in a value of the constant $\varkappa$. Evidently, the statement of Theorem 1 holds for any initial distributions of $x(0)$ because Lemma 3 is valid in the general situation.

Similar result is expected also for the model studied in [7] where a free dynamics is governed by independent random walks.

We hope that Assumption 3 is not essential and an analog of Theorem 1 can be proved also for non-Markovian case under Assumptions A1–A2. The method of proof would be essentially the same but now one should pay additional attention to a general renewal process related with the sequence $\{\tau_n\}$. To show some difficulties which are present in the non-Markovian case we refer to [9].

Anouther important axis of the future research is the joint distribution of $(d_{ij} = x_i - x_j,\ i \neq j)$ which seems to be very complicated due to strong correlations in the synchronized particle system. In the present paper we discussed only one dimensional marginal distributions.

References

1.D.P. Bertsekas and J.N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*, Belmont, Athena Scientific, 1997.

2.D.R. Cox. *Renewal theory*, London, Methuen, 1962.

3.V. Malyshev, A. Manita. Phase transitions in the time synchronization model. *Theory of Probability and its Applications* 50:134–141, 2006.

4.A.G. Malyshkin, Limit dynamics for stochastic models of data exchange in parallel computation networks. *Problems of Information Transmission* 42:234–250. (2006)

5.A. Manita, V. Shcherbakov, Asymptotic analysis of a particle system with mean-field interaction, *Markov Processes Relat. Fields* 11:489–518 (2005)

6.A. Manita, Markov processes in the continuous model of stochastic synchronization. *Russ. Math. Surv.* 61:993–995. (2006)

7.A.D. Manita. Stochastic synchronization in a large system of identical particles. *Theory of Probability and its Applications* 53, 155–161, 2009. (See also http://arxiv.org/abs/math.PR/0606040)

8.A. Manita, Brownian particles interacting via synchronizations. *Communications in Statistics — Theory and Methods*, 2011, v. 40, N. 19-20, p. 3440–3451. (see also http://arxiv.org/abs/1012.3140)

9.A. Manita, On Markovian and non-Markovian models of stochastic synchronization. *Proceedings of The 14th Conference "Applied Stochastic Models and Data Analysis" (ASMDA)*, 2011, Rome, Italy, p. 886–893.
(See also http://goo.gl/8P4ma)

10.A.D.Manita, On phases in evolution of multi-dimensional interacting diffusions with synchronization. In *Contemporary problems of mathematics and mechanics.* Vol. 7. Mathematics. 2011. P. 50-67 (In Russian).
http://mech.math.msu.su/probab/cheb190.pdf#page=50

11.O. Simeone, U. Spagnolini, Y. Bar-Ness, S. Strogatz. Distributed synchronization in wireless networks. *IEEE Signal Processing Magazine*, 2008, V. 25, N. 5, pp. 81–97.

Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013)
International Conference, Mataró (Barcelona), Spain 25 - 28 June 2013

Optimal Scaling with missing categorical data: An application to the study of labour markets in 19th century Spain

Cecilio Mar Molinero[1], and Fabiola Portillo[2]

[1] University of Kent, Canterbury, CT2 7PE, UK, and Universitat Autonoma de Barcelona, Barcelona, Spain (E-mail: C.Mar-Molinero@kent.ac.uk)

[2] Department of Economics and Business, University of La Rioja, C/ Ciguena 60, 26005 Logrono, La Rioja, Spain (E-mail: fabiola.portillo@unirioja.es)

Abstract. The information contained in a database from the Spanish railway company MZA for the period 1882-1889 is explored by means of Optimal Scaling methods. The sample includes 992 employees who joined the Madrid-Atocha workshop during that period and contains missing qualitative values. A technique that combines correspondence analysis with the k-means clustering algorithm is implemented to impute these values, maximizing internal consistency as measured by Guttman's squared correlation ratio. The results show two characteristics observed in other studies of labour relations: the existence of «ports of entry» for workers at low levels of qualification and long-term labour relations.

Keywords: Optimal Scaling Analysis; Missing Categorical Data; Labour markets.

1 Introduction

In the last decades, the importance of the contributions made by R. A. Fisher within the field of Statistics has been reassessed (Hald[19]; Welsh and Robinson[53]). The relevance of Fisher's methodological contributions, and his innovative approaches to probability and inference, have been recognised in this literature. However, his Optimal Scaling (OS) methods –or «the appropriate scoring technique» as Fisher[12] described them– have received lesser attention (Gower[15]). OS methods remain relatively under-utilised when compared with other contributions he made. This is particularly true of economic studies and has been attributed to the way in which Fisher introduced OS. His presentation of the subject is hard to follow within a modern multivariate data analysis context. In this sense, Savage[41] (p. 443) points out that «the world of R. A. Fisher is at once very near to and very far from the world of modern statisticians».

The idea of allocating quantitative values to qualitative concepts by means of OS methods can be traced to the seventh edition of Fisher's[12] Statistical Methods for Research Workers. After this contribution, Hayashi[20], Kruskal[26], De Leeuw

et al.[6], Takane et al.[47], De Leeuw[5], Meulman et al.[34], and Neal and Roberts[35], have developed the methodological tools that make it possible its systematic application in many areas of research.

In this paper we apply OS methods to analyse the introduction of internal or non-competitive labour markets (ILMs) in the Spanish railway company MZA at the end of the 19th century.

It has long been argued that changes in the nature of work, often as a consequence of technological innovation, impact on the organization of labour (Barley and Kunda[2]). Much attention has been paid to the relationships between employment contracts and organizational structures (Turner[49]). Of particular relevance are the changes brought about by industrialization in the late 19th century. At that time, the introduction of new production methods was sometimes accompanied by paternalistic welfare systems, resulting from the personal ethics of factory owners (Fitzgerald[14]). But welfare systems also emanated from economic and social forces since the new industrial concerns needed to train their workers in firm-specific skills. This gave rise to a retention problem (Mackinnon[30]). A solution, particularly suited to large mechanistic organizations operating in stable environments, was the introduction of ILMs (Fitzgerald[13]).

ILMs consist of a set of explicit or implicit rules and procedures governing labour relations, particularly in respect to the recruitment, training, job ladders, pay policies and job security (Lazear and Oyer[28]). In an ILM, new workers join the company at the so-called «ports of entry» (Kerr[25]), which tend to be associated with low qualification levels, and are paid the market rate for the job (Doeringer and Piore[7]). Career progression from low skilled to high skilled jobs would be limited to employees who have been in the firm for some time, in order to use seniority-based promotion ladders, something that the transaction costs model would predict (Sundstrom[46]; Siebert and Addison[44]). Pay policies are consistent with the efficiency wage model in that internal workers receive relatively high rewards with respect to the competitive or external labour market (Idson and Feaster[24]); such rewards may take the form of welfare programmes.

The development of ILMs is often associated with large modern organizations that operate in well-defined and relatively stable markets, and are able to hire workers with long term contracts (Fitzgerald[13]). This situation applies particularly to the iron and steel industry, and to large railway companies (Howlett[21]). During their growth period, railway companies had to recruit, control, train, and retain a large number of workers, becoming more bureaucratic and hierarchical (Drummond[8]). There is general agreement in the ILMs literature that these personnel policies evolved in the 20th century, but their presence in the 19th century is still an issue (Doeringer and Piore[7]). Evidence of ILMs in railway companies, in the UK and Australia, at the late 19th century is found by Drummond[9], Mackinnon[30][31], Savage[42], and Howlett[22][23], but no evidence of ILMs appears to exist for other developed countries during that period. The question is

whether the Spanish railway company MZA was operating ILM regulations at the end of the 19th century.

The MZA company was established in Madrid in 1856, and continued to operate until it was nationalised in 1941, when it became part of RENFE, the public Spanish railway monopoly. Until that moment MZA was the largest or the second largest firm in Spain, and the 10th in Europe (Tedde[48]). The study is based on the personal records of MZA employees.

The sample includes 992 employees who joined the MZA Madrid workshop during the period 1882-1889 and contains missing qualitative values. These had to be estimated using a technique that combines correspondence analysis with the k-means clustering algorithm. The approach maximises data internal consistency as measured by Guttman's squared correlation ratio (Van Buuren and Van Rijckevorsel[50]).

The predictions of the standard theories on labour markets were supported by the data. The results reveal two characteristics observed in other studies of labour relations: the existence of «ports of entry» for workers at low levels of qualification, and long-term labour relations. This paper, therefore, attempts to contribute to the understanding of the introduction of advanced management methods at the beginning of the Second Industrial Revolution, by means of a multivariate statistical approach, which is innovative in this area of research.

The rest of the paper is organised as follows. The description of the data and a brief discussion of the fundamentals of the statistical model employed are followed by the results obtained, which contains several parts. The first analysis subsection describes the way in which categorical missing data imputation took place; the second and the third analysis subsections deal with the main body of the research using the techniques of Categorical Principal Components (CPC) and Property Fitting, respectively. The paper ends with conclusions.

2 Data and methods

The data used in this study was taken from the personal information records of 992 employees of the Spanish railway company MZA hired between 1882 and 1889. Some of the information contained in the files, such as initial salary, is quantitative, and some information, such as reasons for leaving the company, is qualitative. To be able to properly analyse the multivariate information contained in the data set, it is necessary to jointly deal with qualitative and quantitative data and, therefore, with variables measured in several scale types, something that determines the statistical methods to be used (Stevens[45]). Optimal Scaling (OS) methods were employed to quantify qualitative data. This made it possible the subsequent implementation of standard multivariate analysis techniques such as CPC and Property Fitting.

There was a problem of missing categorical data that had to be addressed: out of

the 992 files, only 537 contained complete information. Missing data were estimated using an OS based method appropriate for the imputation of categorical data, MISTRESS, suggested by Van Buuren and Van Rijckevorsel[50].

2.1 Variables

The original information was recoded into eight variables: enrolment age, experience in the firm, initial wage, initial job level, leaving reason, marital status, place of birth, and section in which new workers carried out their activity. The first three variables are quantitative, while the last five are qualitative. Leaving reason was often left blank, and there was complete information for only 537 employees.

The description of initial job performed was carefully recorded in the original personal files. Taking into account job descriptions, as defined by the firm, these jobs were coded into an ordinal variable with six categories or levels, ranging from deputy director of the workshop to apprentice. It is relevant to point out that although entries into the company took place at all levels of hierarchy, most workers were hired with low level qualifications. This is consistent with the existence of «ports of entry» for new workers at low levels of qualification.

It is clearly different to be a skilled carpenter or a skilled mechanical engineer, but it was felt that this difference would be captured by the section in which the work took place. Most skilled workers were hired into jobs traditionally performed by crafts. The management and control of the crafts was a problem that had to be faced by railway companies in the 19th century (Drummond[8]). Unlike other employees, craftsmen had well established working practices, based in their tradition, and craft autonomy had to be reduced by managerial control. To address this situation, companies increased the division of labour in the workshops through a bureaucratization of human resources, this being a common characteristic of ILMs.

The personal files did not register changes that took place during the period of employment, other than the date of its termination and the reasons for this. The period studied suffered from high levels of labour unrest. This resulted in legislative attempts to improve working conditions, which had deteriorated as a consequence of the industrialization process. In this context, the reasons for leaving the firm reflect social tensions arising from the organisation of work. These reasons were classified into: death, retirement, illness, resignation, transfer away from the Madrid-Atocha workshop, redundancy, and disciplinary dismissal. It is to be noticed that the 60.5% of the workers left the company either through resignation or disciplinary dismissal.

Place of birth was recoded into four categories: Madrid, elsewhere in Spain, France, and others. The largest group of workers was from Madrid, where the workshop was situated. Separating Madrid from the rest of Spain made it possible to check differences in education between workers with an urban background, and those with an agricultural or rural industry background. MZA was set up with French capital, and French workers were the largest non-Spanish contingent.

2.2 Optimal Scaling with categorical missing data

The data used in this study contains a mixture of variables measured in nominal, ordinal, and ratio scales. OS is appropriate in this situation since it quantifies qualitative data and, by doing so, makes it possible to apply the standard methods of multivariate analysis. Fisher's[12] OS methodology has now been extended, and can be imbedded within a large set of multivariate analysis approaches, such as regression, canonical correlation analysis, discriminant analysis, principal components analysis, and multiple correspondence analysis.

OS methods treat observations as if they were categorical –either because they are thus, or because they are seen as the result of a finite measurement process, and allocates parameters to each category through an optimisation approach. Fisher's[12] approach is based on the estimation of a linear model of the ANOVA type. Following this approach, let Y be the $(n \times q)$ normalised indicator matrix of the n observations, and q categories that need to be scaled. Let β be a q length vector of the OS parameters to be estimated. The model is:

$$Y\beta = X\underset{\sim}{z} + \varepsilon \,, \qquad [1]$$

where X is the $(n \times p)$ matrix that contains the values of the remaining variables in the model, z a vector of parameters. and ε the random disturbance term.

The parameter vector, β, can be estimated using the Alternating Least Squares (ALS) approach. This requires minimising a loss function, which could be some version of Least Squares; the most popular function being Kruskal's[26] stress. There exist some algorithms that have few local minima problems (Young[54]).

An alternative way of obtaining the values of the OS parameters relies on the Singular Value Decomposition (SVD) of the data matrix:

$$X^T Y = \sum_{k=1}^{r} l_k a_k b_k^T \,, \qquad [2]$$

where $r = \min(p, q)$, $a_k \in \mathbb{R}^p$ y $b_k \in \mathbb{R}^q$ and $l_1 \geq l_2 \geq \ldots \geq l_r \geq 0$. The solution would be contained in b_1, the vector with the highest singular value in the decomposition (Welsh and Robinson[53]). In this study, we follow this approach because it has the advantage that it makes possible to perform hypotheses tests on the OS parameters, as done by Fisher[12] and Barlett[1].

In this study, much effort went into the treatment of missing values, so that these could be estimated without affecting the statistical properties of the data set. The existence of missing data is a common problem when working with data bases. Which method is appropriate to deal with missing values depends on the missingness pattern and mechanism that generates the absences (Little and Rubin[29]). The missing data pattern describes which values are missing in the data, and the mechanism describes the relationship between absent data and variable values.

In this particular case, the only variable with missing values is "leaving reason", hence the missing data pattern is univariate. The absence of leaving reason is significantly related to the values of other variables –experience in the firm, initial job level, place of birth, and section, therefore the mechanism is not Missing Completely at Random (MCAR). For example, leaving reason is missing more frequently amongst low experienced workers employed in low level jobs. One can only imagine that, if a worker does not turn up to work on a particular day, there will be an inclination to wait to see if the absence is temporary or permanent, and that permanent absences may just go unrecorded. On the other hand, death during service is a traumatic event and will probably be immediately recorded. Thus, the absence of the value of a variable may depend on the value of another variable, against the assumption of the MCAR process.

When the data is MCAR, ignoring observations with incomplete data results in correct inferences, although there is some information loss. When the presence or absence of an observation in a particular variable depends on the value of another variable, one should take into account, during model estimation, the process by which observations are missing to avoid bias (Rubin[40]).

If all the variables were measured on a ratio scale, we would be taking advantage of correlations between variables in order to estimate the most likely values for the missing data. In this particular data set, because some of the variables, including leaving reason, are qualitative, imputation of missing values is more complex (Von Hippel[51]; Ferrari and Annoni[11]). Missing data were estimated using an OS based method appropriate for the imputation of categorical data, MISTRESS, suggested by Van Buuren and Van Rijckevorsel[50]. This technique combines correspondence analysis with the *k*-means clustering algorithm is implemented to impute these values, maximizing internal consistency as measured by Guttman's squared correlation ratio.

A several step procedure was followed for the estimation of missing values: we first applied the OS model to the data that excluded employees with missing observations (Portillo et al.[38]); second, correlations between variables were studied; third, the relationships between variables were explored using Path Analysis (Retherford and Choe[39]); and, finally, a homogeneity analysis based procedure due to Van Buuren and Van Rijckevorsel[50] was followed in order to obtain the imputations that were used in the final OS analysis of the data.

3 Results

3.1 Categorical Missing Data

The application of OS to the data set that contains no missing values –i.e., 537 observations– generated a set of correlations between variables. A study of the correlation structure of the data was next performed using Path Analysis. The

AMOS routine of SPSS was run using as endogenous variables experience in the firm, initial salary, and leaving reason. The remaining variables –enrolment age, initial job level, marital status, place of birth, and section of work– were treated as exogenous. The only two relevant path coefficients associated with leaving reason that took values significantly different from zero were the ones that linked leaving reason with enrolment age and with experience in the firm.

The final step was to use the MISTRESS imputation procedure of Van Buuren and Van Rijckevorsel[50]. This procedure maximizes consistency in the data set as measured by Guttman's squared correlation ratio, η^2 (Guttman[18]). The MISTRESS technique combines correspondence analysis methods with the k-means clustering algorithm. The variables found to be associated with leaving reason in the previous analyses –enrolment age, and experience in the firm– were used to obtain estimates of the missing values, whose distribution is presented in table 1.

Table 1. Missing data imputation for employees leaving reason

Leaving reason	Observed		MCAR ‡ expected		MISTRESS Imputed		χ^2–statistic † (p–value)
	Count	%	Count	%	Count	%	
Missing	352	39.6	0	0.0	0	0.0	68.720
Resignation	228	25.6	378	42.5	472	53.1	(≤ 0.01)
Dismissal	80	9.0	132	14.9	98	11.0	
Redundancy	27	3.0	45	5.0	59	6.6	
Transfer	38	4.3	63	7.1	76	8.5	
Illness	38	4.3	63	7.1	39	4.4	
Death	86	9.7	142	16.0	95	10.7	
Retirement	40	4.5	66	7.4	50	5.6	

‡ Missing Completely at Random.
† $\chi 2$ test for the difference between the expected distribution under MCAR mechanism and MISTRESS estimates.

The percentage of missing values was 39.6% with respect to the total in leaving reason, and 4.9% with respect to the total data base. These percentages are within the acceptability limit of unknown entries required to avoid consistency bias, obtained by Van Buuren and Van Rijckevorsel[50] using the bootstrap. Guttman's η^2–statistic took the value 0.68, which is higher than the 0.50 limit required to maximize consistency, also obtained by these two authors using the bootstrap.

3.2 Optimal Scaling and Categorical Principal Components (CPC)

After imputing missing values, the data for the eight variables on 889 workers were analysed using CPC. The first two components accounted for more than 50% of the variation in the data, and the first five components accounted for almost 90%.

Table 2. Optimal Scaling scores for qualitative variables

Variable		Frequency n = 889	Optimal Scaling	
			Scores	F–statistic † (p–value)
Initial job level	Third level	4	2.548	26.358 (≤ 0.01)
	Fourth level	59	2.238	
	Fifth level	353	0.766	
	Sixth level	473	–0.873	
Leaving reason	Resignation	472	–0.601	25.698 (≤ 0.01)
	Disciplinary dismissal	98	–0.438	
	Redundancy	59	–0.224	
	Transfer	76	–0.158	
	Illness	39	2.021	
	Death	95	1.599	
	Retirement	50	2.421	
Marital status	Single	484	–0.903	3.682 (≤ 0.01)
	Married	391	1.035	
	Widower	14	2.323	
Place of birth	Madrid	283	0.692	12.335 (≤ 0.01)
	Rest of Spain	559	0.015	
	France	30	4.257	
	Other countries	17	3.511	
Section	Forge	70	–1.358	15.249 (≤ 0.01)
	Assembly	96	–1.181	
	Boiler making	107	–0.898	
	Foundry-elect. installations	74	–0.738	
	Fitting	64	–0.172	
	Carriages	219	0.352	
	Upholstery- Paintwork	252	0.959	
	Offices	7	1.715	

† Fisher's[12] test for the global significance of OS parameters.

The scores assigned to the categories of each non-numerical variable are given in table 2. Score values are important for interpretation purposes.

Scores in initial job level increase as seniority increases: the lowest score in initial job is associated with apprentices and the highest score is attached to supervisors. Moving on to the scores of the various categories of leaving reason, it is to be noticed that abnormal reasons –resignation, dismissal, and redundancy– show negative scores, while categories that are a normal end to an employment career –illness, death, and retirement– show positive scores. The ordering of the scores for the different categories of place of birth is to be noted. The higher scores are attached to the categories France and other countries. This makes sense in the context of the MZA firm, since it is suspected that the not-Spanish workers were specialists who joined the firm for a short period of time.

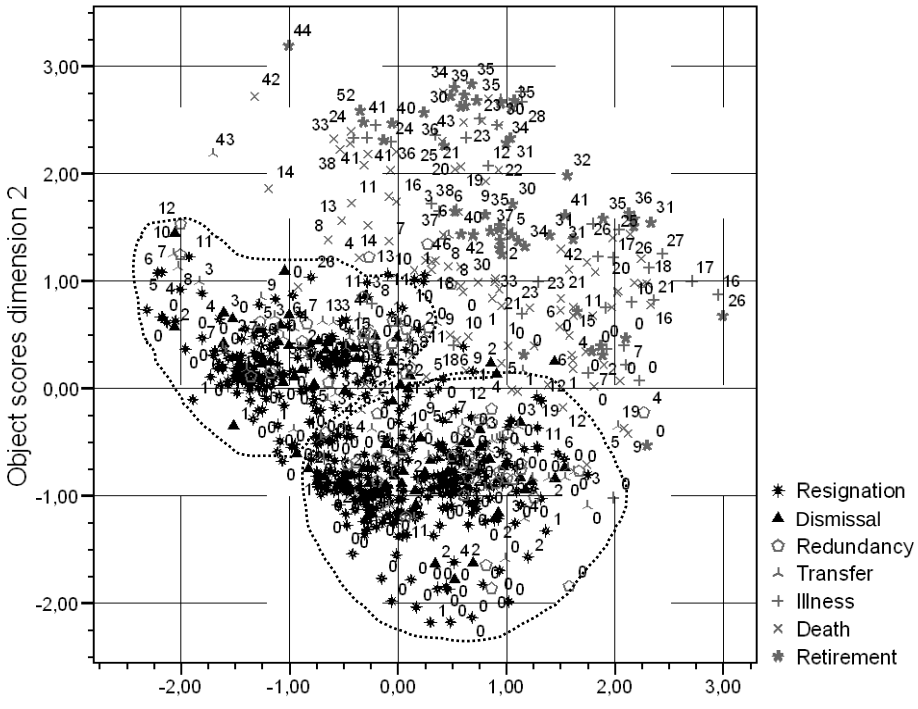


Fig. 1. Workers plotted in the first two categorical principal components by leaving reason with indication of experience in the firm.

Table 2 also shows the result of Fisher's[12] test for the global significance of estimated optimal scores for qualitative variables. The F−statistic for all variables was significant at the 1% level, indicating that all variables included in the model are relevant. So, we conclude that there are highly significant differences amongst the workers hired by MZA, and that these differences are captured by their initial job, leaving reason, marital status, place of birth, and section in which they were employed. Fisher[12] highlights the importance of this result, since it is only under this condition that the scores estimated by the OS model are meaningful.

The variables that load highest in the first component were found to be: enrolment age, marital status, initial job level, and initial salary. Taking into account the quantifications produced by OS and factor loadings, we interpreted the first principal component in terms of experience before the worker joined the firm, or "prior experience". The variables that load high in the second component are experience in the firm, and leaving reason. The second component is, therefore, associated with "permanence in the firm".

The first two categorical principal components are represented in figure 1. In

that figure, the points are labelled with experience in the firm, and with leaving reason. At the top right hand side of figure 1 we find the workers who abandoned the company for causes of retirement, illness, or death. In the same region of the figure we find the workers with a long period of attachment to the company. At the bottom left hand side of figure 1 we find two main groups of workers who left through resignation, dismissal, and redundancy. We see that workers with a short career in the company are also to be found in this area of figure 1.

It is clear from the representation that there is a strong negative association between experience in the company and leaving reason. There are other relationships of interest between the variables, and these will be represented using the method of Property Fitting.

3.3 Property Fitting

Property Fitting is a regression based technique that can be located within the context of Biplots (Gower and Hand[16]). Details of the particular form of the algorithm implemented here can be found in Kruskal and Wish[27], Schiffman et al.[43], and Mar Molinero and Mingers[32]. Property Fitting draws normalized vectors of unit length through the space of the principal components in the direction in which a property of the data grows. The coordinates of vectors can be interpreted as directional cosines. The longer the projection of the vector, the more relevant is the two dimensional plot to the interpretation of the results. The angle between any two vectors depends on the correlation between the variables involved. Acute angles indicate positive correlation, the smaller the angle the higher the correlation. Orthogonal vectors are associated with a lack of correlation.

The vectors obtained are shown in figure 2. Take, for example, the variable enrolment age: the associated vector points towards the right hand side of the figure, indicating that the older a worker was when joining, the more to the right of the configuration he will be plotted. Two further features highlighted in figure 2 are whether the observation contains imputed values or not, and initial job level. Most workers for whom the leaving cause was estimated are plotted towards the left and towards the bottom of the figure. Two groups of workers have been circled in this area. Both groups are similar in the sense that the workers did not stay long in the firm and left the firm by resignation or dismissal. But, as the oriented vectors indicate, the two groups are different in the sense they correspond to workers with different initial salaries and job levels. The first group, situated towards the left, is made up of individuals who did not stay long in the firm, were rather young when they joined, and were hired in low level jobs with low initial salaries. The second group of workers with low job tenure, situated at the bottom of the figure 2, were hired with high initial salaries and at high level jobs. We also see that initial salary, initial job level, section, and place of birth are not related to the absence in the worker's file of the leaving reason. This confirms the validity of the missing data procedure used in the analysis.

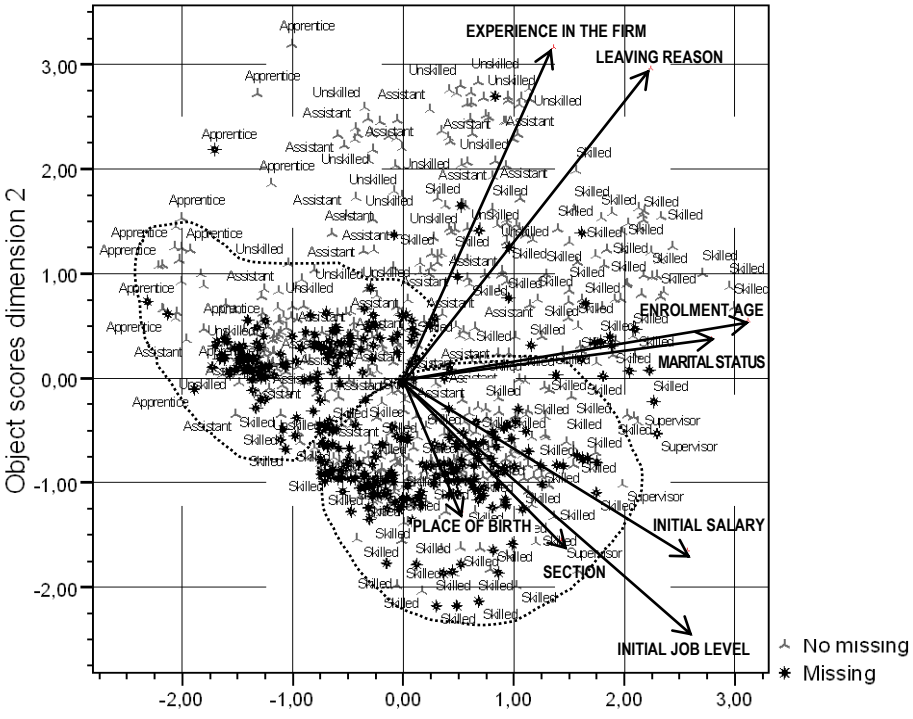


Fig. 2. Workers plotted in the first two categorical principal components by missing data for leaving reason, and projection of Property Fitting vectors

In conclusion, figure 2 jointly represents the workers and the variables measured on them. It would be reasonable to assume that there is association between joining age, and seniority –i.e., that workers who join when young are less well qualified and are assigned lower seniority than workers who join when they are older. This implies a positive correlation between enrolment age, and initial wage. Such positive correlation is present, as can be seen by the acute angle between the vectors enrolment age and initial salary in figure 2. An examination of figure 2 also reveals a positive correlation between enrolment age and initial job performed, indicating that in the firm MZA experienced workers tended to be hired into more skilled jobs. This is consistent with the influence of external labour markets on internal wages through ports of entry.

Two vectors point towards the top right hand side of figure 2; these are associated with experience in the firm, and leaving reason. The angle between these two vectors is acute and small, indicating a strong positive correlation between these two variables. This means that workers who remained in the firm for a long period

of time left it for reasons of retirement, illness or death; and that workers with short periods of employment in the firm left it for reasons of dismissal, redundancy, or resignation. In other words, workers who "survived" the initial period of employment stayed on until retirement or death, or until they had to leave for reasons of illness: they do not resign. These results are consistent with the specific human capital theory. This theory predicts that resignation and dismissal rates decrease as experience in the firm increases (Parsons[37]). However, these findings are at odds with the outcomes to be expected under the deferred compensation model, which states that, once the specific training period is over, salaries exceed worker's marginal productivity, dissuading workers from leaving, but incentivizing dismissals (Medoff and Abraham[33]; and Siebert and Addison[44]).

It is also apparent in figure 3 that leaving reason and initial salary are not associated –the relevant vectors are orthogonal. There is also a lack of association between initial salary and experience in the firm. This is consistent with the view that conditions of employment were determined by the labour market and did not influence workers' permanence in the firm. Permanence in the firm is much better explained by the long-term relationships that ILMs need in order to transform into benefits the costs incurred during the initial employment period. Under the human capital theory, and under the deferred compensation model, ILM relationships act as a mechanism for the selection of the most stable workers (Becker and Stigler[4]; Becker[3]). The explanation is that, since both the firm and the workers share specific training costs, there are incentives not to break the labour relationship through premature dismissals or resignations. Furthermore, the governance structures that arise under ILMs dissuade both workers and employers from adopting opportunist behaviours (Wachter and Wright[52]).

4 Conclusions

In this paper we have addressed methodological and business questions. The methodological questions relate to missing values and to the correct procedures to deal with a mixture of qualitative and quantitative variables. The treatment of missing data is not straightforward. A model based only on observations for which complete information is available may not be representative of the population unless the data is MCAR. In this study we have established that the data was not MCAR, and that the missing data contained valuable information about the process under study. We followed a several step process for missing value imputation that culminates on the maximization of internal consistency. This made it possible to operate with the full data set without dropping any observations.

A further difficulty was the co-existence of qualitative and quantitative variables. OS procedures are appropriate in these cases. OS was introduced by Fisher[12], but it has seldom been applied to the analysis of business problems. In

this paper we have used OS in order to transform qualitative information into quantitative. After this, we have applied standard tools of multivariate analysis. We have also used hypotheses tests originally developed by Fisher[12] in order to establish the relevance of the variables in the model.

Turning now to the business agenda, the study of the way in which ILMs developed and evolved has taken momentum in the last decades (Elvira and Graham[10]; Osterman[36]). Much recent work on ILMs has focused on their evolution and actual implications, arguing that some of the key principles of ILMs have been now dismantled (Grimshaw et al.[17]).

The theory suggests that, when ILMs are present, companies relate to the external market through the so-called «ports of entry». Companies use these in order to hire new workers with no particular training. Workers are expected to undergo an initial period of hardship that will be eventually rewarded by means of wage and social benefits. The theory also predicts that most workers who exit the company in an irregular way do so soon after joining, since opportunity costs can be high for those who have accumulated some experience in the firm.

The main objective of this paper was the study of the introduction of ILMs in the Spanish railway company MZA at the end of the 19th century. The results reveal two characteristics observed in previous studies of labour relations: the existence of «ports of entry» for workers at low levels of qualification, and long-term labour relations. This paper, therefore, attempts to contribute to the understanding of the introduction of advanced management methods at the beginning of the Second Industrial Revolution, by means of a multivariate statistical approach, which is innovative in this area of research.

References

1.Barlett, M. S. The goodness of fit of a single hypothetical discriminant function in the case of several groups. Annals of Eugenics, 16, 199-214, 1951.
2.Barley, S. R. and Kunda, G. Bringing work back in. Organization Science, 12, 76-95, 2001.
3.Becker, G. S. Human capital: A theoretical and empirical analysis. New York: Columbia University Press, 1975.
4.Becker, G. S. and Stigler, G. J. Law enforcement, malfeasance, and compensation of enforces. Journal of Legal Studies, 3, 1-18, 1974.
5.De Leeuw, J. Least Squares Optimal Scaling of Partially Observed Linear Systems. In Recent developments on structural equation models: Theory and applications (eds. K. van Montfort , J. Oud and A. Satorra). Norwell, Mass.: Kluwer, 2004.
6.De Leeuw, J., Young, F. W. and Takane, Y. Additive structure in qualitative data: an alternating least squares method with optimal scaling features. Psychometrika, 41(4), 471-503, 1976.
7.Doeringer, P. B. and Piore, M. J. Internal labour markets and manpower analysis. Lexington, Mass.: Lexington Books, Heath, 1971.
8.Drummond, D. Specifically designed? Employers´ labour strategies and worker responses in British railway workshops, 1838-1914. Business History, 31, 10-11, 1989.
9.Drummond, D. Crewe: Railway town, company and people, 1840-1914. Aldershot: Scholar

Press, 1995.
10.Elvira, M. M. and Graham, M. E. Not just a formality: Pay system formalization and sex-related earnings effect. Organization Science, 13(6) 601-617, 2002.
11.Ferrari, P and Annoni, P. Missing data in Optimal Scaling. Working Paper 2005-19, Dipartimento di Scienze Economiche e Statistiche. Universita di Milano, Milano, 2005.
12.Fisher, R. A Statistical methods for research workers (7th and 8th editions). Edinburgh: Oliver and Boyd, 1938, 1941.
13.Fitzgerald, R. British Labour Management and Industrial Welfare, 1846-1939. London: Croom Helm, 1988.
14.Fitzgerald, R. Employment relations and industrial welfare in Britain: Business ethics versus labor markets. Business and Economic History, 28 167-179, 1999.
15.Gower, J. C. Fisher's Optimal Scores and Multiple Correspondence Analysis. Biometrics, 46, 947-961, 1990.
16.Gower, J. C. and Hand, D. J. Biplots. London: Chapman & Hall, 1996.
17.Grimshaw, D., Ward, K., Rubery, J. and Beynon, H. Organisations and the transformation of the Internal Labout Market. Work, Employment and Society, 15(1), 25-54, 2001.
18.Guttman, L. The quantification of a class of attributes: A theory and method of scale construction. In The prediction of personal adjustment (P.H. Horst ed.), 319-348. Social Sciences Research Council, New York, 1941.
19.Hald, A. A History of Mathematical Statistics from 1750 to 1930. New York: Wiley, 1998.
20.Hayashi, C. On the predictions of the phenomena from qualitative data and quantifications of qualitative data from the mathematico-statistical point of view. Annals of the Institute of Statistical Mathematics, 3, 69-92, 1952.
21.Howlett, P. Evidence of the existence of an internal labour market in the Great Eastern Railway Company, 1875-1905. Business History, 42(1), 21-40, 2000.
22.Howlett, P. Careers for the unskilled in the Great Eastern Railway Company, 1870-1913. London School of Economics Working Paper Series 63/01, London, 2001.
23.Howlett, P. The internal labour dynamics of the Great Eastern Railway Company, 1870-1913. Economic History Review, LVII(2), 396-422, 2004.
24.Idson, T. and Feaster, D. A selectivity model of employer-size wage differentials. Journal of Labor Economics, 8(1), 99-122, 1990.
25.Kerr, C. The Balkinization of Labor Markets. In Labor mobility and economic opportunity (E. W. Bakke ed.), pp. 92-110. Cambridge, MA: MIT Press, 1954.
26.Kruskal, J. B. Multidimensional Scaling by optimizing goodness of fit to a non-metric hypothesis. Psychometrika, 29(1), 1-27, 1964.
27.Kruskal, J. B. and Wish, M. Multidimensional Scaling. London: Sage, 1978.
28.Lazear, E. P. and Oyer, P. Internal and external labor markets: a personnel economics approach. NBER Working Paper Series, No.10192, Cambridge, MA, 2003.
29.Little, R. J. A. and Rubin, D.B. Statistical analysis with missing data. Hoboken, New Jersey: Wiley & Sons Inc, 2002.
30.Mackinnon, M. The Great War and the Canadian labour market: railway workers 1903-39. In Labour Market Evolution (G. Grantham and M. MacKinnon eds.), pp. 205-224. London: Routledge, 1994.
31.Mackinnon, M. Trade Unions and Employment Stability at the Canadian Pacific Railway, 1903-1929. In Origins of the Modern Careers (D. Mitch, J. Brown and M.H.D. van Leeuwen eds.), pp. 126-144. Aldershot: Ashgate, 2004.
32.Mar Molinero, C. and Mingers, J. Mapping MBA Programmes: an alternative analysis. Journal of the Operational Research Society, 58, 874-886, 2007.
33.Medoff, J. L. and Abraham, K. G. Are those paid more really more productive? The case

of experience. Journal of Human Resources, 16(2), 186-216, 1981.
34. Meulman J. J., van der Kooij, A. J. and Heiser, W. J. Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In Handbook of quantitative methodology for the social sciences (D. Kaplan ed.), pp. 49-70. Thousand Oaks, CA: Sage Publications, 2004.
35. Neal, P. J. and Roberts, G. O. Optimal Scaling for partially updating MCMC algorithms. The Annals of Applied Probability, 16(2), 475-515, 2006.
36. Osterman, P. The Wage Effects of High Performance Work Organization in Manufacturing. Industrial and Labor Relations Review, 59(2) 187-204, 2006.
37. Parsons, D.O. The Employment relationship: Job attachment, work effort and the nature of contracts. In Handbook of Labor Economics (R. Layard and O. Ashenfelter eds.), pp. 789-848. Amsterdam: North-Holland, 1986.
38. Portillo, F., Mar Molinero, C. and Martinez, T. Interpreting a data base of railway workers using optimal scaling techniques. Kent Business School Working Paper Series, 127, Canterbury, UK, 2006.
39. Retherford, R. D. and Choe, M. K. Statistical Models for Causal Analysis. New York: John Wiley & Sons, 1993.
40. Rubin, D. B. Inference and missing data. Biometrika, 63 581-592, 1976.
41. Savage, L. J. On reading R.A. Fisher. The Annals of Statistics, 4, 441-500, 1976.
42. Savage, M. Discipline, surveillance and the career: Employment on the Great Western Railway 1833-1914. In Foucault, management and organisation theory (A. McKinlay and K. Starkey eds.), pp. 65-92. London: Sage, 1998.
43. Schiffman, S. S., Reynolds, M. L. and Young, F. W. Introduction to Multidimensional Scaling: Theory, Methods and Applications. London: Academic Press, 1981.
44. Siebert, W. S. and Addison, J. T. Internal labour markets: causes and consequences. Oxford Review of Economic Policy, 7(1), 76-92, 1991.
45. Stevens, S. S. Mathematics, measurement, and psychophysics. In Handbook of Experimental Psychology (S. S. Stevens ed.), pp. 1-49. New York: Wiley, 1951.
46. Sundstrom, W. A. Half a carrer: discrimination & Railroad Internal Labor Markets. Industrial Relations, 29(3), 423-440, 1990.
47. Takane, Y., Young, F. W. and De Leeuw, J. An individual differences additive model: An Alternanting Least Squares method with Optimal Scaling features. Psychometrika, 42, 7-67, 1980.
48. Tedde, P. La expansión de las grandes compañías ferroviarias españolas: Norte, MZA y Andaluces. In La empresa en la historia de España F. Comín and P. Martín-Aceña eds.), pp. 265-284. Madrid: Civitas, 1996.
49. Turner, T. Internal labour markets and employment Systems. International Journal of Manpower, 15(1), 15-26, 1994.
50. Van Buuren, S. and Van Rijckevorsel, J. L. A. Imputation of missing categorical data by maximizing internal consistency. Psychometrika, 57, 567-580, 1992.
51. Von Hippel, P. T. Biases in SPSS 12.0 missing value analysis. The American Statistician, 58 160-164, 2004.
52. Wachter, M. L. and Wright, R. D. The economics of internal labour markets. Industrial Relations, 29(2), 240-262, 1990.
53. Welsh, A. H. and Robinson, J. Fisher and inference for scores. International Statistical Review, 73, 131-150, 2005.
54. Young, F. Quantitative analysis of qualitative data. Psychometrika, 46(4), 357-388, 1981.

Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013)
International Conference, Mataró (Barcelona), Spain 25 - 28 June 2013

Early Forecasting of Parliamentary Seats Distribution: the Representative Polling Stations Method

Ana Martín-Arroyuelos[1], Karmele Fernández-Aguirre[2] and Juan Ignacio Modroño-Herrán[3]

Departamento de Economía Aplicada III, Facultad de Ciencias Económicas y Empresariales, Universidad del País Vasco (UPV/EHU), Avda. Lehendakari Aguirre 83, 48015 Bilbao, Spain.

[1] E-mail: ana.martin@ehu.es
[2] E-mail: karmele.fernandez@ehu.es
[3] E-mail: juan.modrono@ehu.es

Abstract. On election nights, once the polling stations are closed, there is an evident interest in early estimation of the number of parliamentary seats obtained by each political party, based on actual counting of votes. This paper focuses on the representative polling stations method to forecast the aforementioned number of seats. This method involves selecting a reduced group of stations whose aggregate result turns out to be a sufficiently accurate prediction of the final distribution of seats. This paper presents a method based on past results which is both useful to select such representative polling stations and then able to choose the needed proportion of them that should be considered. Emphasis is set out on forecasting accuracy.

Keywords: Parliamentary seats prediction, Representative polling stations, Distances, Linear Regression

1 Introduction

It is widely known that preliminary forecasts of parliamentary seats are disclosed as soon as possible after polling station closures on election days. Such forecasts are made on actual polling stations results as they are progressively becoming available from official sources. If data inflow was purely random since a very early stage, it would suffice to use probabilistic sampling techniques in order to obtain such forecasts. However, data inflow is all but random due to the fact that the results received refer to polling stations where votes have been fully counted. The speed in the count process depends on different and uncontrollable factors. As a result, it is often observed in practice that initial forecasts significantly differ from the final results, where intermediate predictions show clear variations on the initial estimates. Methods in the literature addressing these issues include Bernardo [1], Bernardo and Girón [3] and Pavía, Muñiz and Álvarez [8].

In this paper we use the representative polling stations method. The main reason behind this choice is that it is shown that if representative stations per electoral district are chosen in an adequate way, a reduced number of them is sufficient to achieve an almost error-free forecast of the final results. Thus, in Martín, Fernández and García [7] this method was compared to the polling station segmentation method which consists on building groups of stations with similar vote proportions using Correspondence Analysis combined with Automatic Classification (Lebart [6], Greenacre [5]). Such comparison was tested using the Juntas Generales (a sort of parliament at the province level) election data for the basque province of Bizkaia in 2007, obtaining the forecast of such groups. A comparison was also run with respect to the forecast obtained from no segmentation whatsoever. The representative poll stations method clearly outperformed these other methods.

This work is focused on the representative polling stations method behavior in order to extract conclusions on the proportions of representative stations to be chosen in order to get accurate estimations of both the final voting proportions and the parliamentary seats distribution obtained thereof. This proportion is a relevant issue as long as the consideration of a too small number of stations may imply that possible erratic variations in some of them have a substantial effect on the final forecast. On the other hand, given that obtaining a reduced number or representative stations is actually the core objective of this method, it is of key interest to know which proportion of stations makes the addition of new stations irrelevant (if not counter-productive) for significant prediction improvements.

We will apply the proposed method using information from the period 2000-2012 to get forecasts on the last seven elections run in the Bizkaia province in this work. In detail, these are the General (nationwide) Elections of 2008 and 2011, the Basque Parliament (regional) Elections of 2005, 2009 and 2012 and the Juntas Generales (local, province level) Elections of 2007 and 2011. The data is available from the

Basque Government Department of Security web site (http://www.euskadi.net/elecciones/). In the procedure, we will be using four previous consecutive elections for results forecasting of the next election which means that the oldest elections are used only as predictors, whereas the newer ones are both used as predictors and as forecasting targets (except the last election).

When considering consecutive elections it is evident that there are changes in polling stations composition, due to entries and exits, and in the actual number and coding of stations. New stations appear, some existing ones are eliminated and sometimes stations split up or merge. This makes necessary some homogenization work among elections analyzed together.

This work is organized as follows. After this first introduction section, we provide details on the method chosen for stations homogenization, prior to any analysis, in the second section. In the third section we explain briefly the representative polling stations method and the criteria used for station selection. Later, in the fourth section we explain how a reduced number of representative polling stations is needed for seats forecasting. At this point, some regularities are observed in this method, which consequently allows the proposal of a procedure to establish the proportion of representative stations to be chosen depending on electoral district characteristics. Next, in order to test the reliability of the procedure, we will estimate voting proportions for each political party and the number of parliamentary seats obtained by each party for all elections considered in this work. Finally, we draw conclusions based on the results obtained.

2 Homogenization of polling stations across different elections

We detail in this section the way polling stations have been homogenized, given that several different past elections have been considered for parliamentary seats forecasting. We take as reference the polling stations of the last year, that is, the election where seats prediction is to be obtained. The homogenization procedure is performed according to the following rules:

1. When a station splits up into two new ones, we assume that the proportions of votes for the political parties in the new stations are the same as those from the original one. If two stations split up into three new ones, then the voting proportions of the three are the mean of the original two stations, weighted by census size.
2. When two stations merge, we simply assign the weighted mean of the two to the resulting one, being the weight the respective censuses.
3. If a brand new station arises, the results for the previous years are imputed using the average of stations in the district.
4. If a station disappears, there are two possible alternatives. Sometimes the disappearance occurs at the same time that one adjacent station splits up into two ones; in this case the absent station is assigned to the unfolded one. If such unfolding does not occur there is no clue on where to assign that station and it is simply deleted; however, this is very infrequent.

All these changes have been performed backwards in time, being the polling station distribution of the last election, the forecast target, the one considered for analysis without any modifications.

In forecasting of election results based on past elections it is statistically evident that considering a large number of previous elections is important. At the same time, it is noted that when the time gap between elections increases, the differences in stations distribution also increases, as one would probably expect. This makes it difficult to consider dependencies between elections celebrated too far in time. We have finally considered 5 consecutive elections for this reason. It should also be noted that increasing time gaps between elections modifies the inner composition of station voters, simply by demographic reasons.

3 Method of the representative polling stations

This method consists on the detection, inside each electoral district, of polling stations with a similar distribution to the district average and the selection of a small number of them to forecast the number of parliamentary seats to be assigned to the different political parties.

The procedure to get the predictions is as follows. The sums of all votes for each party for all polling stations are calculated and the proportions of them with respect to all valid votes (excluding blank ones) are also calculated. These proportions are also computed for their corresponding electoral districts.

In order to get the stations with distribution of votes closer to the district average we use the distances between the votes of the station and those of the district such station belongs to.

Thus, if $(p_1, p_2, \ldots, p_k)$ is the vector containing the number of votes obtained by the k parties in the district and $(p_{i1}, p_{i2}, \ldots, p_{ik})$ the vector containing votes obtained by the k parties in the i-th polling station belonging to the district, the interest is in the calculation of the distance between both vectors. In this work we have used three different distance measures, as specified next.

a. Euclidean Distance: $D^2_{eucl,i} = \sum_{j=1}^{k} (p_{ij} - p_j)^2$

b. Chi-square Distance: $D^2_{chi,i} = \sum_{j=1}^{k} \frac{(p_{ij} - p_j)^2}{p_j}$

c. Discrepancy Measure (Bernardo [1]): $D_{discrep,i} = \sum_{j=1}^{k} p_j \log\left(\frac{p_j}{p_{ij}}\right)$

Given that we consider information relative to the four elections prior to the one we want to obtain predictions for, we calculate, for every such election year, distances relative to differences in voting proportions from each polling station to the voting proportions in the final district results. Then, we sum up the four values obtained from each of such past elections. We intend to avoid the inclusion of stations that get a result, by chance, similar to the average; we try to favor stations that are permanently similar. We do not go earlier than 4 years back as the composition and distribution of the polling stations begins to change considerably.

As the distance measures considered are different, we get three different sets of polling stations. These sets are not mutually exclusive, bearing some elements in common, that is, stations that are selected with more than one distance measure.

4 Is a small number of representative polling stations enough for forecasting?

In this section we justify why a low number of representative stations is needed per district for parliamentary seats forecasting, using the election data of the Bizkaia province in the period 2000-2012. We also show the little relevance of the distance choice for the selection of the number of representative stations.

In this analysis twenty possible cases are available given that, for each election, we have considered full results for the province and partial results correspondent to the four existing electoral districts. The relevance of full and partial results depends on which kind of parliament (nationwide, local, ...) the election refers to. In this way, we have 5 cases per election.

For each combination of district-election and distance, the following steps have been followed:

1. The thirty most representative polling stations are initially selected, being so those with the smallest distance to the result of the corresponding district.
2. The estimation of the voting proportions for all parties is done as a sequence of cumulative frequencies. The first estimation uses the most representative station, the second the two most representative ones, the third the three most representative ones, and so on, up to the thirty most representative ones.
3. We draw a time series plot of the differences between the estimated voting proportions and the actual voting proportion obtained in the district for each party, where time is replaced by the number of stations used in the estimation. This plot allows for the selection of the minimum number of stations needed to get a good forecast of the final result. Selecting a higher number of stations eventually worsens the results prediction, as the newly added stations are clearly less representative than the first ones.

As an illustration, we display the estimated-actual voting proportions differences plots obtained for parties obtaining parliamentary representation (pnv, bildu, psoe, pp, iu and upyd) in the last election considered (Basque parliament, 2012). We have considered the three measures of distance mentioned above (Euclidean, chi-square and Bernardo's) for representative stations selection and displayed one plot per distance measure. 1% and 2% bounds for the differences are plotted as horizontal dashed lines.

We make use of the three plots (Fig. 1, Fig. 2 and Fig. 3) in order to select the number of representative polling stations to get a more reliable number. Otherwise it might be the case that the optimum number of polling stations was missed, should it depend on a particular distance measure. When simultaneous consideration of these three distances is incorporated, this possibility gets reduced.

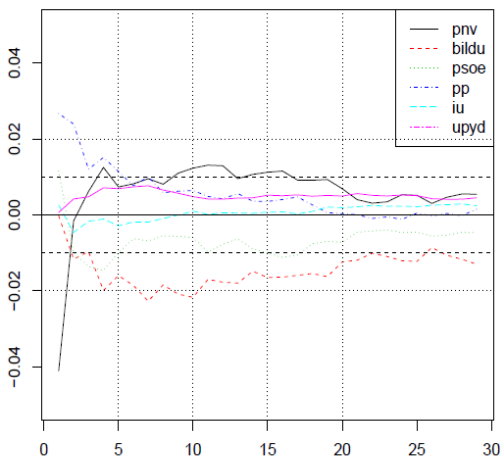


Fig. 1: Differences between actual and forecast results by number of Representative Stations 2012. Euclidean Distance.

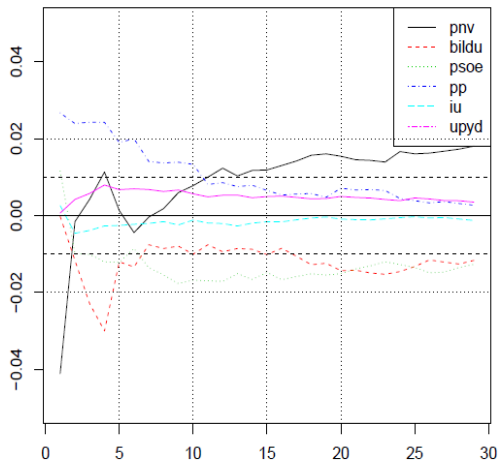


Fig. 2: Differences between actual and forecast results by number of Representative Stations 2012. Chi-square Distance.

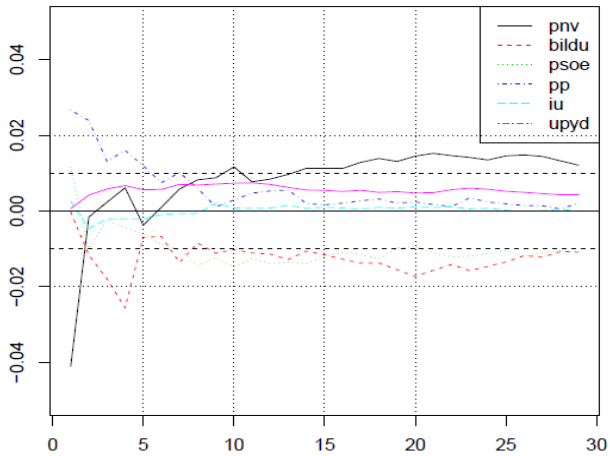


Fig. 3: Differences between actual and forecast results by number of Representative Stations 2012. Bernardo's discrepancy measure.

If we select the Euclidean distance (Basque parliament elections feature a single district) 6 polling stations are sufficient to get an actual-forecast difference below 1% for all political parties, with the exception of one (bildu). When considering the chi-square distance, such difference is achieved with 7 polling stations, being one party (pp) slightly badly predicted. Finally, choosing Bernardo's distance the 1% difference is met with 6 stations. We believe that the first 6 most representative polling stations are likely enough for seat forecasting, as it is the average of the needed stations considering the three criteria.

In this particular case, as in most others, the number of selected stations is very similar, irrespective of the distance criterion chosen. However, sometimes clear differences arise. As a general rule, such differences do not show a pattern making one of them preferable to any other.

5 Optimum proportion of representative polling stations for forecasting

Firstly, it is observed that the proportion of representative polling stations to be considered in each case is different for different electoral districts.

The first hypothesis set out in this paper is that one of the factors behind the selection of the number of representative polling stations is electoral district size. It seems evident that the proportion of stations decreases with district size. This is a logical result, as this means that if there are many stations inside a district, it is more likely that we are able to find a larger number of stations featuring voting distributions tightly closed to the actual average distribution, so a good approximation to the actual results is achieved earlier.

A second hypothesis is that result variability may also affect the representative polling stations number, for each district. In order to do that, we define a variability measure which is equal to the average of the distances of all polling stations with respect to the actual result obtained for the previous 4 past elections. This variability measure is computed for the three distances specified in section 3.

The obtained results show that, considering electoral districts of similar size, when the variability measure increases the proportion of stations to be considered also increases.

At this point we want to establish the adequate proportion of representative polling stations to be considered, as a function of the variability and the size of the electoral district. Data on size and variability are computed for the last 7 elections which, when considering the existing districts, results in a sample of 30 observations of these variables. To get such proportion, we obtain the forecasts from a linear regression run on the observations excluding those from the election to be forecasted, where the dependent variable is the proportion and the independent variables the size and the variability of the electoral district, in the form:

$$Proportion = \alpha + \beta\ size + \gamma\ var(dist)$$

where "*dist*" is one of Euclidean (eu), chi-square or Bernardo's. In order to select which is best, different criteria could be used. If we choose the criterion of the highest coefficient of determination in such regressions, the Euclidean distance turns out to be the winner. To be coherent with the distance measure used, the selection of the representative polling stations will also be done with this distance. Nevertheless, using the other two alternative distances, the results are still very similar.

6 Application of the method for all elections considered

Next we show, to illustrate this method, the results obtained using the representative polling stations that should have been selected for early forecasting on election nights for the last three elections celebrated in the Basque province of Bizkaia, that is, Juntas Generales of 2011, General Elections of 2011 and Basque Parliament Elections of 2012. For Juntas Generales elections, the province is divided into four electoral districts so we get forecasts for each district separately. The other elections are run with a single electoral district and get a unique forecast. Actual-forecast parliament seats are given for the three elections through Tables 1-6 respectively.

For Juntas Generales 2011, the linear regression for forecasting the proportion/number of representative polling stations is

$$Proportion = 0.0182 - 2.08528*10^{-5} * size + 0.0929 * var(eu) \qquad R^2 = 0.4468$$

and the actual-forecast parliament seats are in Tables 1-4, one per district.

	Voting proportions		Parliamentary Seats	
Party	**Actual**	**Forecast**	**Actual**	***Forecast***
PNV	38.93	39.98	7	8
PP	19.22	19.51	3	3
PSE	16.21	16.09	3	3
BILDU	15.18	14.22	3	2
EB	4.03	4.05	0	0

Table 1: Actual-Forecast results, Juntas Generales 2011 (District 1). 7 Polling Stations. Var(eu): 0.0745

	Voting proportions		Parliamentary Seats	
Party	**Actual**	**Forecast**	**Actual**	***Forecast***
PNV	41.12	41.55	6	6
PP	13.22	13.79	2	2
PSE	9.34	9.59	1	1
BILDU	29.10	26.32	4	4
EB	2.19	2.53	0	0

Table 2: Actual-Forecast results, Juntas Generales 2011 (District 2). 8 Polling Stations. Var(eu): 0.1255

	Voting proportions		Parliamentary Seats	
Party	**Actual**	**Forecast**	**Actual**	***Forecast***
PNV	35.51	35.59	4	4
PP	10.36	11.19	1	1
PSE	17.53	18.89	1	2
BILDU	27.36	24.09	3	2
EB	3.73	4.69	0	0

Table 3: Actual-Forecast results, Juntas Generales 2011 (District 3). 9 Polling Stations. Var(eu): 0.1985

	Voting proportions		Parliamentary Seats	
Party	**Actual**	**Forecast**	**Actual**	***Forecast***
PNV	35.84	35.60	5	5
PP	12.10	12.07	2	2
PSE	25.76	25.40	4	4
BILDU	16.19	16.11	2	2
EB	4.49	5.45	0	0

Table 4: Actual-Forecast results, Juntas Generales 2011 (District 4). 12 Polling Stations. Var(eu): 0.2223

For General Elections 2011, the linear regression for forecasting the proportion/number of representative polling stations is

$$Proportion = 0.0156 - 1.8343*10^{-5} * size + 0.0881 * var(eu) \qquad R^2 = 0.5961$$

and the actual-forecast parliament seats are in Table 5.

	Voting proportions		Parliamentary Seats	
Party	**Actual**	**Forecast**	**Actual**	***Forecast***
PNV	32.99	34.58	3	3
AMAIUR	19.41	17.82	2	1
PSE	21.63	20.36	2	2

PP	17.93	18.51	1	2

Table 5: Actual-Forecast results, General Elections 2011. 8 Polling Stations. Var(eu): 0.1891

Finally, for Basque Parliament Elections 2012, the linear regression for forecasting the proportion/number of representative polling stations is

$$Proportion = 0.0208 - 2.1787*10^{-5} * size + 0.0876 * var(eu) \qquad R^2 = 0.4907$$

and the actual-forecast parliament seats are in Table 6.

	Voting proportions		**Parliamentary Seats**	
Party	**Actual**	**Forecast**	**Actual**	***Forecast***
PNV	38.42	39.23	11	11
BILDU	21.47	19.63	6	6
PSOE	18.95	18.40	5	5
PP	11.81	12.39	3	3

Table 6: Actual-Forecast results, Basque Parliament 2012. 8 Polling Stations. Var(eu): 0.1898

In all three past linear regressions the number of observations available are 30 and all coefficients are at least 5% significant.

Considering the proportion of representative polling stations needed, this number goes from 7 to 12 for the elections analyzed. For the 2011 election where the province is divided into 4 electoral districts, the number of polling stations per district is around 350, which means a proportion of representative polling stations of around 3%. For the other 2 elections, with unique district of size next to 1500, this proportion is as low as 0.55%. The forecasting accuracy is either exact in terms of parliamentary seats allocation or with an error of just one seat mistakenly assigned to the wrong party.

As a summary, we can say that this method is very reliable in this case. It approximates very precisely the proportion of votes obtained and, in fact, the number of estimated parliamentary seats assigned to each party matches exactly the actual result finally obtained.

Conclusions

In this paper we use the representative polling stations method to forecast the number of parliamentary seats obtained by each political party in a democratic election.

Using the information relative to elections in Bizkaia from 2000 to 2012, it is observed that the proportion of polling stations to be retained for forecasting purposes is not always the same. In particular, it has been observed that there exists an inverse relationship between that proportion and the total size of the electoral district, in such a way that when the electoral district is the whole province, the proportion is much smaller than it is in the case that it is divided into several smaller districts. On the other hand, it is also observed that there is a direct relationship between the degree of variability (or heterogeneity) inside the district and the proportion of stations to be retained. That is, if inside the district there are stations with very different results, more stations are needed to make a good representation of the whole district adequately.

Based on this evidence, we have estimated the relationship between the proportion of stations to be retained and the two factors found behind it, that is, district size and heterogeneity degree. Upon application of this result to the elections subject to analysis, it is observed that the method behaves adequately, at the same time that never appears a difference greater than two parliamentary seats (one by one) between the forecast of seats we obtain and the actual distribution of seats finally obtained. Effectiveness of the method in other electoral systems is yet to be assessed and remains for future work.

Given the effectiveness of the method proposed, we advocate its use in practice along with some way of speeding up the vote counting process just in those stations selected as representative in each electoral

district. This could be, for instance, the use of electronic ballot systems, provided that the drawbacks of these systems can be wiped out (Fernández [4]). If such electronic systems are considered unacceptable we recommend a faster manual counting process for selected stations, perhaps by assigning more personnel to this task. This would not be feasible for the whole electoral territory, or a large proportion of polling stations, but it would be so for the minimal proportions of representative polling stations that this method requires for parliamentary seats prediction.

As a secondary use of this method, we suggest the possibility of using sampling mainly on representative polling stations for prior-to-election polls, perhaps by telephone interviewing, rather than relying solely on sampling techniques on the whole census of voters taking into account demographic and/or geographical considerations. At the same time, it is probably not convenient making public which polling stations are going to be considered as representative, as voters in the corresponding areas might be influenced by this information in a way that could affect their so-much-sought representation.

Acknowledgements

Financial help from the Departamento de Educación, Política Lingüística y Cultura from Gobierno Vasco, Basque University System Research Groups IT-783-13 and IT-567-13 and BETS: Bilbao Economic Theory and Statistics (UFI 11/46), is gratefully acknowledged.

References

1. Bernardo, J.M. (1984), "Monitoring the 1982 Spanish Socialist Victory: A Bayesian Analysis" en *Journal of the American Statistical Association*, vol. 79, pp. 510-515.

2. Bernardo, J.M. (2009), "Probabilidad y Estadística en los Procesos Electorales" en Marrero, M.I., Roche, J. (coord.), *Horizontes Matemáticos*, Universidad de la Laguna, Servicio de Publicaciones, pp. 33-48.

3. Bernardo, J.M., Girón F. J. (1992), "Robust Sequential Prediction from Non-random Samples: the Election Night Forecasting Case" en *Bayesian Statistics*, vol. 4, pp. 61-77.

4. Fernández, R.M. (2001), "El voto electrónico: el caso vasco" en *Revista de Estudios Políticos* (Nueva Época), vol. 112, pp. 199-236.

5. Greenacre, M. J. (2008), *La práctica del análisis de correspondencias*, Ed. Fundación BBVA

6. Lebart, L. (1994), "Complementary use of Correspondence analysis and Cluster Analysis" en Greenacre, M. J. and Blasius, J. (Eds.): *Correspondence Analysis in the Social Sciences*. Academic Press. London, pp.162-178.

7. Martín, A., Fernández, K., García, C. (2008), "Métodos de predicción de escaños en sufragios electorales mediante segmentación y mesas representativas", disponible en http://www.iesa.csic.es/es/congresoencuestas/comunicacionesdefinitivo.pdf pp. 175-183

8. Pavía, J.M., Muñiz, P., Alvarez, J.A. (2001), "Proyecciones en la noche electoral" en *Estadística Española*, vol. 43, pp. 225-239.

Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013)
International Conference, Mataró (Barcelona), Spain 25 - 28 June 2013

Couple-wise divorce rates in Japan

Shigeru Mase[1] and Kouji Yoshimoto[2]

[1] Department of Mathematical and Computer Sciences, Tokyo Institute of Technology, Tokyo, Japan
(E-mail: mase@is.titech.ac.jp)
[2] Department of Mathematical and Computer Sciences, Tokyo Institute of Technology, Tokyo, Japan
(E-mail: yoshimoto.k.aa@m.titech.ac.jp)

Abstract. In Japan, ratios of numbers of divorce to those of marriage per year exceed one-third from 2000 onward. This fact is frequently referred to as if one-third of married couples will divorce in future. This interpretation is doubtful since numbers of marriage has been rapidly decreasing in Japan because of declining birth rates and, on the other hand, numbers of divorce include those of married couples from older generations with higher birth rates. What are true couple-wise divorce rates, that is, rates of couples married in one year who will divorce afterward, in Japan? Estimates of these rates are important because short marriage durations will frequently result in children with insufficient parental cares and financial supports.

In order to estimate couple-wise divorce rates, we have to forecast numbers of divorce in future for couples married each year. This can be done using methods to construct life table. The result shows that actual couple-wise divorce rates will be over one-third for couples married from 1997 onward, contradictory to our first hypotheses that actual couple-wise divorce rates are much less than one-third. Moreover, it will be almost 40% for those married in 2002.
Keywords: couple-wise divorce rates, vital statistics, Japan.

1 Introduction

Divorce as well as marriage is one of important demographic factors. It lessens not only numbers of babies born but also increases children with insufficient parental and financial cares. In this paper, three kinds of divorce rates are discussed and compared, that is,

- mean annual no. of divorces per 1,000 populations,
- ratios of no. of divorces to that of marriages per year,
- ratios that couples married in a year who will divorce afterward.

The first one is called crude divorce rates (CDR) in literature and is used for cross-country comparison. There seem no established terms for the rest and, for convenience, we refer them to as annual divorce rates (ADR) and couple-wise divorce rates (CWDR) respectively in the sequel.

In Japan, ADRs have been larger than one-third since 1997 and this fact is frequently quoted as if one-third of married couples will divorce in future, that is, ADR is confused with CWDR. This interpretation is doubtful. Although annual ratios of death to birth varied from 80.8% to 111.48% from 2000 to 2010 in Japan, "person-wise death rates" should be always 100%. Numbers of

marriage has been rapidly decreasing in Japan because of declining birth rates and, on the other hand, numbers of divorce include those of married couples from older generations with higher birth rates. What are true CWDRs, that is, ratios of couples married in a year who will divorce afterward, in Japan?

In order to study this question, it is necessary to predict numbers of couples who married in a year and will divorce afterward. This resembles construction of life tables and can be treated in a similar way. Main differences are that one may not marry, may marry and divorce several times, and cannot divorce after his death.

In the following, we study the "life table" of divorce in Japan using divorce data of vital statistics of the Ministry of Health, Labour and Welfare, Japan.

2 Dataset

Basic data used is the vital statistics of of the Ministry of Health, Labour and Welfare, Japan[5]. Divorce data is summarized in "Vital Divorce Statistics 2009"[6]. This dataset consists of number of divorced couples classified according to their marriage duration (in year) for 1947–2010. Several missing data was taken from the website[5].

We should remark the followings about this dataset:

- Only years of divorce are given and we cannot recover corresponding wedding years. For example, couples who divorced in 2010 with marriage duration 0 (i.e. less than one year) might marry either in 2009 or 2010.
- Couple's previous marital status are unknown.
- Numbers of couples with marriage durations over 20 years are added up till 1989 except 1970, 1975 and 1980.
- International couples one of which are Japanese are included.
- Years of divorce mean those when couples began to live separately and marriage durations mean periods couples lived together.
- There are relatively few couples marriage durations of which are unknown.

We constructed working data consisting of wedding years and divorce years for all couples in the original data as follows:

- Estimated marriage durations in intervals

$$I : [20, 25), [25, 30), [30, 35), [35, \infty)$$

  for those with divorce years before 1990 so that their proportions are the same as those of (nearest) 1970,1975 and 1980.
- Applied monotone spline fitting[1] to cumulative frequencies of marriage durations and, then, estimated numbers with marriage duration $d = 0, 1, 2, \ldots, 55$ for each divorce year 1947–2010. Here the largest duration $d = 55$ was chosen so that it is long enough and estimated marriage durations became almost decreasing for $d \geq 30$.
- Relatively few data for which marriage duration are unknown were distributed to $d = 0, 1, 2, \ldots, 55$ proportionally.

- For each couple with divorce year n and marriage duration n, their divorce day $e = 1, 2, \ldots, 365$ was chosen uniformly in the year n and, then, corresponding wedding day (the day starting to live together) $m = 1, 2, \ldots, 365$ was chosen uniformly in 365 days ending with e. 27 leap days were taking account of.

Thus we could get data of numbers of divorced couples classified by both wedding years and marriage durations.

3 Probability law of marriage durations

The basis of life table is the assumption that life spans obey probability laws. Before starting analysis, it is worth checking if marriage durations also obey probability laws since marriages and divorces seem more subjective matters than births and deaths and may be influenced heavily by current socio-economic conditions. Above all, one may not marry and, if married, one may not divorce or may die before divorce. The left of Fig. 1 shows relative frequencies of marriage durations of couples who divorced in 2001–2010 and the right is the corresponding probabilities estimated by least square method. It is seen that marriage durations shows almost the same patterns at least for these period. Couples with marriage duration $d = 3$ has the highest probability 0.0797. The probability for $d \leq 5$ and $d \leq 9$ is 0.330 and 0.533 respectively. The median and the mean is 8.26 and 11.0 respectively.

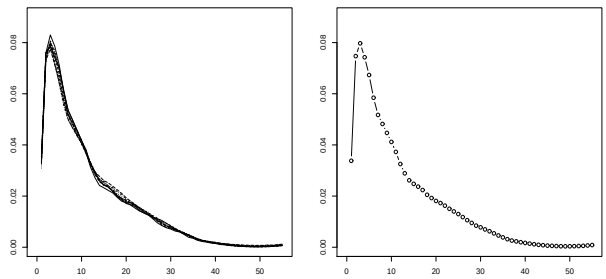


Figure 1. Relative frequencies of durations of marriage of couples who divorced from 2001 to 2010 (left) and estimates of corresponding probabilities by least square method (right)

4 CDRs and ADRs of Japan

The left of Fig. 2 is plots of populations and ADRs of Japan for 1951–2010. The right is plots of CDRs and marriage rates per 1,000 population. Both CDRs and ADRs were calculated by yearly data and show fairly similar variations.

CDRs vary from 1.96 to 2.27 for 2001–2010. According to the international comparison around 2007, Japanese CDR is 26th. Russian and American CDR, which was first and fourth, are about 4 and 5. Korean CDR 2.6 is 9th and the highest in Asia. Chinese CDR 1.6 is 32th. CDRs in Japan took a peak value 1.53 in 1899, became 0.99 in 1920, varied around 1.0 from 1947–1950,

decreased gradually afterward and showed 0.73 in 1963 which was the lowest value after the World War II.

ADRs vary from 0.346 to 0.383 for 2001–2010. This is the reason of the widespread belief that one-third of newly married couples in Japan will divorce afterward.

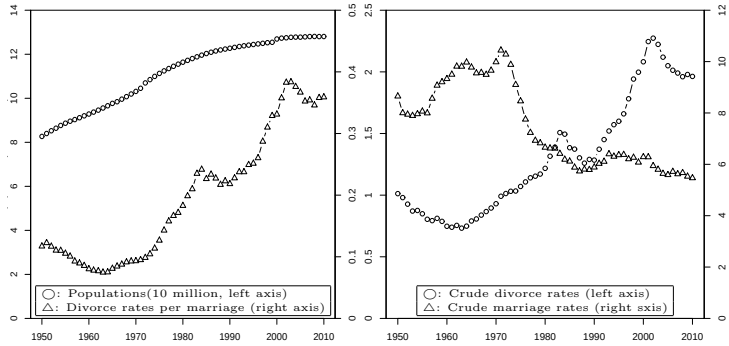


Figure 2. Left: Populations (per 10 million) and ratios of divorces and marriages(%) for 1950–2010, Right: Crude divorce rates and marriage rates (per 1,000) for 1950–2010.

5 Couple-wise divorce rates in Japan

In life table analysis, one is interested in the decrease of a population with the same birth year. In the same way, we can consider a "life table of marriage", that is, the decrease of a population of couples by divorce with the same wedding year. In particular, it is important to guess CWDR, rates that couples who will divorce eventually. Although both life tables are similar in nature, there are several distinctions. Although rates that person dies eventually is 100%, there are no such values for CWDRs decided beforehand. One may divorce several times and there is possibility that CWDRs for first marriage are different for those of remarried couples. One couple may not divorce or may die before divorce.

In life table analysis, Lexis diagram[2] is well-known. Fig. 3 is the Lexis diagram for marriage duration. It shows numbers of couples with x-axis being wedding years and y-axis being marriage durations. In true Lexis diagrams, x-axis is for years of birth and y-axis for ages at death. Numbers with $x = 1$ corresponds to those couples who married and divorced in the same year and, therefore, are smaller than others. The shaded area in Fig. 3 is the area where we have data. There are no available data for the upper-left triangular area. The lower-right triangular area belongs to future and we have to predict their values using existing ones. In order to predict the CWDR for couples married in 2010, it is necessary to predict their numbers of divorce after $d = 2, 3, \ldots, 55$ year.

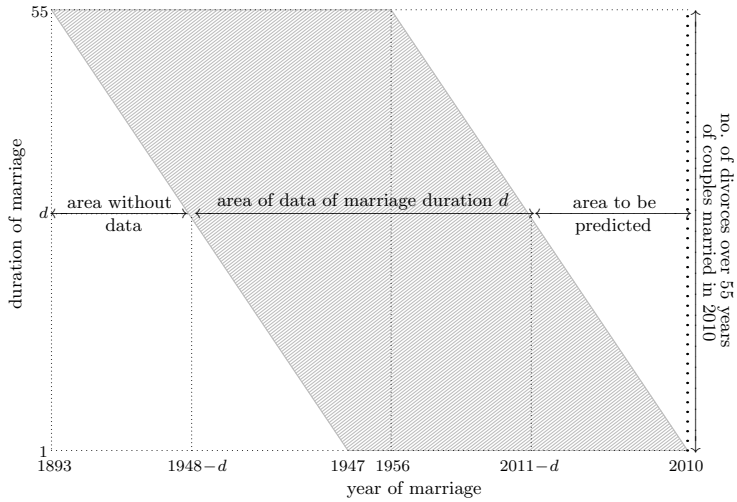


Figure 3. Lexis diagram of marriage and divorce

The procedure to predict CWDRs of couples married in 1947–2010 using spline function fitting is as follows:

1. Fit cubic splines to times series of divorce numbers with length 64
$$X^{(d)}_{1948-d}, X^{(d)}_{1949-d}, \ldots, X^{(d)}_{2011-d}$$
for each marriage duration $d = 2, 3, \ldots, 55$.
2. Predict numbers of divorce after $2012 - d$
$$\hat{X}^{(d)}_{2012-d}, \hat{X}^{(d)}_{2013-d}, \ldots, \hat{X}^{(d)}_{2010}$$
using thus fitted spline functions.
3. Add 55 values of predicted divorce numbers and known divorce numbers before 2010 and get total divorce numbers of couples married in 1947–2010. Dividing them by corresponding marriage numbers, we get CWDR.

Predicted values using cubic splines may vary considerably depending on nodes used. Comparison of predicted values for nodes 25(5)55 and 64 showed almost the same predicted values on average. But individual values may be non-realistic, such as negative values, depending on node numbers. Cubic splines are mainly tools for interpolation but not for prediction. So results by spline prediction seem not reliable as seen in Fig. 4.

A more reliable prediction method for time series X_t is the linear state space (STL) model[4]. This model has the following structure:

$$X_t = M_t + S_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2_\epsilon),$$
$$S_{t+1} = -S_t - S_{t-1} - \ldots - S_{t-s+2} + w_t, \quad w_t \sim N(0, \sigma^2_w)$$

where S_t is a periodic component with period s, M_t is a local trend. ϵ_t and w_y are iid Gaussian errors with respective variances σ^2_ϵ and σ^2_w. σ^2_w may be equal to 0 and there may not exist a periodic component S_t at all.

Table 1 and Fig. 4 shows results of CWDR prediction. Table 2 shows predicted results of numbers of marriage (per 10,000 population), numbers of divorces (per 10,000 population) and CWDRs (%) for 2011–2020 by STL models. CWDRs increases gradually around 40% The variation of CWDRs by STL model is fairly similar to that of ADRs. On the other hand, the variation of CWDRs by spline method differs from them and almost monotonically increases. Table 3 is a summary of how they increase.

6 Conclusion

From Fig. 4, predictions by STL model seem plausible although there may be other prediction methods. Contrary to our first opinion, CWDRs of Japan are larger than 1/3 from 1997. Moreover, it is about 40% for couples married in 2001–2010.

Changes of CWDRs anticipate those of ADRs by a few years and larger than ADRs by 2.38–5.90% since 1966. Fig. 5 shows differences of CWDRs and ADRs. It has three peaks at 1968,1976 and 1997. First two peaks seems to be results of increases of marriages of females and males of the first baby boomers respectively. The differences became smaller gradually in 1879–1993. This seems a result of subsequent increase of divorces and gradual increase of divorces of middle and older age. The third peak seems to be a result of the second marriage boom (marriages of children of the first baby boomers) accelerated by a decrease of younger populations. The differences became again smaller after 2000.

From Fig. 6, we can see a fairly similar pattern of increase of divorce regardless of marriage durations. In particular, a rapid increase from 1990 and a stagnation and/or a decrease from 2001 are common. As Table 2 suggests, this stagnation seems to continue in future. A long-standing effect of baby boomers after the World War II has been almost diluted. It becomes more difficult to divorce as well as marriage after a long recession. Also it may be a result of new social norms on marriages and families. In fact, it has been now common in Japan for youngs to live together without getting married and get married only after having children. Also it has been becoming common not to marry at all.

Table 1. Summary of divorce rates (1947–2010)

	min	median	max	mean
CWDR(LST)	9.22% (1961)	22.0% (1980)	40.1% (2002)	22.2%
CWDR(spline)	11.0% (1958)	21.7% (1980)	38.7% (2003)	24.2%
ADR	7.54% (1963)	18.95% (1980)	37.5% (2003)	19.6%
CDR	0.732 (1963)	1.22 (1980)	2.27 (2002)	1.31

Table 2. Predictions by STL model of number of marriages and divorces and corresponding CWDR for 2010–2020

	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
marriage	69.1	68.3	67.4	66.5	65.6	64.7	63.8	63.0	62.1	61.2
divorce	26.2	26.1	25.9	25.8	25.7	25.6	25.5	25.4	25.2	25.1
CWDR	37.9	38.2	38.5	38.8	39.2	39.5	39.9	40.3	40.7	41.0

Table 3. Summary of how three divorce rates increase for 1951–2010

	CWDR by spline	CWDR by STL	ADR
$\geq 20\%$	1976∼	1978∼	1982∼
$\geq 25\%$	1985∼	1983∼(except 1988)	1993∼
$\geq 30\%$	1988∼	1996∼	1998∼
$\geq 1/3$	1994∼	1997∼	2000∼
$\geq 35\%$	1996∼(except 1999,2000)	1998∼	2001∼
$\geq 40\%$	none	2002	none

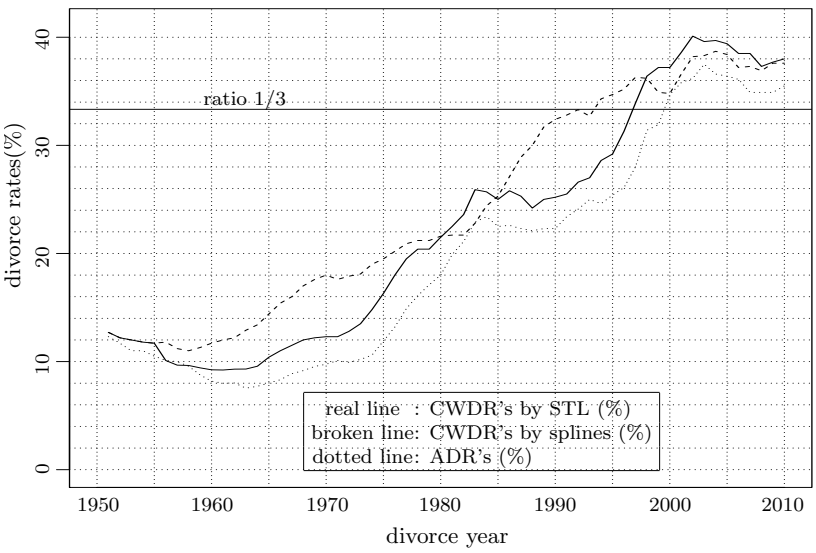


Figure 4. Plots of CWDRs by STL and spline models and ADRs for couples married in 1951–2010

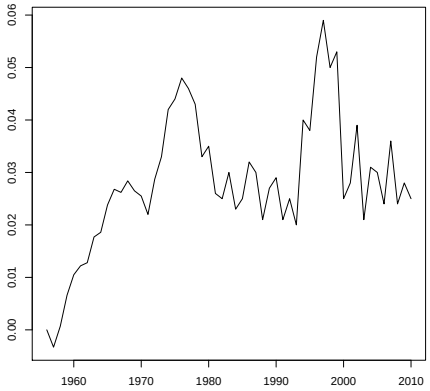


Figure 5. Plots of differences of CWDRs and ADRs for 1956–2010

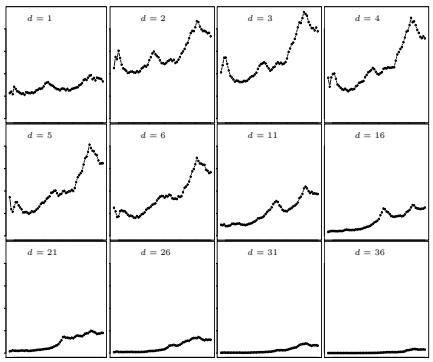


Figure 6. Plots of numbers of divorces with marriage duration d for divorce years 1947–2010

References

1. F. N. Fritsch and R. E. Carlson. Monotone piecewise cubic interpolation, *SIAM Journal on Numerical Analysis*, 17, 238–246, 1980.
2. H. U. Gerber. *Life Insurance Mathematics*, 3rd ed., Springer-Verlag, Berlin Heidelberg, 1997.
3. P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall, 1994.
4. A. C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge, 1989.
5. The Statistics Bureau and the Director-General for Policy Planning of Japan Statistics and Information Department Minister's Secretariat, *e-Stat*, The Ministry of Internal Affairs and Communications, Japan, 2012.
`http://www.stat.go.jp/english/index.htm`
6. Health and Welfare Statistics Association. *Vital Divorce Statistics 2009* (in Japanese), the Ministry of Health, Labour and Welfare, Japan, 2010.
`http://www.hws-kyokai.or.jp/english-index.htm`

Data analysis and reliability models of pumping stations of the haoud Elhamra-Bejaia pipeline (Algeria)

[1*]Medjoudj Rafik, Imine Aghiles[2], [1]Medjoudj Rabah, [2]Aissani Djamil
[1]Electrical Engineering Department, University of Bejaia, Algeria
[2]LAMOS Laboratory, University of Bejaia, Algeria

Abstract, in this paper, a case study for reliability data analysis of pumping stations of a pipeline is introduced. By modeling the times to failure and the mean times between failures of components at these stations, the Weibull distributions were obtained on the basis of Kolmogorov Smirnov tests. When analyzing the shape parameters of the obtained distributions, it was observed that their values are less than unity, indicating that, theoretically, the considered components are in a youthful period or that the components have been operating for about 25 years. These statements are contradictory with practical considerations (e.g., an aging period) and the shape parameters values should be greater than the unity. Regarding also the multiple failure causes of the considered components, the adjustment of data using mixed Weibull distributions is deemed adequate by the tests and the shape parameters values are globally greater than the unity and the trend confirms the nature of aging equipment. The reliability evaluation of a station's components and the reliability change for different variants of the pipeline have been ascertained.

Keyword: Pumping station, Reliability, Weibull distribution.

1 Introduction

Reparable electrical components were usually modeled using two state (working-failure) diagrams. Nowadays, this assumption is relaxed in favor of multi-degraded items. In reference [1], the authors stated in their review of a Markov process that deterioration of the equipment is modeled as occurring in a limited number of steps. It is assumed that the duration of each stage of deterioration as well as times for repairing failed components is exponentially distributed. During the last decade, the Weibull-Markov approach has been applied to electrical components and systems interruption modeling and maintenance applications. Introduced by Van Castaren [2, 3] for state probabilities, transition probabilities and frequencies assessment and cost evaluation, it was reproduced by Pivatolo [4] for system generation reliability evaluation. Recently we have compared both the Markov method and the Weibull-Markov approach in interruption modeling of electrical distribution systems. It has been proved that the results obtained using this latter were more realistic [5, 6]. This paper develops reliability models of the components constituting the different pumping stations of the pipeline Haoud Elhemra-Bejaia and finds reliabilities of three different configurations of the same stations. As for the statistical treatments, two modeling methods are used and compared: the first uses simple distributions, however the second introduces the mixed distributions.

2 Description of the system

2.1 Haoud Elhamra Bejaia pipeline

Sonatrach (National Society of Transportation and Marketing of Hydrocarbons) was born December 31, 1963. It was considered to be the instrument of state intervention in the oil sector, alongside French companies. Until 1966, it was confined to the mission of the transport and marketing. But then it became 'National Society for the Research, Production, Transportation, Processing and Marketing of Hydrocarbons'. Currently, it is the first company in Africa, It is ranked 12th among global oil companies. It is the second explorer of LNG and LPG and the third largest explorer of natural gas in the world. Its global production (all products) was estimated to 230 million tons in 2006. Our study concerns one part of the oil transportation by pipeline in Algeria, the line HEH (Houdh El Hamra in the sahara) – Bejaia (petroleum port). The maximum flow is estimated to 2540 m^3/hr and the actual transit between 700m^3/hr and 1800m^3/hr. Depending on the demand and the use of principal stations, the flow can reach 2000m^3/hr. The pipeline ″ OB1 24 / 22 ″ is the first structure transport via pipeline installed in Algeria, by the oil company managers of SOPEG in 1959. It connects the storage facility (HEH) and marine terminal (Bejaia). The pipeline diameter OB1 24 `` (609.60 mm) and HEH-Bejaia has a total length of 660,721 km and it is composed of two sections: The diameter of the first section line between HEH and neck Selatna is “24″” (609.60 mm) for the length of 533,217 km. The second one, with diameter 22″ (558.8 mm), is connecting the neck Selatna and marine terminal on a length of 127,551 km. HEH-Bejaia line, consists of : A departure terminal HEH-SP1; three (3) main pumping stations SP1 bis, SP2 and SP3; four (4) intermediate pump stations (satellites) SPA, SPB, SPC and SPD; terminal arrival Bejaia; an oil port. The station SPA (Touggourt, ALT = 89.5, PK = 135129, the station SPB (Oumach, ALT = 83, PK = 189.83), the station TPS (Oued El Fida, ALT = 411.75, PK = 443.083) and the station SPD (Beni-Mansour, ALT = 356, PK =573.62) each have a turbo-pump and auxiliaries consisting of a combustion turbine “THMs” and “Guinard” Pump.

2.2 Operation mode of the pipeline

To perform a reliability analysis, it is necessary to know the functional and technical specifications of the system components. The line can operate in three modes.

Each mode is characterized by the speed of operation and the elements to start. The operation mode is represented as a variant of the system. The following table (1) represents the different operation modes of the system.

Mode	Flow (m^3/h)	Tons (MTA)*	Number of groups per station
1st Mode	700 to 800	5.52	1EPS/SP1, 1MPS/SP2, 1 MPS /SP3
2sd Mode	1400	9.65	2 EPS/SP1, 2 MPS/SP2, 2MPS/SP3
3th Mode	1700 to 1800	12.41	3 EPS/SP1,1 TP(TP1 or TP2) / SP1bis, 3 MPS/SP2,3 MPS/SP3

Table 1. Operation modes of the pipeline

Starting with an annual quota set by the marketing department, we determine the hourly flow transported by the pipeline. It will be calculated based on the density of the product to transit. To attain 16 MT / year of crude oil, its density must be 779 Kg/m^3. We will have an hourly capacity of:

$$\frac{6.10^9 / 365 \times 24 \quad \text{Kg / h}}{779 \quad \text{Kg / m}^3} = 2354 \quad \text{m}^3 / h$$

To determine the starting elements in the system, in order to attain that flow along the line from the departure terminal SP1, it must take into account the speed provided by the elements of line, and the pressure provided by each station.

3. Reliability of the each variant of the system

The objective of this analysis is to reduce the overall system to a set of functional variants of the system. From a reliability point of view, each variant is composed of a set of elements, where the failure of each of them causes the failure of the variant (i.e., series configurations).

R_{TP} (t) : Reliability of the turbo-pump of type THM (satellite stations and TPC)
R_{MPS} (t) : Reliability of motor Pump set
R_{EPS} (t) : Reliability of electro-pump set
R_{vari} (t) : Reliability of variant i

Variant 1:
This variant represents the operation of stations working with an hourly capacity of 800 m3 / h. In this case, each station of the principal stations operates with one equipment only among the five. One Electro Pump Set in SP1 and 1 motor-pump set in two other stations (SP2, SP3) will operate. So this variant represents a series configuration of all stations and each station represents the configuration of (1/5). Its reliability is presented as follows:

R_{var1} (t) = R_{SP1}(t) × R_{SP2}(t) × R_{sp3}(t)

R_{SP1} (t) = $C_5^1 R_{EPS}(t) \times [1-R_{EPS}(t)]^4 \times C_5^2 R^2_{EPS}(t) \times [1-R_{EPS}(t)]^3 + C_5^3 R^3_{EPS}(t) \times [1-R_{EPS}(t)]^2 + C_5^4 R^4_{EPS}(t) \times [1-R_{EPS}(t)] + C_5^5 R^5_{EPS}(t)$.

$R_{SP2}(t) = C_5^1 R_{MPS2}(t) \times [1-R_{MPS2}(t)]^4 \times C_5^2 R^2_{MPS2}(t) \times [1-R_{MPS2}(t)]^3 + C_5^3 R^3_{MPS2}(t) \times [1-R_{MPS2}(t)]^2 + C_5^4 R^4_{MPS2}(t) \times [1-R_{MPS2}(t)] + C_5^5 R^5_{MPS2}(t)$.

$R_{SP3}(t) = C_5^1 R_{MPS3}(t) \times [1-R_{MPS3}(t)]^4 \times C_5^2 R^2_{MPS3}(t) \times [1-R_{MPS3}(t)]^3 + C_5^3 R^3_{MPS3}(t) \times [1-R_{MPS3}(t)]^2 + C_5^4 R^4_{MPS3}(t) \times [1-R_{MPS3}(t)] + C_5^5 R^5_{MPS3}(t)$.

Variant 2 :

This configuration allow to the system to bring the flow to 1400 m^3/h. The SP1 requires 3 EPS and 2 MPS for each of the two other stations. The SP1 station represents the 3/5 configuration. The SP2 and SP3 each represent the 2/5 configuration. The reliability is given by:

$R_{var2}(t) = R_{SP1}(t) \times R_{SP2}(t) \times R_{SP3}(t)$

$R_{SP1}(t) = C_5^3 R^3_{EPS}(t) \times [1-R_{EPS}(t)]^2 + C_5^4 R^4_{EPS}(t) \times [1-R_{EPS}(t)] + C_5^5 R^5_{EPS}(t)$.

$R_{SP2}(t) = C_5^2 R^2_{MPS2}(t) \times [1-R_{MPS2}(t)]^3 + C_5^3 R^3_{MPS2}(t) \times [1-R_{MPS2}(t)]^2 + C_5^4 R^4_{MPS2}(t) \times [1-R_{MPS2}(t)] + C_5^5 R^5_{MPS2}(t)$.

$R_{SP3}(t) = C_5^2 R^2_{MPS3}(t) \times [1-R_{MPS3}(t)]^3 + C_5^3 R^3_{MPS3}(t) \times [1-R_{MPS3}(t)]^2 + C_5^4 R^4_{MPS3}(t) \times [1-R_{MPS3}(t)] + C_5^5 R^5_{MPS3}(t)$.

Variant 3:

In this variant, the SP1 station operates with three EPS. For the stations SP2 and SP3 two MPS are planned, and one TP for the two other stations, namely SP1bis and SPB.

$R_{var2}(t) = R_{SP1}(t) \times R_{SP2}(t) \times R_{SP1bis}(t) \times R_{SP3}(t) \times R_{SPB}(t)$

$R_{SP1}(t) = C_5^3 R^3_{EPS}(t) \times [1-R_{EPS}(t)]^2 + C_5^4 R^4_{EPS}(t) \times [1-R_{EPS}(t)] + C_5^5 R^5_{EPS}(t)$.

$R_{SP2}(t) = C_5^3 R^3_{MPS2}(t) \times [1-R_{MPS2}(t)]^2 + C_5^4 R^4_{MPS2}(t) \times [1-R_{MPS2}(t)] + C_5^5 R^5_{MPS2}(t)$.

$R_{SP3}(t) = C_5^3 R^3_{MPS3}(t) \times [1-R_{MPS3}(t)]^2 + C_5^4 R^4_{MPS3}(t) \times [1-R_{MPS3}(t)] + C_5^5 R^5_{MPS3}(t)$.

$R_{SP1bis}(t) = R^3_{TP1}(t)$

$R_{SP1B}(t) = R^3_{TP2}(t) \times [1-R_{MPS3}(t)]^2 + C_5^4 R^4_{MPS3}(t) \times [1-R_{MPS3}(t)] + C_5^5 R^5_{MPS3}(t)$.

4. Data analysis

The data collected from the rural industry was a very important step in our work. These data concerns the different mean times of the components of the

different stations, and the period is from 1/01/1997 to 12/31/2007 or a period of 11 years.

4.1 Adjustment tests

The following table shows the results of adjustment tests of the mean lifetimes of the different components by using the Weibull distribution.

Variables		n	Adjusted law	Parameters		Test of K-S		Validation test
						D_n	$d_{n,0.05}$	
X1	SP1 Booster	3	Weibull	β=1.1139846	η=9194.127654	0.171	0.172	Not rejected
			Exponential	λ=0:0001192653		0.332	0.172	Not rejected
X2	SP1 GEP	3	Weibull	β=5.348812	η=2600.770844	0.076	0.172	Not rejected
			Exponential	λ=0.0004214963		0.202	0.172	Not rejected
X3	SP1bis TP	77	Weibull	β=0.64560084	η=612.0831252	0.103	0.172	Not rejected
			Exponential	λ=0.0011918398		0.122	0.172	Not rejected
X4	SP2 Booster	6	Weibull	β=0.8119306	η=7762.198197	0.106	0.270	Not rejected
			Exponential	λ=0.0001252087		0.119	0.270	Not rejected
X5	SP2 MPS	29	Weibull	β=0.6284866	η=1266.52	0.107	0.309	Not rejected
			Exponential	λ=0.0005792701		0.456	0.309	Not rejected
X6	SP3 Booster	8	Weibull	β=0.6313168	η=4401.447566	0.107	0.328	Not rejected
			Exponential	λ=0.000149379		0.136	0.328	Not rejected
X7	SP3 MPS	13	Weibull	β=1.3268163	η=4212.132403	0.219	0.294	Not rejected
			Exponential	λ=0.0002756515		0.477	0.294	Not rejected
X8	SPB TP	114	Weibull	β=0.58311194	η=276.9759413	0.219	0.294	Not rejected
			Exponential	λ=00022518519		0.477	0.294	Rejected

Table 2. Results of adjusted test of mean times between failures using parametric laws

After analyzing the results obtained by the adjustment tests, it was observed that their values are less than unity indicating that, theoretically, the considered components are in a youthful period, which is contradictory with practical considerations (aging period). For this reasons we have proceeded to an adjustment test using the mixed Weibull function as the distribution function of the mean lifetimes of the components.

Variables		n	Adjusted law	Parameters		Test of K-S		Validation test
						D_n	$d_{n,0.05}$	
X1	SP1 Booster	3	-	-	-	-	-	-
X2	SP1 GEP	3	-	-	-	-	-	-
X3	SP1bis TP	77	Mixed Weibull	β_1= 1.007182 β_2= 1.141648 P_2= 0.4010147	η_1= 96.94063 η_2= 1566.4557 P_2= 0.5989853	0.134	0.148	Not rejected
X4	SP2 Booster	6	-	-	-	-	-	-
X5	SP2 MPS	29	Mixed Weibull	β_1= 1.1562567 β_2= 0.9798344 P_2= 0.5006783	η_1= 342.6612 η_2= 4113.4872 P_2= 0.4993217	0.149	0.24	Not rejected
X6	SP3 Booster	8		-	-	-	-	-
X7	SP3 MPS	13	Mixed Weibull	β_1=5.36067 β_2=2.553703 P_2=0.381716	η_1= 669.1127 η_2=5150.0732 P_2= 0.618284	0.231	0.361	Not rejected
X8	SPB TP	114	Mixed Weibull	β_1=1.020364 β_2=1.224585 P_2=0.5612362	η_1= 71.5822 η_2= 1037.4013 P_2= 0.4387638	0.084	0.121	Not rejected

Table 3. Results of adjusted test of mean times between failures using mixed Weibull distribution

5. Reliability study of the system

5.1 Turbo-pumps and motor-pumps reliabilities

The figures (1) represents the changing reliabilities of different components of the pumping stations. Their expressions are given as follows:

$$R_{TP\,1bis} = 0.4010147 \quad \exp\left[-\left(\frac{t}{16.94063}\right)^{1.007182}\right] + 0.5989853 \quad \exp\left[-\left(\frac{t}{1566.45570}\right)^{1.141048}\right]$$

$$R_{MPS\,2} = 0.5006783 \quad \exp\left[-\left(\frac{t}{342.6612}\right)^{1.1562567}\right] + 0.4993217 \quad \exp\left[-\left(\frac{t}{4113.4872}\right)^{0.9798344}\right]$$

$$R_{MPS\,3} = 0.381716 \quad \exp\left[-\left(\frac{t}{669.1127}\right)^{5.360671}\right] + 0.6182845 \quad \exp\left[-\left(\frac{t}{5150.07322}\right)^{2.553703}\right]$$

$$R_{TPSPB} = 0.5612362 \quad \exp\left[-\left(\frac{t}{71.5822}\right)^{1.020364}\right] + 0.4387638 \quad \exp\left[-\left(\frac{t}{1037.4013}\right)^{1.224585}\right]$$

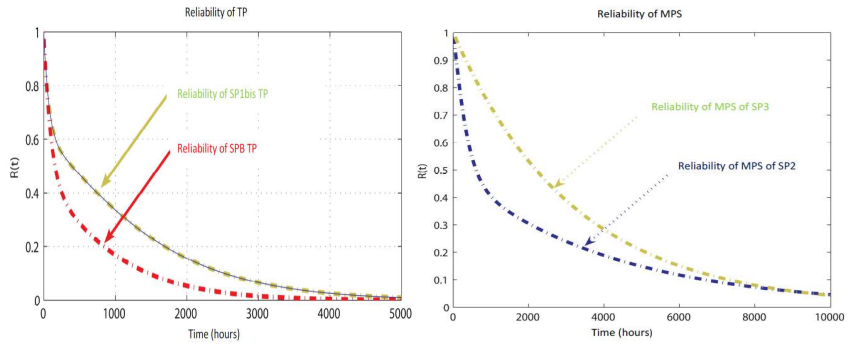


Figure1: Change in reliability of turbo-pumps and motor-pumps

The figure (1) shows respectively the reliabilities of turbo-pumps and the motor-pumps installed in the pumping stations. We remark that the and the motor-pumps are more reliable than the turbo-pumps, this because the stations contain 5 motor-pump sets installed in parallel, While it contains only one turbo-pump in the stations SP1bis and SPB.

. **5.1 Reliability of different variants of the pipeline**

The figure (3) shows the reliabilities of the different variants of the pipeline.

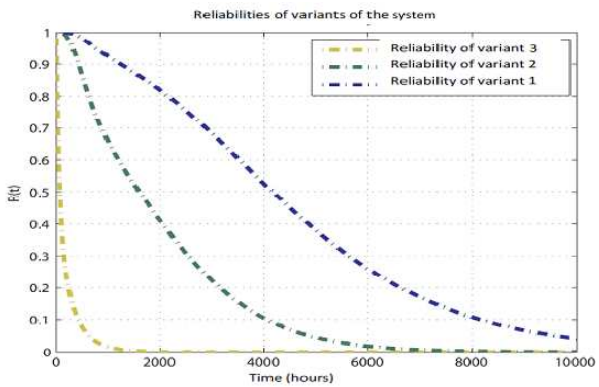


Figure1: Reliabilities of three variants of the system

We remark that the first variant is more reliable, these because it operates by only one component in each station and the four others are in standby.

Conclusion

Pumping station components are an important part of the oil transportation network, and they are often subject to failures due to environment and exploitation conditions. In this work, three variants of system reliability were given taking into account the state of the components of the pumping stations. A real case study of an oil pipeline was considered to explain the strategy used variants of the system, and we observe that the reliability of variant three is the most important, based on the statistical treatments of data collected in a rural industry. The results were presented and discussed, and the reliability of the components were determined under real functioning conditions.

References

1.G.K Chan, Asgarpoor, S. Optimum maintenance policy with Markov process. *Electric Power Systems Research*, 76, 452-456, 2006.

2.J Van Castaren,. Reliability assessment in electrical power systems: the Weibull-Markov stochastic model. *IEEE Trans on Industry Applications*, 36(3), 911-915, 2000.

3.J Van Castaren,. Assessment of interruption costs in electric power systems using Weibull- Markov model. *PhD Thesis, Department of electric power engineering: Chalmers University of technology, Sweden, 2003*.

4.A Pievatolo,. The downtime distribution after a failure of a system with multistate independent components. *IMATI-CNR Technical report, Presented at the Mathematical Methods on Reliability conference, MMR*, Glasgow, 2007.

5.R Medjoudj,., Aissani, D., Boubakeur , A and Haim , K.D., Interruption modelling in electrical power distribution systems using the Weibull-Markov Model. *Journal of Risk and Reliability*, DOI:10.1243/1748006XJRR215, 2009.

6.S-H Jin, Lu, J-G., Wan, Y-P and Sun, S-G., Application of the Bayes' theory to the failure rate evaluation of electrical apparatus. *Journal of Zhejiang University Science A*8(3): 501-505, 2007,.

Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013)
International Conference, Mataró (Barcelona), Spain 25 - 28 June 2013

MDG and 1000 Day Window: Health and Mortality Disparities in India

Barun Kumar Mukhopadhyay

Retired Scientist, Population Studies Unit, Indian Statistical Institute
Kolkata, India
Email: barun_mukhopadhyay@yahoo.com

Abstract: Achieving universal access to reproductive health by 2015 is one of the two targets of *Goal 5 - Improving Maternal Health* - of the eight Millennium Development Goals (UN, 2000). The scientists, economists and health experts of the International Food Policy Research Institute (IFPRI), USA (2012) are of the opinion and agree that improving nutrition during the critical 1,000 day window is one of the best investments to achieve lasting progress in global health and development. Considering the *Goal 5* of MDG and objective of IFPRI, an attempt is made in the paper to investigate situation prevailing in India in relation to health of mother, infant and children under 2 years of age. At the outset it must be mentioned that India's concern is much voiced in international arena perhaps because of her overwhelming population of more than 1.2 billion mark (1,210,193,422 in Census, 2011). And India is likely to surpass China in population by 2030 (Rosenberg, 2012). IFPRI (2012) made some comment about India that 1998 to 2005, India's gross domestic product (GDP) grew by 40 per cent, yet the number of children who were stunted (generally defined as being significantly below the median height for their age) declined only from 51 to 44.9 per cent, and those underweight from 42.7 to 40.4 per cent. It is very unlikely that India having boast of the largest democracy and second largest food producing country in the world still are struggling to overcome the curse of ill health of mother and children, in particular, more about gender bias. It is well known that there are countries in sub-saharan regions, in particular and other under developed region in the world are still clutched with antecedent of health hazards, hunger etc. Contrary, India having glimpses of her past and present history in many fields but lag behind health, mortality, literacy, poverty and many others simply because of her heterogeneous character of people which is found less in general in other countries in the world. In this juncture, the paper tries to analyze Indian situation from different sources of data. The findings show predicaments from different major states, in general but boast of having good condition in other smaller states as well.
Keywords: MDG, Empowered Action Group, Index of Dissimilarity

1 Introduction

Achieving universal access to reproductive health by 2015 is one of the two targets of *Goal 5 - Improving Maternal Health* - of the eight Millennium Development Goals (UN, 2000). And the scientists, economists and health experts of the International Food Policy Research Institute (IFPRI), USA (2012) are of the opinion and agree that improving nutrition during the critical 1,000 day window is one of the best investments to achieve lasting progress in global health and development. IFPRI remarked Indian green revolution at the same time economic development without development being activated in health of child and mother. Considering the *Goal 5* of MDG and objective of IFPRI, a joint analysis of the two schemes is made in the paper to investigate situation prevailing in India in relation to death of mother, infant and children under 2 years of age for the major states in India.

A perpetual attempt is made in India to reduce the maternal mortality. WHO (Network), 199,000 maternal deaths in South Asia, nearly 74 per cent would be accounted for by India. This amounts to about 140,000 maternal deaths. India is a country with diversity. But this diversity is not the "unity in diversity" rather a much problematic situation by which India suffers. There are states with much development whereas majority of big states without development. As a result overall India is defamed in many country around the world. Because of heterogeneity of the people of India overall estimates of MMR remain as high as around 300 in 2003 and reduced to 212 during 2007- 2009. In so far as the data are concerned, experts in the RG office of the GOI in association with many other scientists realised the grim situation prevailing in India which must be curbed. How long the process of state level stratification will continue without people's stratification. As data of later kind is lacking, the present paper likewise attempts to analyse state wise situation in order to know the different situation prevailing in the different major states in India so that action may be done to improve by inputting different ingredients.

The critical 1,000 day window envisages mother health as well as infants, in particular and children under 2 years of age. The situation prevailing about mother has been explained, to some extent above. But the level of infant mortality rates in India is also very high as compared to neighbouring countries. The differentials in the statuses of infancy period is subjected to their morbidity and mortality at the cost of people's lack of awareness, education, mental set up nutritional food, prophylaxis against nutritional anemia so on and so forth. These are all responsible for in some way or other to have raised the infant deaths in India. Moreover gender disparity right from the embrayo stage at the mother's womb to infancy to adulthood is widely accepted fact as far as India is concerned. It has long been observed that India's population shows an unusually masculine composition. This was generally attributed to higher mortality among females than males in contrast to what has been observed in the western world. However, since the 1990s, the young ages show greater masculinity than in the past and a steep rise in the sex ratio at birth has been observed. There is evidence that sex-selective abortions are widely practiced, facilitated by easy availability of the technology of pre- natal sex-detection and access to medical termination of pregnancy, at least in some parts of the country (Kulkarni, 2007).

2 Sources of Data

i) Special Bulletin on Maternal Mortality in India, 2007-2009, Office of Registrar General, New Delhi-110 066.

ii) Statistical Outline of India, 2007-2008, TATA Services Limited, Department of Economics and Statistics, Bombay House, Mumbai-400 001.

iii) DLHS 3, District level household and facility survey, 2007-2008, International Institute for Population Sciences (IIPS), Mumbai-400 004.

iv) Statistical Abstract of India 2007, Central Statistical Organisation (CSO), Ministry of Statistics and Programme Implementation, New Delhi-110 066.

v) Census of India for different years

vi) Sample Registration System for different years and

vii) NFHS at different time point of Government of India.

3 Methodology

The 1000 Day window is: 273 days of gestation period, 365 days of infancy and again 365 days of children under two years of age above infancy. The first period of gestation signifies the health statuses of mother as well as baby in the womb, the second one the infancy which is quite related to very common vocabulary "infant mortality rate" and finally children under two years of age above infancy. The study is done as par the above criteria. The gestation period has been taken up in terms of differentials of maternal mortality ratios among the states. MMR may give some idea of lack of intake of nutrient food, breast feeding and medicine and so on during ante natal and post natal periods by the population of the different states. Moreover, second category may be studied through infant mortality rates, gender bias among the infants which is very common among people of some of the states. The gender differential is measured through boys/girls ratio of each subject under study. An Index of Dissimilarity (ID) is developed to examine the differentials among population. The formula for the gender differences was originally established by Mukhopadhyay and Majumdar (2012) is as follows:

$$\text{Index of Dissimilarity} = \frac{1}{n} = \sum | R_x - 1 |$$

where R_x is the ratio of male by female, x is for different aspect, n the number of observation. As the Institute is of the opinion that if the 1000 days are safe, much of the problem of malnutrition, ill health could be curbed if importance is put on these days of population. What they actually said is important here to quote, "The 1,000 days between a woman's pregnancy and her child's 2nd birthday offer a unique window of opportunity to shape healthier and more prosperous futures. The right nutrition during this 1,000 day window can have a profound impact on a child's ability to grow, learn, and rise out of poverty**.** It can also shape a society's long-term health, stability and prosperity"**.**

4 Analysis of the data

At the initial stage of the analysis, a table may be given for overall India about the level and trend in MMR in order to cope up of the MDG program of the UN. Later, the analysis may be done for the level of MMR in respect of the major states in India. There are other states (small) and some Union Territories (UTs) which are not considered here. From the major states much of the finding are usually obtained.

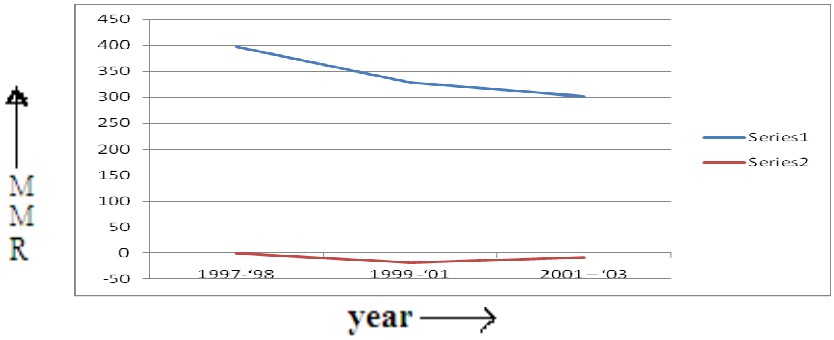
5 Overall level and trend of Maternal Mortality Rate in India

The latest figure of MMR is 212 of India during 2007-2009 (RG, 2011). To explain a few words (Table 1 given below) from the declining values of maternal mortality ratios from 398 in 1997-1998 to 301 in 2001-2003, i.e., during the period of seven years is of the order of around 25 per cent. An earlier study

Table 1 Maternal mortality ratios and per cent change in India , 1997-2003

Year	Maternal mortality ratios	Per cent changes in MMR
1997-'98	398	-
1999 - '01	327	-17.84
2001 — '03	301	-8.00

estimated a declining trend from around 1355 maternal deaths per every 100 thousand live births during 1957-'60 to about 330 maternal deaths, in the year 2002 (Ranjan and Gulati, 2008). In NFHS studies maternal mortality rates had not decline during nineties, rather increased from 424 in 1992-'93 (NFHS-1) (95 per cent CI: 324-524) to 540 in 1998-'99 (NFHS 2) (95 per cent CI: 428- 653). The large confidence intervals were due to smaller number of maternal deaths in these surveys. In this respect a downward linear trend was found during 1966 to 1992 in rural areas of India (Mukhopadhyay, 1996) from estimated figures of 676 in 1966 to 352 in 1992. The linear trend was tested through best fit of the data ($p<0.00$).



Series 1: Absolute values; Series 2: Percentage change

Fig.1. MMR values with trend

6 State MMRs and Trends

Table 2 Maternal mortality ratios of the major states in India and rate of change, 1997-'03 03

Major States	Maternal Mortality Ratio			Rate of Change (percent)	
	1997-98	1999-2001	2001-2003	(1997-98)-(1999-2001)	(1999-2001)-(2001-2003)
Andhra Pradesh	197	220	195	-11.68	-11.36
Assam	568	398	490	-29.93	+23.12
Bihar/Jharkhand	531	400	371	-24.67	-7.25
Gujarat	46	202	172	_	-14.85
Haryana	136	176	162	+29.41	-7.95
Karnataka	245	266	228	+8.57	-14.29
Kerala	150	149	110	-0.67	-26.17
Madhya Pd/Chhattisgarh	441	407	379	-7.71	-6.87
Maharashtra	166	169	149	+1.81	-11.83
Orissa	346	424	358	+24.54	-15.57
Punjab	280	177	178	-36.79	+0.56
Rajasthan	508	501	445	-1.38	-11.18
Tamil Nadu	131	167	134	+27.48	-19.76
UttarPd./Uttaranchal	606	539	517	-11.06	-4.08
West Bengal	303	218	194	-28.05	-11.01
India	398	327	301	-17.84	-7.95

A perusal of the above table reveals a large extent of differences in the level of MMR among the states in 1997-1998 at the initial point of time, i.e., 1997-1998. States like Tamil Nadu ,Haryana and Kerala depicted lower level of MMR being respectively 131, 136 and 150 in 1997-'98. On the other hand, states like Uttar Pradesh/Uttaranchal, Assam, Bihar/Jharkhand and Rajasthan possessed the level at much higher values at the rate higher than 500 mark. Madhya Pradesh/Chhattisgarh depicts slight lower level of 441 which is itself still much high. The levels of the remaining states lie in between these two categories of states. The figure for Haryana being 46 is a doubtful one. Here the contribution of the states in their number of population in overall India seems very high for Uttar Pradesh/Uttaranchal, Bihar/Jharkhand, Rajasthan (Fig.1 as shown in the **Appendix**).

Now the trend of the values during the seven year periods into two periods of 1997-2001 and 2001-2003 clearly show some trend of declining values except in some states where upward trend was observed. As the period of study is very short, hence no definite conclusion can be made so far as trend analysis is concerned. Due to paucity of data, that cannot be done in a proper manner. However, apart from all predicaments, it may be said that a trend of decline at one hand is observed among all the states and on the other the variations are found among the states. As such any developmental programme, if it is done this must go through the in depth analysis before actual application. Otherwise only on the basis of overall Indian values, it would be a waste either of energy or of money etc., as far as national policy is concerned.

After analyzing the data on MMR with respect to MDG, it is now necessary to analyse the data of infant mortality rates at an initial stage giving parity with the 1000 Day Window.

7 Infant Mortality Rates in India with Trend and Gender Bias

The male-female infant mortality rates from 1985 to 2000 are given in Table 3. The values of IMF are declining in both the cases of sexes more or less over the entire period from 1985 to 2000. The CIA World Fact book (Network system, 2011) shows, male: 46.18 deaths/1,000 live births and female: 49.14 deaths/1,000 live births in India, 2011. From these figures the value of DC in 2011 is found to be 0.940 which is still not favouring females. From the table the ratios are found higher than unity in some stray cases where female mortality is less, otherwise in most years the values indicate higher infant mortality rates among females than males

8 Infant Mortality Rates in the States

The table given below shows the level of infant mortality rates of the 15 major states of India. The differentials in the statuses of infancy period is subjected to their morbidity and mortality at the cost of people's lack of awareness, education, mental set up nutritional food, prophylaxis against nutritional anemia so on and so forth. These are all responsible for in some way or other to have raised the infant deaths for which the values of which are put in the table given below for those states in India in 2001.Moreover, these differentials are expected to be similar still after one decade, that is in 2012. Again, the state, Kerala possessed the minimum value of only 11 infant deaths meaning thereby good health of infants, to a great extent, with less malnutrition, good ante natal and post natal cares to have been taken and many other measures rightly and in time seriously have been adopted by the people with their advanced mental set up and awareness, in particular and state as well. The children in that state cannot be said to have been stunted as being significantly below the median height for their age and underweight. Odisha's figure is of the order of 90 may be taken as an exceptional level. Whereas big states like Madhya Pradesh (86), Uttar Pradesh (82), Rajasthan (79) and Assam (73) have very high infant mortality rates, the other states like Andhra Pradesh (66), Haryana (65), Bihar (62) possess moderately high values than the states, Maharashtra (45), Tamil Nadu (49), West Bengal and Punjab (51 each) and Karnataka (58). A picture is given below about these values of these states and India as well.

Table 3 Statuses of infant mortality rates in the major states, 2001

Major states	IMR	Major states	IMR	Major states	IMR	Major states	IMR
Andhra Pd.	66	Haryana	65	**Maharashtra**	45	**Tamil Nadu**	49
Assam	73	Karnataka	58	**Odisha**	90	**Uttar Pradesh**	82
Bihar	62	Kerala	11	**Punjab**	51	**West Bengal**	51
Gujarat	60	Madhya Pd.	86	**Rajasthan**	79		

From the differential scenario of data and figure, there must have been some diversities in the states in India which might have played a great role on the overall level of IMF. However, this picture gives the impression of variations in the values of infant mortality rates among the major states in India.

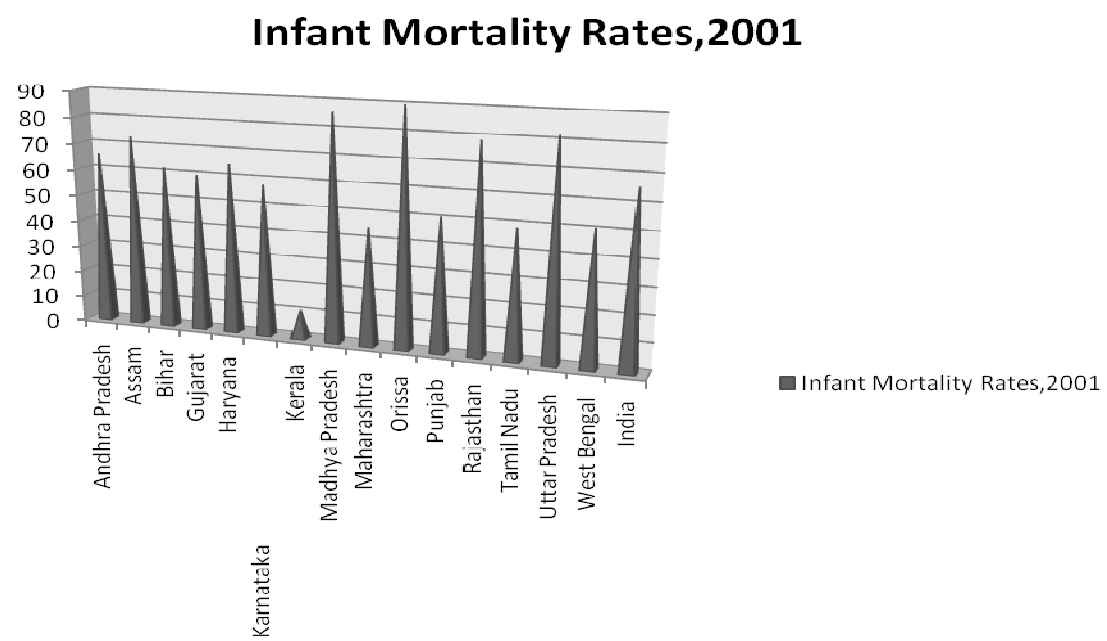


Figure 2: Infant Mortality Rates, States, India, 2001

9 Different 'over sex-ratio' at birth for overall India during 1982-2005

An attempt is made in the present paragraph to know overall values of sex-ratio at birth (Male births per 100 female births) in India. This paragraph is given here to understand the level of sex-ratio at birth for India never achieved the international standard of either 105 or 106 or 107. Scientists very commonly sometimes use sex-ratio at birth to be 105 while estimating demographic parameters without verifying the exact figure at the time of their research.

Table 4 Sex-ratio at birth at different time point during 1982-2005, India

Year		1982-'84	1983-'85	1984-'86	1985-'87	1986-'88	1988-'90	1990-92
Sex-ratio at birth		109.8	110.4	109.5	109.6	109.8	109.8	111.1
1991-93	1992-94	1993-95	1994-96	1995-97	1996-98	1998-'00	1999-'01	2000-'02
111.9	113.0	113.8	113.3	112.2	111.0	111.4	111.9	112.1
1991-93	2001-03	2002-04	2003-05					
111.9	113.3	113.4	113.6					

The table 5 above shows how the values are too high at different time point. From these high figures one may find some distinct difference between the past and the present values. As for example, during 1982 to 1990 the figures are slightly less values around 110 on an average but after that they started increasing from 111 to 114 during 1991 to 2005. It simply may show more and more female feticide and infanticide have occurred over time.

10 The third category study

Now coming to the third category of population age one, that is those below the age of 2 years as prescribed by IFPRI, the analysis is done on the basis of data on deaths below the age of 5 among the states in rural and urban areas in India in 2002. From these figures the condition of under 2 age population has been attempted here.

Table 5 Per cent under five death rate by sex and residence, 2002

States in India	Rural		Urban	
	Male	**Female**	**Male**	**Female**
Madhya Pradesh	**25.7**	**30.3**	**16.7**	**13.5**
Uttar Pradesh	**23.8**	**28.0**	**18.2**	**15.7**
Gujarat	**19.4**	**25.0**	**9.9**	**9.8**
Assam	**22.7**	**24.9**	**9.0**	**10.9**
Rajasthan	**22.9**	**23.9**	**13.3**	**18.4**
Odisha	**27.8**	**23.2**	**12.1**	**18.0**
Haryana	**16.0**	**22.2**	**10.2**	**14.4**
Punjab	**11.9**	**21.5**	**8.0**	**13.4**

Contd. from earlier page

Contd. in next page

Bihar	**15.6**	**19.7**	**14.1**	**12.9**
Andhra Pradesh	**17.1**	**18.5**	**10.0**	**5.9**
Karnataka	**18.3**	**17.4**	**6.9**	**6.6**
West Bengal	**13.7**	**12.9**	**10.2**	**6.0**
Maharashtra	**11.9**	**11.3**	**8.5**	**7.9**
Tamil Nadu	**12.8**	**11.3**	**7.9**	**6.8**
Kerala	**1.8**	**2.8**	**2.5**	**1.4**

The table given above depicts the deaths of children under age five of the major states in India in 2002. The bottom line figures are for Kerala are the lowest in rural and urban areas of the order of 1.8 for male in rural areas and 2.8 for female in the same areas. The similar low figures are in urban areas as well. There are other states like Tamil Nadu, Maharashtra and West Bengal possessing low average figures around 10 in all the cases. The position of other major states are comparatively higher values as shown in the table. As far as gender bias of deaths below the age of 5 among the states in rural and urban areas in India in 2002 there are gaps. The figures overall show less gender bias among mostly in Kerala and to some extent in other developed states like Tamil Nadu, Maharashtra, West Bengal both in rural and urban areas. In Punjab although rural figure gives lower value (11.9) but urban figure shows almost double. In undeveloped states the figures are much higher namely Madhya Pradesh, Uttar Pradesh, Rajasthan, Assam and Odisha. The figures for Andhra Pradesh, Karnataka, Bihar, Haryana are lying within these two categories.

Finally a comparison of India with South Asian countries in respect of MMR and some health indicators of children under 5 years of age has been done using the following table.

11 Comparison of health status among the South Asian Countries

Three indicators namely maternal mortality rate, per cent children under 5 severely underweight and per cent children under 5 moderately underweight have been considered for the analysis among the countries including **India, Sri Lanka, Pakistan, Maldives, B' Desh, Nepal, Bhutan, Myanmar and Afghanistan.** The table given below gives the diversified scenario among the nine countries except Afghanistan about which incomplete data are available.

Table 6 Health indicators of South Asian Countries

South Asean Countries	Health Indicators		
	MMR	%Children<5 severely underweight	% Children<5 moderately or severely underweight
India	200*	15.8**	43.5**
Sri Lanka	35.0*	3.7[d]	21.6[e]
Pakistan	260*	12.7[a]	31.3[b]
Maldives	60*	3.3[c]	17.8[c]
Nepal	170*	10.6**	38.8**
Bangladesh	240*	11.8[d]	41.3[d]
Bhutan	180*	3.2*	12.7*
Myanmar	200*	5.6*	29.6[e]
Afghanistan	460*	na	na

*2010, a>2002, b>2001, c>2009, d>2007, e>2003
**2006

Three indicators namely maternal mortality rate, per cent children under 5 severely underweight and per cent children under 5 moderately underweight have been shown in the table above. The Indian position is worth noticeable. While MMR is concerned majority of the states have reduced their rates except Afghanistan, Pakistan the rates are high and to some extent Bangladesh too. India's position is mediocre. Sri Lanka and Maldives deserve high ranking position. In terms of the figures for under 5 children's health condition in both cases are overwhelmingly high in India. From above figures it is clear that India is to traverse more paths in order to achieve the goal of MDG and Window of 1000 Day.

12 The Root Cause of Overall Situation

There are several reasons for the downfall of India's position in respect of health of mother and child. The most vulnerable point to be noted here that with many sensitive people of India being perpetually defamed in international arena in its many aspects including the present one at the cost of a big chunk of people with their many social taboos, prejudices and backwardness in their mind and thoughts. The root cause of these deficiencies in Indian people cannot be opened before the general masses who are still in darkness that why India's position is so worse when at the same time state like Kerala and some other states in south as well as in west, north or even in east boast of having very good position as compared to major big states contributing overwhelmingly the overall Indian status for which a section of good states suffer. These are all culminating from the very nature of India with much heterogeneities in its health consciousness of their mothers and children (Figures II, III, IV and V as shown in the **Appendix I**) apart from their culture, language, social systems, attitudes towards social responsibilities and so on. In

this diversified nature of the people of India an example of use of different languages as their mother tongue in different states is quoted below in table 8. There are other heterogeneities as well. Ironically the age at marriage in

India, especially in rural areas is still very low. It is, to some extent, also low even among educated youths enjoying urbanities with high fertilities due to non practice of FP methods in their early marital life if investigated.

Table 7 People in India with their languages spoken in millions in different states, 2008

Language	Speakers (in millions, 2001)	State(s)
Assamese	13	Assam, Arunachal Pradesh
Bengali	83	West Bengal, Tripura, Andaman & Nicobar Islands
Gujarati	46	Dadra and Nagar Haveli, Daman and Diu, Gujarat
Hindi	258–422	Andaman and Nicobar Islands, Arunachal Pradesh, Bihar, Chandigarh, Chhattisgarh, the national capital territory of Delhi, Haryana, Himachal Pradesh, Jharkhand, Madhya Pradesh, Rajasthan, Uttar Pradesh and Uttarakhand
Kannada	38	Karnataka
Kashmiri	5.5	Jammu and Kashmir
Konkani	2.5–7.6	Goa, Karnataka, Maharashtra, Kerala
Malayalam	33	Kerala, Andaman and Nicobar Islands, Lakshadweep, Puducherry
Marathi	72	Maharashtra, Goa, Dadra & Nagar Haveli, Daman and Diu, Madhya Pradesh
Oriya	32	Odisha
Punjabi	29	Chandigarh, Delhi, Haryana, Punjab
Tamil	61	Tamil Nadu, Andaman & Nicobar Islands, Puducherry
Telugu	74	Andaman & Nicobar Islands, Andhra Pradesh, Puducherry
Urdu	52	Jammu and Kashmir, Andhra Pradesh, Delhi, Bihar, Uttar Pradesh and Uttarakhand

From the above table it may be pointed out that Hindi is the language spoken by most of the major states in India namely Bihar, Chandigarh, Chhattisgarh, the national capital territory of Delhi, Haryana, Himachal Pradesh, Jharkhand, Madhya Pradesh, Rajasthan, Uttar Pradesh and Uttarakhand. No doubt this language is the national language, but unfortunately these states except some smaller ones also contribute maximum in the Indian population with highest MMR, IMF and many other social, demographic characteristics. As a result Indian position could not be remedied unless these people improve their status. No doubt there are other states as well where improvement is also is to be done, not to great extent. Moreover it is also to be mentioned here that India as a democratic country. People can have the right to live in their own choice of states sometimes for their livelihood and other purposes. If detailed analysis is done taking separately each state, it may be found cross language system prevails, which means People from other states migrated in West Bengal. Similarly in other states also. Heterogeneities within the state are not because of entire aborigines but a large amount of migrated people in the state as well. The culture, language, lifestyle are different among the people which means that geographic diversities are not the only ones. But it is the people's diversity above all. In this connection government usually adopt different strategies and targets in order to curb the problem, - *Population Policy (NPP), 2000, the National Health Policy (NHP), 2002, the Reproductive and Child Health (RCH) Programme (initiated since 1996) and the National Rural Health Mission (NRHM) 2005-2012. Between 4 and 5 million women in India suffer from ill- health associated with childbearing India's commitment to the eight Millennium Development Goals (MDG) set by the United Nations, among which is the goal of reduction in maternal mortality ratio (MMR) by three-fourths by 2015, will therefore require the necessary political will in the implementation of policy and programmes to safeguard maternal health* (Das et al, 2007).
Time has come now to undertake chemotherapy-like treatment to kill the dead cell from the body in each step so that all the bad prejudices, dogmas, unawareness about the situation prevailed among people irrespective of education, social economic status residing in India spreading over different regions, states, districts, blocks and villages are eradicated. But that is a huge task, alternatively GOI have classified the states into three categories. The first category comprises the "Empowered Action Group" (EAG) states of Bihar and Jharkhand, Madhya Pradesh and

Chhattisgarh, Odisha, Rajasthan, Uttar Pradesh and Uttaranchal. Assam is also included in this list. Nearly two-third of the maternal deaths in the country are reported to occur in the EAG states and Assam. Among the three categories, this category accounts for nearly 47 per cent of the births, no doubt with stunted births as well. Numbers of children who were stunted (generally defined as being significantly below the median height for their age) declined only from 51 to 44.9 percent, and those underweight from 42.7 to 40.4 percent during 1998 to 2005 (IFPRA, 2012) . Moreover, these states have had, historically higher child mortality, higher poverty levels and lower life-expectancy and other indicators than other states. The second category (Region) covers the "Southern" states of Andhra Pradesh, Karnataka, Kerala and Tamil Nadu. These states traditionally have had better child mortality and other health indicators. The remaining major states formed the third category (Region) and have been classified as 'Others' (Office of the RG in India, 2006) including Maharashtra, Punjab, West Bengal and some other with their mediocre status.

In this respect it is envisaged that although MMR declined from about 400 maternal mortality rate in 1997-'98 to about 300 in 2001-'03 and 212 during 2007-2009 gives satisfaction but at the same time tells that reducing MMR to 195 by 2012 of National Commission of Population (NCP) and National Rural Health Mission (NRHM) and to 109 by 2015 of the Millennium Development Goals (MDG) is going to be a real challenge. Particularly when most of the deaths occurred in the states included in the "Empowered Action Group" (EAG) of states an Assam. For further decline, rapid progress in health sector schemes would be needed in these states, in particular. And, these states are thus focus of the National Rural Health Mission (NRHM) as are quoted elsewhere (Government of India,2006). But Government of India with joint meeting of the different international bodies like UN, UNICEF in its report mentioned in 2010 the target is to achieve at least 135 not 109 keeping in view of the present position.

13 A Proposed Model on Demographic aspect

If a model of demographic identity system is thought to be initiated in India so that every individual will be directly connected with the government machinery so that social responsibility of the people in India would be more proactive for the benefit of the nation building. Moreover any new item occurs in respect of any individual that must either be reported to the concerned department so that the change will have to be included in the corresponding records maintained in the computer network system. It will be easy to follow up each individual on regular basis after certain specified interval of time. Moreover each individual will be liable to the prescribed schedule introduced time to time by the GOI. The National Population Register (NPR) though initiated by the Registrar General and Census Commissioner of India before another new system in the form of "AADHAAR" (Unique Identification Authority of India, 2011) to reach each individual through e-governance as and when necessary for the real benefit of mankind but the demographic data in the NPR are not complete as it seems from the website (see Appendix II). Before this GOI have already introduced Epic card also with the photo of the respective voter in order to get rid of malpractices in the election process. Then it can also be done for detailed demographic identity of the population may it be included in the NPR or a separate card may be introduced as par present model .

14 Conclusion

MDG (UN, 2000),- when developed countries registered substantial reduction in maternal morbidity and mortality, developing countries in particular constituting ninety nine per cent of the total maternal deaths which are occurring throughout the world. India is still lagging behind the target of 109 by the MDG later changed to 135 in 2015 by the GOI with the consultation of UN. International Food Policy Research Institute (IFPRI), USA (2012) recently prescribed 1000 Day Window improving nutrition during the 1000 day of mother and child. IFPRI makes some comment about India, in particular that in spite of India's green revolution and economic development is lagging behind health especially of child and mother and other. India with 2nd largest food producing country with highest democracy in the world, her overall estimates of MMR remain as high as around 212 during 2007-2009 only because of India having much heterogeneities in her culture, language, social values, etc apart from general consensus of geographical diversities. As such, GOI classified the states into different categories as mentioned in earlier paragraph in order to curb the situation by giving emphasis on respective group of states. And the present study also pointed out diversities in different aspect of health of mother and child vis-à-vis diversities among states. From the heterogeneity point of view of Indian people, keeping aside the usual attempt on states, districts, region

basis a new model on individual basis study has been proposed in the present attempt in the form of introducing identity card system of Indians in order to capture their socio economic, demographic statuses for effective planning.

15 References

1. Bread for the World Institute, 2012. International Food Policy Research Institute, IFPRI , USA.

2. Das N.P., Shah, U and Patel, R. 2007. Emerging Causes and Determinants of Maternal Mortality in India: Based on Large Scale Surveys since the 1990s.

3. Dasgupta, Monika, 1987. Selective Discrimination against Female Children in Rural Punjab, India, Population and Development Reviews, 13, The Population Council, USA.

4. Govt. of India, 2006. Maternal mortality in India, 1997-2003, Trends, causes and risk factors, Office of the ***Registrar General***, New Delhi, pp 1-18.

5. Govt. of India, 2006. AADHAAR, Unique Identification Authority of India, Bengaluru.

6. Mukhopadhyay, B.K. and Majumdar P.K. (2012). Status of Gender-Differentials and Trends in India : Population, Health, Education and Employment in Pal, M ; Pathak, P ; Bharati, P ; Ghosh, M; Gender Issues and Empowerment of Women, Nova Science Publishers, INC, New York.

7. Mukhopadhyay Barun Kumar, 1996. Trends and developments in maternal mortality vis-a-vis a differential health socio-economic infrastructural transformations among the major states in India, Socio-Economic Transformation in India, Ray, B (ed.), ***Kanishka Publishers, Distributors***, New Delhi, pp 79-110.

8. Network System, 2011. The CIA World Fact book.

9. P.M. Kulkarni, 2007. Estimation of Missing Girls at Birth and Juvenile Ages in India, A paper prepared for the United Nations Population Fund (UNFPA), India.

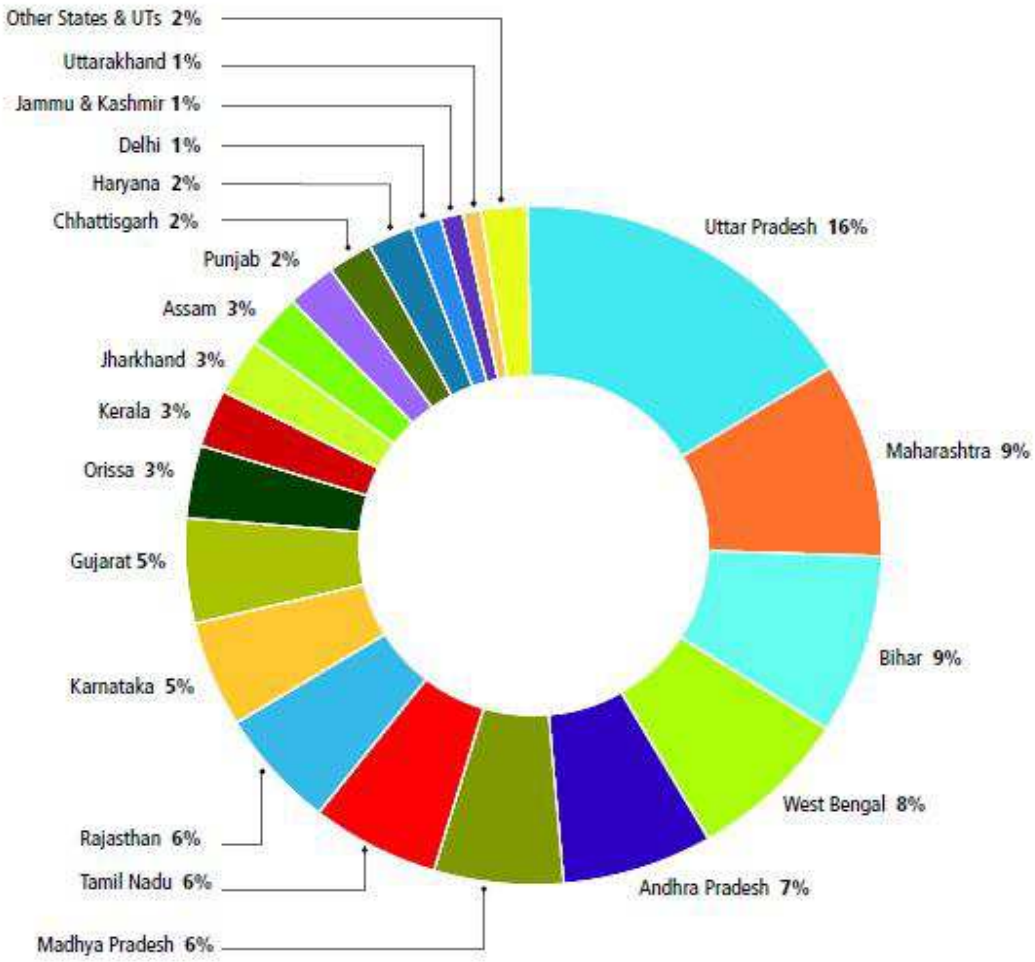
10. Ranjan, A. Gulati, S.C, 2008. INDIA, the state of population, 2007, ***Oxford University Press***, New Delhi, pp1-215.

11. RG, 2011. Special Bulletin on Maternal Mortality in India 2007-09, SRS, New Delhi.

Appendix I

Pictures shown below are differentials among state proportions

Figure I: Proportions of populations in states and UTIs, Census 2001



Contd. Appendix I

Figure II: Ante natal check up of expectant mothers

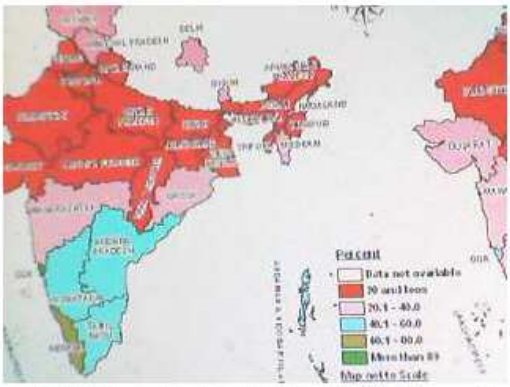


Figure III: Institutional delivery, DLHS, 2007-09

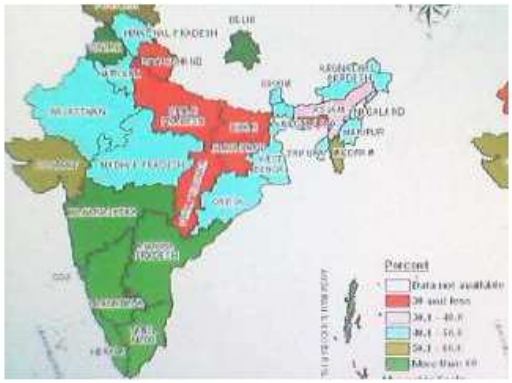


Figure IV: Full vaccination of children (12-23 months)

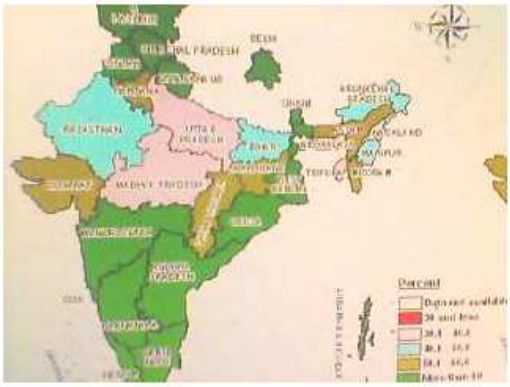
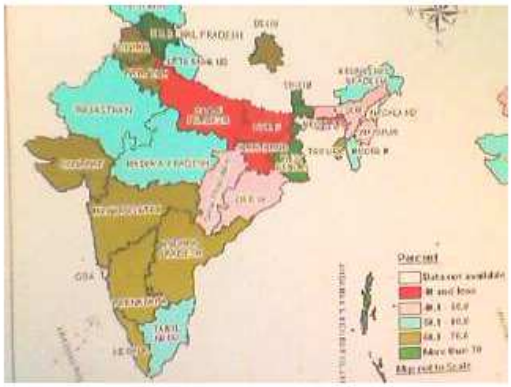


Figure V: Current use of FP methods, both, DLHS, 2007-09



APPENDIX II

NATIONAL POPULATION REGISTER (NPR)

The National Population Register (NPR) is a Register of usual residents of the country. It is being prepared at the local (Village level), sub District (Tehsil/Taluk level), District, State and National level under provisions of the Citizenship Act 1955 and the Citizenship (Registration of Citizens and issue of National Identity Cards) Rules, 2003.

The NPR would have the following details of every usual resident in the country (i) Name of Person, (ii) Relationship to head, (iii) Father's name, (iv) Mother's name (v) Spouse's name (if married), (vi) Sex, (vii) Date of Birth, (viii) Marital Status, (ix) Place of Birth, (x) Nationality as declared, (xi) Present address of usual residence, (xii) Duration of stay at present address, (xiii) Permanent residential address, (xiv) Occupation/ Activity, (xv) Educational qualification Also, three biometrics namely photograph, 10 finger prints and 2 iris would be collected for persons of age 5 years and above.

A proposal to issue Resident Identity Cards to all usual residents in the NPR of 18 years of age and above is under consideration of the Government. This Card would be a smart Card and would bear the Aadhaar number.

The NPR is a register of usual residents. The data collected in NPR will be sent to UIDAI for de-duplication and issue of Aadhaar Number. Thus the register will contain three elements of data – (i) demographic data, (ii) biometric data and (iii) the Aadhaar (UID Number).

A proposal to issue Resident Identity Cards to all usual residents in the NPR of 18 years of age and above is under consideration of the Government. This Card would be a smart Card and would bear the Aadhaar number.

Reach and Utilization of Maternal and Child Health Services in India

P. Murugesan
Research Officer and Biostatistician
Department of Fertility Studies
International Institute for Population Sciences (IIPS),
Deonar, Mumbai-India 400 088
E-mail: aasai_muru@yahoo.com

Summery

Maternal and Child Health has remained an integral part of the Family Welfare Programmes in India. The National Family Health Survey-3 (2005-06) provides enormous data related to Maternal and Child Health care- antenatal, delivery and post natal care for the women while for child care variables like child immunization-use of BCG, DPT 1 to 3 and Polio 1-3 etc. are available. The present paper therefore aims to provide the prevailing situation regarding reach and utilization of the MCH services in India as well as among the different states. It was observed that in India around 77 percent women received antenatal care for their most resent birth doing the five years residing the survey. Interestingly the rate of increase was higher in rural area as compared with urban areas. State-wise variations were noticed with regard reach and utilization of antenatal care services. Further, highest cases of safe delivery assisted by health personnel were found for the state of Tamil Nadu (91 percent) and lowest for Nagaland (12 percent). With regard to child care services percentage values obtained for fully vaccinated children range from 81 percent from Tamil Nadu and 21 percent for Nagaland. Besides providing state wise comparative picture with regard to MCH services the paper intends do discus factors affecting utilization of MCH services among the regions of the country by using logistic regression analysis. In addition attempt would be made to discus some of the plausible reasons for the variation in the utilization of MCH services.

Keywords: Reach and Utilization of Maternal and Child Health Services

Introduction

In India, the striking interregional diversity is an important confounding factor. As such, the current research analyzes data for India as a whole and for the states of Bihar, Madhya Pradesh, Rajasthan, and Uttar Pradesh in the north and Andhra Pradesh, Karnataka, Kerala, and Tamil Nadu in the south. These two groups of states are distinctly different socio-economically and culturally and are fairly representative of the north-south dichotomy observed by Dyson and Moore (1983). Southern women typically enjoy greater freedom, an outcome of the Dravidian culture, and higher levels of literacy, education, and employment. Northern women are strongly subjected to the traditional conservatism of the Mogul legacy and are predominantly illiterate, less educated, and less likely to work outside the home.

This research sets out to investigate the broad nature of the association between utilization of maternal and child health (MCH) services in India. Using data from the 2005-06 National Family Health Survey (NFHS-3), we examine the relationship between maternal schooling and factors known to reduce the risks of maternal and child mortality—utilization of antenatal and delivery care services, utilization of child immunization services, and treatment of childhood diseases. We hypothesize that the practices of educated mothers, with regard to pregnancy, childbirth, immunization, and management of childhood diseases such as diarrhea and acute respiratory infection (ARI), are quite different from those of their uneducated counterparts. In addition, the extent to which the impact of maternal schooling on the utilization of MCH services is confounded by other socioeconomic and demographic influences is examined.

Cross-country comparisons using large data sets, such as the World Fertility Survey and the Demographic and Health Surveys, have shown that education in general and female education in particular exert a very strong influence in reducing child morbidity and mortality (Boerma et al. 1990; Bicego and Boerma 1993; Caldwell and Caldwell 1990; Hobcraft, McDonald, and Rutstein 1985; Murthi, Guio, and Dreze 1995). At the micro level, more in-depth quantitative and qualitative research that examines women's health-enhancing behavior has arrived at similar conclusions (Bhuiya and Streatfield 1991; Bourne and Walker 1991).

In investigating the pathways of influence, research confirms that the causal linkages between these two factors are far from clear and that this relationship is simply not a reflection of a co-occurrence of education with other socioeconomic variables (Desai 1994; Hobcraft 1993). In spite of methodological problems associated with the measurement of maternal mortality and morbidity, several studies have shown a strong relationship between maternal mortality and morbidity and the absence of prenatal care. They have also shown that and that utilization of prenatal care is dependent on, among other factors, maternal education (Monteith et al. 1987; Okafor 1991; Wong et al. 1987). Studies of factors that influence the utilization of modern delivery-care services are, however, scarce. Several other studies have been carried out to explain how maternal education may influence child health, mostly within the conceptual framework put forward by Mosley and Chen (1984), who argue that mortality is the outcome of a combination of social, economic, biological, and environmental factors and that these factors operate through a set of proximate determinants.

Three broad pathways of influence, linking maternal schooling to child mortality, that result in greater utilization of modern health services have been suggested: educated women are better able to break away from tradition to utilize modern means of safeguarding their own health and that of their children (Caldwell and Caldwell 1988; Cleland 1990); educated women are better able to utilize what is available in the community to their advantage (Barrera 1990;

Caldwell 1990; Goodburn, Ebrahim, and Senapati 1990); and educated women may be able to make independent decisions regarding their own and their children's health leading to greater utilization of modern health facilities (Caldwell 1979; Caldwell 1986).

Objectives

The objective t of the present study is aims to provide the prevailing situation regarding reach and utilization of the maternal and child health services in India. And the specific objectives are as follows:

1. To study the reach and utilization of maternal health services in India and its major states.

2. To study the reach and utilization of child health services in India and its major states and

3. To examine the factors affecting the maternal and child health services in India and its states.

Data and Methodology

The NFHS-3 gathered during 2005-06 information on various aspects of maternal-care utilization has been used in India as well as the major states. Specifically, for each live birth in the five years preceding the survey, a woman was asked if she had received antenatal care. If she did, she was asked who administered the care, how many months pregnant she was when she first received antenatal care, how many antenatal care visits she had in all, whether she had received an injection to prevent tetanus during her pregnancy and, if so, how many injections she had received, whether she had received iron/folic-acid tablets while she was pregnant, where she gave birth, and who assisted with the delivery.

The base sample for our study of maternal-care utilization constitutes 49,369 children born in the period 1–42 months prior to the survey. These were children born to ever-married women aged 15–49 for whom information on health was obtained. For child health-care utilization, only the 45,363 children who were alive at the time of the interview are considered. In this analysis, we include five main indicators of maternal-care utilization—antenatal care, tetanus toxoid immunization, receipt of iron/folic-acid tablets, place of delivery, and delivery care. The child health-care indicators of interest are the percentage of children born in the four years preceding the survey who suffered

from ARI in the two weeks preceding the date of interview and were taken to a health facility for treatment, the percentage of children who suffered from diarrhoea in the two weeks prior to the interview and were taken to a health facility for treatment or were treated with ORS and/or RHF, and the percentage of children age 12–23 months who were fully vaccinated.

Information on maternal and child health care is birth based, that is, it is collected for each child born to respondents in the four years preceding the interview. The denominator differs for each dependent variable. For maternal health-care indicators, the denominator is all live births in the four years preceding the survey. For ARI, diarrhoea, and treatment with ORS and/or RHF, it is the number of children who were actually reported sick with the illness in the two weeks before the interview. There are five dichotomous dependent variables for maternal health-care indicators: whether or not mothers received at least two doses of tetanus toxoid vaccine; whether or not mothers received iron/folic-acid tablets; whether or not mothers received antenatal care from a health professional in an institutional setting, from a health worker at home, or from any other person whether or not mothers received delivery care from a trained health professional; and whether or not the birth was delivered in a public or private medical institution. Trained health professionals include doctors, nurses, trained nurse/midwives, and other formally trained personnel. The dichotomous dependent variables for child health-care indicators are: whether or not sick children suffering from symptoms of ARI and/or diarrhoea were taken to a health facility or health provider for treatment; whether children sick with diarrhoea were treated with ORS and/or RHF; and whether children age 12–23 months at the time of interview are fully vaccinated.

A health facility or provider includes government or municipal hospitals, primary health centers, subcenters, mobile clinics, village health guides, and government paramedics in the public sector and private hospitals or clinics, private doctors, mobile clinics, and community health workers in the private sector. A child is considered as having been taken to a health facility if he/she received care from one or more of the above sources. Children are considered fully vaccinated if they received the tuberculosis (BCG) and measles vaccinations and all three doses of the DPT (diphtheria, pertussis, tetanus) and polio vaccines. The dichotomous dependent variable in this case is children fully vaccinated or not. The age group 12–23 months is recommended by the World Health Organisation (WHO) as the age group to examine because it covers children older than 11 months (by which age they should have been fully vaccinated) and because it contains the most recent information (older children may not have been covered by more recent health programmes). Even though the sample is not restricted to one child per woman, it is unlikely to over represent women with higher fertility (who are more likely to be rural and less educated) because there are very few women who had more than one birth in the four years before the survey.

The dependant variables are maternal health care services and child health care services. Several independent variables, both socioeconomic and demographic, which could potentially influence the outcome were included in the analysis: mother's education, place of residence, employment status, caste, and religion and child's sex, birth order, and age (applicable to child health-care indicators only) at the time of interview. Mother's age at the time of birth of her child was not included since it is highly correlated with birth order. Mother's education was categorized into three groups: illiterate, literate less than middle school complete and middle school complete and above. Place of residence was urban or rural. Mother's employment status was categorized into working and not working outside the home. Even though women working for cash may be in a particularly strong position, socio-economically and otherwise, because of the small numbers working outside the home (and even smaller numbers working for cash), it made sense statistically to group women into just these two categories. Caste was divided into two groups: those belonging to scheduled castes and scheduled tribes and those not belonging to such groups. Religion was classified into Hindu, Muslim, and others. Birth order was broken down into three categories: 1, 2–3, and 4+; and child's age in months was grouped into five categories: less than 6 months, 6–11, 12–23, 24–35, and 36–47 months.

A logistic regression is run for each of the two dependent dichotomous variables. This technique examines the potential strength of education when the influence of other socioeconomic and demographic variables is controlled. The model is based on forward stepwise selection using the Wald statistic for deleting variables that exceed the 0.1 cutoff values. The same models are applied at the national level only.

Result and Discussions

The present paper made an attempt to discus two components i.e. Maternal and Child health care services. These two components are analyzed one by one using NFHS-3 all India data and the same is provided below.

Maternal Health Care Services

The Table.1 shows India as well as states differentials in five maternal care services for birth during the five years preceding the survey. These indicators together summarize the extent to which different states have progressed toward achieving safe motherhood goals at all three stages of the birth process: antenatal, delivery and postnatal. The first indicator is a summary antenatal care for the last live birth in the five years preceding the survey, mother received three or more antenatal check-ups(with the first check-up within the first trimester of pregnancy), received two or more tetanus toxoid injections, and took iron and folic acid tablets or syrup for three or more moths. The next two indictors pertain to care during delivery and show the percentage of births

delivered in medical institutions and deliveries assisted by health personnel. The last two indicators pertain to postnatal car for mothers and show the percentage of deliveries with a postpartum check-up within 42 days of the birth and within two days of birth.

For India as a whole, mothers of only 15 percent of births received all of the required components of antenatal care services. This indicator ranges from a high of 64 percent in Kerala and 56 percent in Goa to a low only 2 percent in Nagaland and 4 percent in Uttar Pradesh. Other states that performs almost all as poorly as Uttar Pradesh and Nagaland on this indicator include Bihar, Arunachal Pradesh, Madhya Pradesh, Jharkhand, Meghalaya, Rajasthan and Mizoram, where only 6-9 percent of women received the recommended components of antenatal care. Kerala , followed closely by Goa, also outperform all other states in terms of delivery care, with nearly all deliveries taking place in medical institutions and a similarly high percentage if deliveries assisted by a health professional. Tamil Nadu, with 88 percent of births delivered in medical institutions and 91 percent assisted by health professional, ranks third among the states on these delivery care indicators. By contrast, only 12-20 percent of births are delivered in medical institutions in Nagaland, Chhattisgarh, and Bihar. Only 25-29 percent of deliveries are assisted by health are assisted by health professionals in Nagaland, Uttar Pradesh, Jharkhand and Bihar.

Tamil Nadu, where 91 percent of deliveries have a postnatal check-up within 42 days of birth and 87 percent within two days, tops the list of states with regard to both of the postnatal care services. Kerala and Goa are the other two states which perform quite well on the postnatal care services.

An examination of the performance of each state on the different safe motherhood indicators shows that several states consistently perform well below the national average one each of the five services. This list includes Rajasthan in the North Region, all states in the Central Region, Bihar and Jharkhand in the East Region, and Arunachal Pradesh, Assam, Meghalaya, and Nagaland in the North East Region. Uttarachal also performs poorly on all the indicators except antenatal care, which is slightly higher than the national average. In contrast, Mizoram performs above the national average on the delivery care service and the postnatal care services, but poorly on the antenatal care service.

In case of the multivariate analysis of logistic regression models the dependent variable is considered as maternal health care services. The maternal health care services are all three types of services are included (antenatal, delivery and postnatal cares). And the independent variables were considered is mothers socio-economic and demographic background variables. The logistic models were carried out only for India as a whole and are presented in the Table.2 providing Adds Ratio or Expected B values of the dependent variables with the mother's socio-economic predictors. It observes that compared with all the background variables the mother's education has high significance in all the three factors of maternal health care services. Another very important finding was that mother's religions have significance (99 percent level of significance) with assistance during delivery and postnatal care services.

Table.1 Maternal Health Care Services for births during the five years preceding the survey by State.

India /Major States	Percentage who received all recommended types of ANC care	Percentage of births delivered in a health facility	Percentage of deliveries assisted by health personnel	Percentage of deliveries with a postnatal check-up	Percentage of deliveries with a postnatal check-up within two days of birth	Total Number of Women
India	15.0	38.7	46.6	41.2	37.3	29995
North						
Delhi	29.0	58.9	64.1	60.9	58.4	797
Haryana	14.7	35.7	48.9	57.6	55.9	783
Himachal Pradesh	17.4	43.0	47.8	50.6	43.2	668
Jammu & Kashmir	17.5	50.2	56.5	51.6	48.4	748
Punjab	19.8	51.3	68.2	63.7	62.0	837
Raiasthan	8.6	29.6	41.0	31.8	28.9	1043
Uttaranchal	16.1	32.6	38.5	35.8	32.4	644
Central						
Chattisgarh	11.3	14.3	41.6	36.3	28.4	1040
Madhya Pradesh	7.2	26.2	32.7	33.8	28.5	1753
Uttar Pradesh	4.1	20.6	27.2	14.9	13.3	3196
East						
Bihar	5.8	19.9	29.3	17.8	15.9	588
Jharkhand	7.5	18.3	27.8	19.6	17.0	716
Orissa	18.4	35.6	44.0	40.9	33.3	1166
West Bengal	12.3	42.0	47.6	44.3	40.7	1699
Northeast						
Arunachal Pradesh	6.5	28.5	30.2	23.7	22.7	344
Assam	9.6	22.4	31.0	15.9	13.9	887
Manipur	10.5	45.9	59.0	50.1	46.4	1267
Maghalaya	8.1	29.0	31.1	33.2	28.8	540
Mizoram	8.7	59.8	65.4	53.5	50.6	453
Nagaland	1.9	11.6	24.7	11.8	10.6	866
Sikkim	27.2	47.2	53.7	52.4	44.9	474
Tripura	10.6	46.9	48.8	33.7	30.3	410
West						
Goa	55.7	92.3	94.0	82.8	75.5	784
Gujarat	25.6	52.7	63.0	61.4	56.5	949
Maharashtra	21.6	64.6	68.7	64.0	58.7	2171
South						
Andhra Pradesh	28.2	64.4	74.9	73.3	64.1	1604
Karnataka	29.6	64.7	69.7	66.9	58.5	1421
Kerala	63.6	99.3	99.4	87.4	84.9	822
Tamil Nadu	34.0	87.8	90.6	91.3	87.2	1325

Source: NFHS-3

Table2. Logistic Models of dichotomous variables of Received Antenatal Care, Assistance received during Pregnancy and Received Postnatal Cares (Dependent Variables) with Mothers socio-economic background variables

Background Variables	Odds Ratio or Expected Beta Values with Significance Levels			
	Received Ante Natal Cares	**Assistance for Delivery**	**Received Postnatal Cares**	**Reference Categories**
Place of Residence				
Rural	1.11*	1.13	1.13*	Urban
Current Age				
21 - 30 Years	2.36***	0.50*	1.61*	Below 20 Years
31 – 40 Years	1.75**	.063*	1.71*	
41 and Above	1.48*	0.89	1.37	
Religion				
Muslim	1.24*	0.48***	0.78*	Hindu
Christian	1.05	0.29***	0.68**	
Others	0.84	1.58*	0.40***	
Education				
Primary Complete	0.46***	4.21***	0.36***	Illiterate
SSC complete	0.68*	2.69**	0.59**	
HSC Complete	0.66*	3.21**	0.48***	
Degree and Above	0.82	1.79*	0.60**	
Partner Education				
Primary Complete	0.91	1.13	0.76	Illiterate
SSC complete	1.28	0.82	0.88	
HSC Complete	1.34	0.54	0.88	
Degree and Above	1.38	0.28*	0.75	
Respondent Occupation				
Working	1.02	0.00	1.23***	Not working
Partner Occupation				
Working	0.23**	1.86*	0.73	Not working
Constant	0.11***	.01***	0.21***	

Note: *** (1 % level of significance), ** (5 % level of significance), * (10 % level of significance).

1. Dependent Variable 0 = Not received, 1= Received any Antenatal Care
2. Dependent Variable 0 = Not Assisted, 1= Assisted during Delivery
3. Dependent Variable 0 = Not received, 1= Receive any Postnatal Care

Child Health Care Services

Table.3 shows vaccination coverage rates for each recommended vaccination and the percentage of mothers showing a for children 12-23 months in India and each state. There are considerable interstate differentials in the coverage rates for different vaccinations and for children receiving all vaccinations. The percentage of children who are fully vaccinated ranges from 21 percent in Nagaland to 81 percent in Tamil Nadu. Tamil Nadu, Goa, Kerala and Himachal Pradesh stand out in full immunized. Among the more populous states, Uttar Pradesh (23 percent), Rajasthan (27 percent), Assam (31 percent), Bihar (33 percent), Jharkhand (34 percent), and Madhya Pradesh (40 percent) stand out as having a much lower percentage of children fully vaccinated than the national average of 44 percent. As these states account for nearly one-third of the total population of the country, their low vaccination coverage pulls down the coverage rate for the country as a whole. In addition to Nagaland and Assam, some of the other northeastern states (Arunachal Pradesh and Meghalaya) also have a relatively poor record on vaccination coverage.

A similar picture emerges with respect to individual vaccinations. In Tamil Nadu, Himachal Pradesh, Goa, Kerala, Sikkim, and Maharashtra, the coverage for BCG and at least the first doses of DPT and polio is generally in excess of 90 percent and in some cases, nearly universal. In Tamil Nadu and Goa, measles coverage is also above 90 percent. However, in most of the states, there is a considerable drop from the second to the third dose for both DPT and polio, and in almost every state fewer children have received measles vaccine than any of the other vaccinations except polio 0. The polio 0 is the polio vaccination given at the time of birth.

The analysis of logistic regression models were also carried out to examine the factor affecting the child health care services. In this models there are four dependent variables or indicators of child health information have been considered with the mothers socio-economic background predictor variable of place of the residence, current age, religion, education of the other, partner education, mothers occupation and partners ‘occupation were considered in the analysis. The mothers education has found very high significant (99 percent level) with all the four indicators. Similar pattern also observed in the case of current age of the mother except the measles vaccination of the child. In other words the literate mothers all most double as compared with the illiterate mothers in all the receiving child heath care services in India.

Table.3 Percentage of children age 12-23 months who received specific vaccinations at any time before the survey (according to a vaccination card or mother's report)

India /Major States	BCG	DPT			Polio				Measles	Fully Vaccinated
		1	2	3	0	1	2	3		
India	78.1	76.0	66.7	55.3	48.4	93.1	88.8	78.2	58.8	43.5
North										
Delhi	87.0	83.4	80.5	71.7	70.4	88.5	86.5	79.1	78.2	63.2
Haryana	84.9	83.8	81.0	74.2	52.7	92.2	91.3	82.8	75.5	65.3
Himachal Pradesh	97.2	96.6	91.9	85.1	67.1	96.8	94.6	88.6	86.3	74.2
Jammu & Kashmir	90.9	90.5	88.8	84.5	48.3	95.1	93.8	82.2	78.3	66.7
Punjab	88.0	85.9	80.4	70.5	65.6	90.1	86.7	75.9	78.0	60.1
Raiasthan	68.5	65.0	53.2	38.7	30.0	93.0	84.0	65.2	42.7	26.5
Uttaranchal	83.5	81.04	76.4	67.1	51.8	89.1	84.5	80.3	71.6	60.0
Central										
Chattisgarh	84.6	87.2	77.4	62.8	37.0	96.7	93.8	85.1	62.5	48.7
Madhya Pradesh	80.5	76.0	63.7	49.8	41.3	94.0	88.4	75.6	61.4	40.3
Uttar Pradesh	61.0	55.7	43.6	30.0	34.4	94.6	92.3	87.6	37.7	23.0
East										
Bihar	64.7	65.2	55.5	46.1	30.5	90.6	87.5	82.4	40.4	32.8
Jharkhand	72.7	66.0	53.2	40.3	25.2	93.4	87.2	79.3	47.6	34.2
Orissa	83.6	83.6	77.6	67.9	38.5	85.7	80.3	65.1	66.5	51.8
West Bengal	90.1	89.7	83.2	71.5	53.4	93.2	88.6	80.7	74.7	64.3
Northeast										
Arunachal Pradesh	57.7	57.0	48.4	39.3	34.3	72.6	65.5	55.8	38.3	28.4
Assam	62.4	66.7	56.2	44.9	27.5	81.6	72.7	59.0	37.4	31.4
Manipur	80.0	77.4	72.3	61.2	23.1	93.5	90.2	77.5	52.8	46.8
Maghalaya	65.9	62.0	56.0	47.3	31.0	81.5	74.2	56.6	43.8	32.9
Mizoram	86.4	89.1	84.5	66.8	46.4	89.0	83.7	63.5	69.5	46.5
Nagaland	46.3	47.5	36.3	28.7	13.2	79.8	68.4	46.2	27.3	21.0
Sikkim	95.9	94.9	91.2	84.3	63.4	94.0	91.2	85.6	83.1	69.6
Tripura	81.1	80.2	76.0	60.2	56.0	84.7	77.8	65.3	59.9	49.7
West										
Goa	96.8	95.7	92.6	87.5	85.6	98.6	94.0	87.2	91.2	78.6
Gujarat	86.4	82.2	73.4	61.4	59.9	92.6	83.5	65.3	65.7	45.2
Maharashtra	95.3	94.3	86.8	76.1	71.7	95.9	91.7	73.4	84.7	58.8
South										
Andhra Pradesh	92.9	92.6	76.4	61.4	28.3	96.2	94.5	79.2	69.4	46.0
Karnataka	87.8	86.7	81.5	74.0	75.1	91.8	87.9	73.8	72.0	55.0
Kerala	96.3	94.0	90.8	84.0	86.7	94.5	88.6	83.1	82.1	75.3
Tamil Nadu	99.5	98.9	97.7	95.7	94.5	99.6	96.3	87.8	92.5	80.9

in India and by state, 2005-06.

Table 4. Logistic Models of dichotomous variables of Received BCG, Received All DPTs, Received All Polios and Received Measles (Dependent Variables) with Mothers socio-economic background variables

Background Variables	Odds Ratio or Expected Beta Values with Significance Levels				
	Received BCG	**Received All DPTs**	**Received All Polios**	**Received Measles**	**Reference Categories**
Place of Residence Rural	1.31***	0.98	0.91***	1.25***	Urban
Current Age					
21 - 30 Years	1.16***	16.72***	4.56***	0.52***	Below 20 Years
31 – 40 Years	1.42***	1.21***	6.01***	1.02	
41 and Above	1.22*	12.28***	9.04***	1.15*	
Religion					
Muslim	1.038	0.92*	1.021	0.94	Hindu
Christian	0.58***	0.87***	1.12***	0.61***	
Others	0.35***	0.87**	1.059	042***	
Education					
Primary Complete	0.10***	0.18***	0.31***	030***	Illiterate
SSC complete	0.19***	0.24***	0.31***	049***	
HSC Complete	0.21***	0.29***	0.37***	053***	
Degree and Above	0.38***	0.44***	0.48***	0.79***	
Partner Education					
Primary Complete	0.85	0.79**	0.81**	085*	Illiterate
SSC complete	1.15	0.94	0.82**	1.11	
HSC Complete	1.34**	01.11	1.00	1.22*	
Degree and Above	1.70***	1.20*	1.07	1.28*	
Respondent Occupation Working	3.72	8.70	12.62	6.15	Not working
Partner Occupation Working	0.982	0.810**	0.82	.0.88	Not working
Constant	12.18***	0.015***	0.02***	2.56***	

Note: *** (1 % level of significance), ** (5 % level of significance), * (10 % level of significance).

1. Dependent Variable 0 = Not received, 1= Received BCG
2. Dependent Variable 0 = Not Assisted, 1= Received All DPTs
3. Dependent Variable 0 = Not received, 1= Received ALL Polios
4. Dependent Variable 0 = Not received, 1= Received Measles

Conclusions

To improve the availability of and access to quality health care, especially for those residing in rural areas, the poor, women, and children, the government recently launched the National Rural Health Mission for the 2005-2012 period. One of the important goals of the National Rural Health Mission is to provide access to improved health care at the household level through female Accredited Social Activities (ASHA), who act as an interface between the community and the public the health system. The ASHA acts as a bridge between the ANM and the village, and she is accountable to the Panchayat. She helps promote referrals for universal immunization, escort services for RCH, construction of household toilets, and other health care delivery programmes.

An examination of the performance of each state on the different safe motherhood indicators shows that several states consistently perform well below the national average one each of the five services. The logistic regression analysis has been observes that compared with all the background variables the mother's education has high significance in all the three factors of maternal health care services. Another very important finding was that mother's religions have significance (99 percent level of significance) with assistance during delivery and postnatal care services.

The percentage of children who are fully vaccinated ranges from 21 percent in Nagaland to 81 percent in Tamil Nadu. Tamil Nadu, Goa, Kerala and Himachal Pradesh stand out in full immunized. Among the more populous states, Uttar Pradesh (23 percent), Rajasthan (27 percent), Assam (31 percent), Bihar (33 percent), Jharkhand (34 percent), and Madhya Pradesh (40 percent) stand out as having a much lower percentage of children fully vaccinated than the national average of 44 percent. The mother's education has found very high significant (99 percent level) with all the four indicators of child immunizations was found by using logistic regression analysis.

REFERENCES

Barrera, A. 1990. The role of maternal schooling and its interaction with public health programs in child health production. *Journal of Development Economics* 32:69–91.

Bhuiya, Abbas, and Kim Streatfield. 1991. Mothers' education and survival of female children in a rural area of Bangladesh. *Population Studies* 45:253–64.

Bicego, George, and J. Ties Boerma. 1993. Maternal education and child survival: A comparative study of survey data from 17 countries. *Social Science and Medicine* 36:1207–27.

Boerma, J. Ties, Elisabeth A. Sommerfelt, Shea O. Rutstein, and Guillermo Rojas. 1990. *Immunization: Levels, trends, and differentials.* DHS Comparative Studies, No. 1. Columbia, Maryland: Institute for Resource Development.

Bourne, Katherine L., and George M. Walker, Jr. 1991. The differential effect of mothers' education on mortality of boys and girls in India. *Population Studies* 45:203–19.

Caldwell, John. 1979. Education as a factor in mortality decline: An examination of Nigerian data. *Population Studies* 33:395–413.

_____. 1986. Routes to low mortality in poor countries. *Population and Development Review* 12:171–220.

_____. 1990. Cultural and social factors influencing mortality in developing countries. *The Annals of the American Academy of Political and Social Science* 510:44–59.

Caldwell, John, and Pat Caldwell. 1988. *Women's position and child mortality and morbidity in LDC's.* Research Paper. Canberra: Department of Demography, Research School of Social Sciences, Australian National University.

Caldwell, Pat, and John C. Caldwell. 1990. *Gender implications for survival in South Asia*. Health Transition Working Paper No. 7. Canberra: National Center for Epidemiology and Population Health, Australian National University.

Cleland, John. 1990. Maternal education and child survival: Further evidence and explanations. In *What we know about health transition: The cultural, social, and behavioral determinants of health,* pp. 400–419. John Caldwell, Sally Findley, Pat Caldwell, Gigi Santow, Wendy Cosford, Jennifer Braid, and Daphne Broers-Freeman, eds. Vol. 1. Canberra: Australian National University.

Cleland, John, and Jerome K. van Ginneken. 1988. Maternal education and child survival in developing countries: The search for pathways of influence. *Social Science and Medicine* 27:1357–68.

Desai, Sonalde. 1994. *Maternal education and child health: Evidence and ideology.* Paper presented at the IUSSP Seminar on Women, Poverty, and Demographic Change, Oaxaca, Mexico, October 2–28.

Dyson, Tim, and M. Moore. 1983. On kinship structure, female autonomy, and demographic behavior in India. *Population and Development Review* 9:35–60.

Fauveau, V., M. Koenig, J. Chakraborty, and A. Chowdhury. 1988. Causes of maternal mortality in rural Bangladesh: 1976–1985. *Bulletin of the World Health Organization* 66:643–51.

Goodburn E., G. J. Ebrahim, and Sishir Senapati. 1990. Strategies educated mothers use to ensure the health of their children. *Journal of Tropical Pediatrics* 36:235–39.

Govindasamy, Pavalavalli, and B. M. Ramesh. 1996. *Maternal education and gender bias in child care practices in India.* Paper presented at the Annual Meeting of the Population Association of America, New Orleans, May 9–11.

Govindasamy, Pavalavalli, M. Kathryn Stewart, Shea O. Rutstein, J. Ties Boerma, and A. Elisabeth Sommerfelt. 1993. *High-risk births and maternity care*. DHS Comparative Studies, No. 8. Columbia, Maryland: Macro International Inc.

Hobcraft, John. 1993. Women's education, child welfare, and child survival: A review of the evidence. *Health Transition Review* 3:159–75.

Hobcraft, John, J. McDonald, and S. O. Rutstein. 1985. Socioeconomic factors in infant and child mortality: A cross-national comparison. *Population Studies* 38: 193–223.

International Institute for Population Sciences (IIPS). 1995. *National Family Health Survey (MCH and Family Planning), India 1992–93*. Bombay: IIPS.

Monteith, Richard S., Charles W. Warren, Egberto Stanziola, Ricardo Lopez Urzua, and Mark W. Oberle. 1987. Use of maternal and child health services and immunization coverage in Panama and Guatemala. *Bulletin of the Pan American Health Organization* 21:1–15.

Mosley, W. H., and L. C. Chen. 1984. An analytical framework for the study of child survival in developing countries. *Population and Development Review* 10 (Supplement): 25–45.

Murthi, Mamta, Anne-Catherine Guio, and Jean Dreze. 1995. Mortality, fertility, and gender bias in India: A district-level analysis. *Population and Development Review* 21:745–82.

Okafor, C. B. 1991. Availability and use of services for maternal and child health care in rural Nigeria. *International Journal of Gynecology and Obstetrics* 34:331–46.

Rosenzweig, M., and T. Paul Schultz. 1982. Child mortality and fertility in Colombia: Individual and community effects. *Health Policy and Education* 2:305–348.

Schultz, T. Paul. 1984. Studying the impact of household economic and community variables on child mortality. *Population and Development Review* (Supplement) 10:215–35.

UNICEF. 1990. *Development goals and strategies for children in the 1990s*. New York: UNICEF.

Ware, Helen. 1984. Effects of maternal education, women's roles, and child care on child mortality. *Population and Development Review* (Supplement) 10:191– 214.

Westoff, Charles. 1986. *Technical Notes No. 1.* Demographic and Health Surveys. Columbia, Maryland: Institute for Resource Development/Westinghouse.

Wong, Emelita L., Barry M. Popkin, David K. Guilkey, and John S. Akin. 1987. Accessibility, quality of care, and prenatal care use in the Philippines. *Social Science and Medicine* 24:927–44.

Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013)
International Conference, Mataró (Barcelona), Spain 25 - 28 June 2013

Regression Models for Repairable Systems

Petr Novák

Charles University in Prague, Faculty of Mathematics and Physics, Department of Probability and Mathematical Statistics, Sokolovská 83, 186 75 Praha 8, Czech Republic
(E-mail: novakp@karlin.mff.cuni.cz)

Abstract. When operating a device which is a subject to degradation, we want to estimate the distribution of the time to failure for maintenance optimization. Our aim is to describe the dependency of the failure time distribution on applicable regression variables. Models commonly used in survival analysis, such as the Cox model or the Accelerated failure time model, need to be adjusted to accommodate repairs and maintenance. For instance, we may use the number of repairs or maintenance actions or their cost as time-varying covariates. In this work we describe such models and demonstrate their application on real data.
Keywords: Reliability analysis, Repair models, Regression.

1 Introduction

We study data describing a service record of one ore more devices which degrade over time. When a device breaks down, it is necessary to perform a repair. We want to avoid breakdowns by performing preventive maintenance, and to optimize the maintenance costs, it is desirable to estimate the time to failure distribution with the help of available information. In this work we focus on methods of modeling the life time of the device with available regression models of survival analysis with suitable covariates. The models need to be adjusted to accommodate recurring repairs and maintenance actions. One such approach was described by Percy and Kobbacy[8] and Percy and Alkali[6] for the Cox proportional hazards model with covariates multiplicatively influencing a parametric baseline hazard. In a similar way, we show the use the Accelerated failure time model with time-varying covariates (Lin and Ying[4]), which states that the covariates influence multiplicatively the flow of the internal time of the device. Further, we show methods of estimating the cumulative baseline hazard nonparametrically if we have data on more devices, which allows us to estimate the regression parameters without assumptions on the shape of the baseline. Finally, we show the application of all described methods on real data from oil industry.

2 Modeling the life time of one device

Let $T_1, ..., T_n$ be random variables representing the ordered times of actions (repair or maintenance). Denote $\Delta_1, ..., \Delta_n$ the indicators whether in j-th time a repair ($\Delta_j = 1$) or a maintenance ($\Delta_j = 0$) was performed and let $X(t)$ be an explanatory variable, possibly time-varying.

We work with counting processes denoting the number of repairs and maintenance actions up to time t:

$$N_\bullet(t) = \sum_{j=1}^{n} I(T_j \le t, \Delta_j = 1), \qquad M_\bullet(t) = \sum_{j=1}^{n} I(T_j \le t, \Delta_j = 0).$$

Denote the hazard function

$$\lambda(t) = \lim_{h \to 0} P(N_\bullet(t+h) - N_\bullet(t) \ge 1 | \mathcal{H}(t))/h$$

where $\mathcal{H}(t)$ is the history of events up to time t. Further denote the cumulative hazard function $\Lambda(t) = \int_0^t \lambda(s)ds$ and $S(t) = \exp(-\Lambda(t))$ and $f(t) = -S'(t)$ corresponding survival function and density of the time to failure distribution. We assume that a repair returns the device to working state and that it affects the hazard function. The aim is to determine whether repairs and maintenance actions increase or decrease the hazard and by how much. We parametrize the hazard function and estimate the parameters using maximum likelihood method. The likelihood can be written as

$$L = \prod_{j=1}^{n} \left(\frac{f(T_j^-)}{S(T_{j-1})} \right)^{\Delta_j} \left(\frac{S(T_j)}{S(T_{j-1})} \right)^{1-\Delta_j} = \prod_{j=1}^{n} \lambda(T_j^-)^{\Delta_j} \cdot S(T_n)$$

and the log-likelihood has the form

$$l = \sum_{j=1}^{n} \Delta_j \log \lambda(T_j^-) - \int_0^{T_n} \lambda(t)dt.$$

Cox model

In the Cox model the covariates affect the hazard function multiplicatively. We assume that each repair or maintenance action multiplicatively increases or decreases the baseline hazard and so do other explanatory variables. We work with the hazard function in the form (Percy and Alkali[6])

$$\lambda(t) = \lambda_0(t) e^{M_\bullet(t)\rho + N_\bullet(t)\sigma + X(t)\beta} = \lambda_0(t)(e^\rho)^{M_\bullet(t)}(e^\sigma)^{N_\bullet(t)}(e^\beta)^{X(t)}.$$

As the explanatory variable $X(t)$ we can use for instance the cost of the last repair. If the covariate values change only in the times of observed events and the baseline hazard $\lambda_0(t)$ is parametric, it is possible to insert the hazard function into the log-likelihood and maximize.

Accelerated failure time model

We can also assume that each repair or maintenance causes that the internal time of the device flows faster or slower (Accelerated Failure Time model, AFT). We use the time transformation (Lin and Ying[4])

$$t \to \int_0^t e^{M_\bullet(s)\rho + N_\bullet(s)\sigma + X(s)\beta} ds =: h(t, \boldsymbol{\beta}),$$

where $\boldsymbol{\beta} = (\rho, \sigma, \beta)^T$. The hazard function has the form

$$\lambda(t) = \lambda_0(h(t, \boldsymbol{\beta}))e^{M_\bullet(t)\rho + N_\bullet(t)\sigma + X(t)\beta}.$$

If the baseline hazard function is constant (corresponding with the exponential distribution), both models coincide.

3 Inference when observing more devices

If we have data on n independent devices, we work with joint likelihood. We can either parametrize the baseline hazard and proceed as above, or estimate the baseline nonparametrically. This may be desirable, since we do not need to pose any assumptions on the form of the baseline and focus on the regression parameters. Let us have $\lambda_i(t)$, T_{ij}, Δ_{ij}, $j = 1, ...n_i$ and $X_i(t)$ the hazard function, times of events, repair indicators and covariate values for the i-th device respectively. Denote

$$N_{ij}(t) = \Delta_{ij} I(T_{ij} \leq t),$$
$$M_{ij}(t) = (1 - \Delta_{ij}) I(T_{ij} \leq t),$$
$$Y_{ij}(t) = I(T_{i,j-1} < t \leq T_{ij}).$$

We get the log-likelihood in form

$$l = \sum_{ij} \int_0^\infty \Big(\log \lambda_i(t^-) dN_{ij}(t) - Y_{ij}(t) \lambda_i(t^-) dt \Big).$$

The hazard function λ_i will contain the counts of repairs and maintenance actions $N_{i\bullet}$ and $M_{i\bullet}$, where $\bullet$ means the sum over corresponding index.

Semiparametric Cox model

Denote $\mathbf{X}_i^T(t) = (N_{i\bullet}(t), M_{i\bullet}(t), X_i(t))$. Then the likelihood and score function under the Cox model are

$$l = \sum_{ij} \int_0^\infty \Big((\log \lambda_0(t^-) + \mathbf{X}_i^T(t^-)\boldsymbol{\beta}) dN_{ij}(t) - Y_{ij}(t) e^{\mathbf{X}_i^T(t^-)\boldsymbol{\beta}} \lambda_0(t^-) dt \Big),$$

$$U(\boldsymbol{\beta}) = \sum_{ij} \int_0^\infty \Big(\mathbf{X}_i^T(t^-) dN_{ij}(t) - Y_{ij}(t) \mathbf{X}_i^T(t^-) e^{\mathbf{X}_i^T(t^-)\boldsymbol{\beta}} d\Lambda_0(t) \Big).$$

The score depends on an unknown cumulative baseline hazard $\Lambda_0(t)$. This can be replaced by the Nelson-Aalen estimate

$$\hat{\Lambda}_0(t, \boldsymbol{\beta}) = \int_0^t \frac{dN_{\bullet\bullet}(s)}{\sum_{ij} e^{\mathbf{X}_i^T(s^-)\boldsymbol{\beta}} Y_{ij}(s)}.$$

Inserting the estimate we get the score function in form

$$U(\boldsymbol{\beta}) = \sum_{ij} \int_0^\infty \left(\mathbf{X}_i(t^-) - \frac{\sum_{ij} \mathbf{X}_i(t^-) e^{\mathbf{X}_i^T(t^-)\boldsymbol{\beta}} Y_{ij}(t)}{\sum_{ij} e^{\mathbf{X}_i^T(t^-)\boldsymbol{\beta}} Y_{ij}(t)} \right) dN_{ij}(t)$$

and we find the parameter estimates by solving the equations $U(\boldsymbol{\beta}) = \mathbf{0}$.

Semiparametric AFT model

For each device we have the time transformation $t \to h_i(t, \boldsymbol{\beta})$. We work with time-transformed processes

$$N^*_{ij}(t, \boldsymbol{\beta}) = \Delta_{ij} I(h_i(T_{ij}, \boldsymbol{\beta}) \leq t),$$

$$M^*_{ij}(t, \boldsymbol{\beta}) = (1 - \Delta_{ij}) I(h_i(T_{ij}, \boldsymbol{\beta}) \leq t),$$

$$Y^*_{ij}(t, \boldsymbol{\beta}) = I(h_i(T_{i,j-1}, \boldsymbol{\beta}) < t \leq h_i(T_{ij}, \boldsymbol{\beta})),$$

$$X^*_i(t, \boldsymbol{\beta}) = X_i(h_i^{-1}(t, \boldsymbol{\beta})).$$

The exact score has a more complicated form, but it can be replaced by the approximate score (Lin and Ying[4])

$$U(\boldsymbol{\beta}) = \sum_{ij} \int_0^\infty \mathbf{X}^*_i(t^-, \boldsymbol{\beta}) \left(dN^*_{ij}(t, \boldsymbol{\beta}) - Y^*_{ij}(t, \boldsymbol{\beta}) d\Lambda_0(t) \right)$$

and we can again insert the estimate of the cumulative baseline hazard function

$$\hat{\Lambda}_0(t, \boldsymbol{\beta}) = \int_0^t \frac{dN^*_{\bullet\bullet}(s, \boldsymbol{\beta})}{\sum_{ij} Y^*_{ij}(t, \boldsymbol{\beta})}.$$

We get

$$U(\boldsymbol{\beta}) = \sum_{ij} \int_0^\infty \left(\mathbf{X}^*_i(t^-, \boldsymbol{\beta}) - \frac{\sum_{ij} \mathbf{X}^*_i(t^-, \boldsymbol{\beta}) Y^*_{ij}(t, \boldsymbol{\beta})}{\sum_{ij} Y^*_{ij}(t, \boldsymbol{\beta})} \right) dN^*_{ij}(t, \boldsymbol{\beta}).$$

Because the score is not continuous in $\boldsymbol{\beta}$, we obtain the parameter estimates by minimizing $\|U(\boldsymbol{\beta})\|$.

4 Modeling lifetime of oil pumps

We explore data on service of oil pumps during several years (Kobbacy *et al.*[1] and Percy and Alkali[7]). For one device we have detailed data on $n_1 = 65$ times of repairs, maintenance actions and the cost of each action in man-hours. This data has been studied by Percy and Alkali[6] using the Cox model. We try to model the lifetime using both the Cox and the AFT model as shown above with various parametrized baseline hazard functions and compare the results. In the parametric case, it is possible to directly maximize the likelihood for all cases and see in which it was largest.

For four other pumps we have only the times of actions at disposal, with $(n_2, ..., n_5) = (51, 90, 30, 30)$. We use both the semiparametric methods and parametrized baseline hazards with the two described models to estimate the regression parameters utilizing data of all the five pumps. The likelihood in semiparametric methods depends on the unknown baseline hazard and therefore is not available for comparison of the used methods.

Parametric modeling of one pump service

We have the times of repairs, maintenance actions and cost of each action for one pump. Using methods from section 2 we estimate the parameters ρ, σ and β in both the Cox model and the AFT model. We try to maximize the likelihood for exponential, Weibull $\lambda_0(t) = a\lambda^a t^{a-1}$, gamma $f(t) \propto t^{a-1}e^{-\lambda t}$, truncated Gumbel $\lambda_0(t) = \lambda a^t$ and log-normal baseline distributions.

Model	λ_0	log - lik	$e^{\hat{\rho}}$	$e^{\hat{\sigma}}$	$e^{\hat{\beta}}$	$\hat{\lambda}$	$\hat{a}$
	Exp.	-213.8	1.407	0.980	1.0066	0.0015	–
Cox	Weibull	-213.5	1.266	0.924	1.0064	0.0017	1.672
	Gamma	-213.8	1.405	0.918	1.0066	0.0016	1.027
	Gumbel	**-210.2**	0.701	0.745	1.0063	0.0006	1.010
	LN	-214.8	1.541	0.913	1.0069	$\hat{\mu}$=6.3	$\hat{\sigma}$=1.66
AFT	Weibull	-212.7	1.278	0.918	1.0061	0.0014	1.639
	Gamma	-213.8	1.418	0.916	1.0066	0.0014	0.918
	Gumbel	**-210.2**	1.318	0.877	1.0050	0.0005	1.001
	LN	-218.1	1.300	1.050	1.0070	$\hat{\mu}$=5.25	$\hat{\sigma}$=0.89

Table 1. The log-likelihood and parameter estimates from parametric models of the lifetime of one oil pump.

Comparing the likelihood values in Table 1 we find that it is highest for both the Cox and AFT model with the truncated Gumbel distribution. Further we see that the more each action did cost, the more it increased the hazard function or accelerated the internal time, because $e^{\hat{\beta}} > 1$. Each man-hour of the action means an increase of hazard or acceleration of time by about $0.5 - 0.7\%$. A repair itself has a positive influence ($e^{\hat{\sigma}} < 1$), with the exception of the AFT model with log-normal baseline distribution, but that is the case with the lowest likelihood value. It is interesting that according to all cases except the Gumbel distribution in Cox model, the maintenance actions tend to have a negative influence ($e^{\hat{\rho}} > 1$). This could be due to repairs often taking much more man-hours than maintenances, resulting in negative influence of both.

Semiparametric modeling of the lifetime of five pumps

For five devices we have only the times of repairs and maintenances available. The data on the cost of the actions was not available for all pumps, therefore we estimate only the regression parameters ρ and σ. We tried the Cox and the AFT models, both parametric with the same baseline distributions as above and semiparametric. In the parametric cases we maximize the log-likelihood whereas in the semiparametric approach we insert the estimate of the cumulative baseline hazard into the score function and solve the equations $U(\boldsymbol{\beta}) = \mathbf{0}$ for the Cox model and minimize $\|U(\boldsymbol{\beta})\|$ for the AFT model.

Model	λ_0	log - lik	$e^{\hat{\rho}}$	$e^{\hat{\sigma}}$	$\hat{\lambda}$	$\hat{a}$
	Exp.	-880.3	0.985	1.016	0.016	–
Cox	Weibull	-880.2	0.976	1.016	0.014	1.063
	Gamma	-880.1	0.988	1.016	0.015	0.811
	Gumbel	-880.3	0.994	1.016	0.016	0.999
	LN	-894.4	1.090	1.016	$\hat{\mu}$=3.22	$\hat{\sigma}$=0.89
AFT	Weibull	-880.2	0.980	1.015	0.014	1.038
	Gamma	-880.1	0.988	1.016	0.015	0.812
	Gumbel	**-875.1**	1.022	1.036	0.013	0.999
	LN	-879.5	1.284	1.158	$\hat{\mu}$=2.67	$\hat{\sigma}$=1.56
Cox	nonparam.	–	1.043	1.020	–	–
AFT	nonparam.	–	1.028	1.084	–	–

Table 2. The log-likelihood and parameter estimates from modeling the lifetime of five pumps.

In Table 2 we see that in all cases a repair increases the hazard or accelerates the internal time ($e^{\hat{\sigma}} > 1$). Among the parametric models, the Gumbel distribution with AFT model has the highest likelihood. In that case and also in the cases with log-normal baseline hazard and the semiparametric models, the maintenance actions have also a negative influence, whereas in the other cases it is positive. In Figure 1 we see the estimates of the cumulative baseline hazard for both Cox and AFT models. The time in the AFT model is on the transformed scale $t \to h(t, \hat{\boldsymbol{\beta}})$.

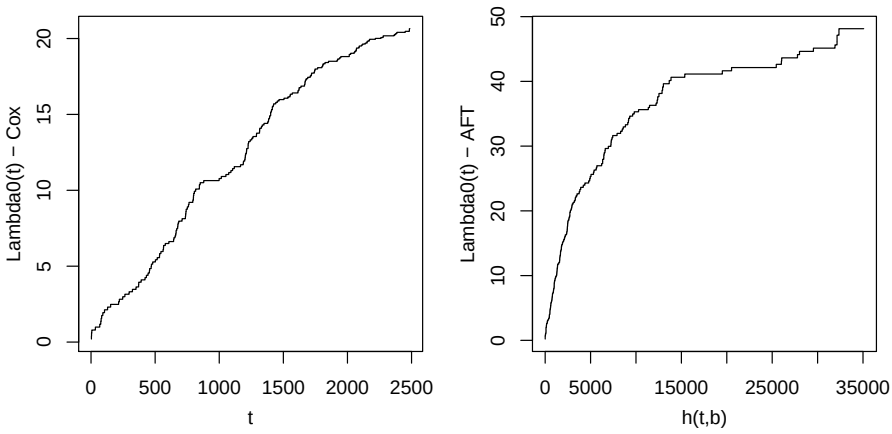


Fig. 1. Estimates of the cumulative baseline hazard in semiparametric Cox and AFT models.

Possible model selection and validation methods

When comparing two models, it is possible to perform a χ^2 test based on scaled deviance $D = 2 \cdot (l_1 - l_2)$ to see for instance whether an added covariate significantly improves a model. For the one pump data using Cox model with Gumbel baseline, no significant improvement was found when adding the influence of the cost of the action (β) to a model containing only the regression parameters ρ and σ (Percy and Alkali[6]), while it can be argued that the covariate still adds some relevant information.

Aside from the likelihood value, we do not have a direct means to determine which model fits the data best, especially when comparing the parametric and semiparametric approaches. For classic survival regression data, goodness-of-fit tests have been developed for parametric versions of both models by Lin and Spiekerman[2], for the semiparametric Cox model by Lin *et al.*[3] and for the semiparametric AFT model by Novák[5]. This methods could be adapted to accommodate repairs and maintenance actions. They are however, based on resampling approach and asymptotic convergence of certain martingale processes, and therefore it remains to be seen how well they would perform in such cases as with above data representing only a few independent devices.

5 Conclusion

We explored methods for modeling the influence of maintenance and repairs on the lifetime of the observed device. In the Cox model the covariates representing the count and size of repairs and maintenance actions influence the hazard function multiplicatively, whereas in the AFT model they accelerate or decelerate the flow of the internal time of the device. When we parametrize the baseline hazard function, the service record of one device is enough to obtain the estimates of the regression parameters. If we have data on more devices, it is possible to estimate the cumulative baseline hazard function nonparametrically. Further research could concern developing goodness-of-fit tests or testing whether a nonparametric estimate may be replaced by a suitable parametrized baseline hazard. It would be also possible to explore other transformations in the accelerated failure time model.

Acknowledgement

This work was supported by grants SVV 261315/2013 and GAUK 11122/2013.

References

1.K.A.H. Kobbacy, B.B. Fawzi, D.F. Percy and H.E. Ascher. A full history proportional hazards model for preventive maintenance scheduling. *Quality and Reliability Engineering Intl.*, 13, 187-198, 1997.

2.D.Y.Lin and C.F. Spiekerman. Model checking techniques for parametric regression with censored data. *Scandinavian Journal of Statistics*, 23, 157–177, 1996.
3.D.Y. Lin, L.J. Wei and Z. Ying. Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* 80(3), 557–572, 1993.
4.D.Y. Lin and Z. Ying. Semiparametric inference for the accelerated life model with time-dependent covariates. *Journal od Statistical Planning and Inference*, 44, 47–63, 1995.
5.P. Novák. Goodness-of-fit test for the Accelerated Failure Time model based on martingale residuals *Kybernetika* 49(1), 40–59, 2013.
6.D.F. Percy and B.M. Alkali. Generalized proportional intensities models for repairable systems. *IMA Journal of Management Mathematics* 17, 171–185, 2005.
7.D.F. Percy and B.M. Alkali. Scheduling preventive maintenance for oil pumps using generalized proportional intensities models. *International Transactions of Operational Research* 14, 547–563, 2007.
8.D.F. Percy and K.A.H. Kobbacy. Using proportional-intensities models to schedule preventive-maintenance intervals *IMA Journal of Mathematics Applied in Business & Industry*, 9, 289–302, 1998.

Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013)
International Conference, Mataró (Barcelona), Spain 25 - 28 June 2013

Stochastic modeling of hydraulic operating parameters in pipeline systems

N.N. Novitsky [1], O.V. Vanteyeva [2]

[1] Energy Systems Institute, Siberian Branch of the Russian Academy of Sciences, 130, Lermontov st., Irkutsk, 664033, Russia (e-mail: pipenet@isem.sei.irk.ru)

[2] Energy Systems Institute, Siberian Branch of the Russian Academy of Sciences, 130, Lermontov st., Irkutsk, 664033, Russia (e-mail: vanteeva@isem.sei.irk.ru)

Introduction. The problems of calculating hydraulic operating parameters are the basic problems in the analysis of operating conditions of pipeline systems when designed, operated, and controlled. These problems are traditionally solved using models and methods, which, however, do not allow us to quantitatively assess the satisfiability of operating conditions when consumption is random, which is typical of many practical situations. This is explained by high complexity and dimensionality of pipeline systems (heat-, water-, gas supply systems, etc.) as modeling objects, excessive efforts necessary to apply general methods of stochastic modeling (such as the Monte-Carlo method), and difficulties in obtaining initial statistical data.

The paper presents an approach, a set of mathematical models and methods for modeling the operating parameters of pipeline systems that were developed in terms of stochastics and dynamics of consumption processes and the established rules of their control, which make it possible to rationally combine the adequacy of modeling and its high computational efforts [1, 2].

Problem statement of the probabilistic calculation of hydraulic operating parameters. Probabilistic description of definite hydraulic operating parameters is reduced to the probability density function, which is denoted here by $p(R,\phi_R)$, where R – the value of a random vector of operating parameters (pressure, flow rate, etc.); ϕ_R – distribution parameters. Most of the practical cases allow us to use the hypothesis about normal distribution of R. Then $\phi_R = \{\bar{R}, C_R\}$ and the probabilistic description of hydraulic operating parameters can be reduced to the specification of values of mathematical expectation ($\bar{R}$) and covariance matrix (C_R) for value *R*.

Not every combination of R components is acceptable, since they should satisfy the equations of flow distribution model $U(R)=0$ (where U – non-linear vector function). These equations result from general physical conservation laws, and hence should be solved deterministically.

The traditional deterministic model of steady hydraulic operating parameters in a pipeline system as a hydraulic circuit with lumped parameters can be represented as [3]

$$U(R) = U(X,Y) = U(x,Q,P,\alpha) = \begin{pmatrix} Ax - Q \\ A^{\mathrm{T}}P - f(x,\alpha) \end{pmatrix} = 0. \qquad (1)$$

Here the first subsystem of equations represents the conditions of material balance at the nodes of hydraulic circuit (equations of the first Kirchhoff law); the second subsystem – the equations of the second Kirchhoff law; X – boundary conditions; Y – unknown operating parameters; T – transposition sign; A – $m \times n$ - incidence matrix with elements $a_{ji} = 1(-1)$, if node j is the initial (end) node for branch i, $a_{ji} = 0$, if branch i is not incident to node j; *m, n* – number of nodes and branches of the hydraulic circuit; x – *n*-dimensional vector of flow rate in branches, Q, P – *m*-dimensional vectors of nodal pressures and flow rates, $f(x,\alpha)$ – *n*-dimensional vector-function with components $f_i(x_i,\alpha_i)$, reflecting the laws of hydraulic flow for the branches; α – n_α-dimensional vector of parameters of these characteristics. For instance, if $f_i(x_i,\alpha_i) = s_i x_i |x_i| - H_i$, then $\alpha_i = \{s_i, H_i\}$, where x_i – flow rate in the *i*-th branch; s_i – hydraulic resistance of the branch; $H_i > 0$ – increase in pressure in the case of an active branch (e.g. a branch representing a pumping station); $H_i = 0$ in the case of a passive branch (e.g. a branch representing a pipeline section). If in (1) all parameters s_i, H_i, $i = \overline{1,n}$ are set deterministically, then $R = (x^{\mathrm{T}}, \bar{Q}^{\mathrm{T}}, \bar{P}^{\mathrm{T}})^{\mathrm{T}}$. Thus, the probabilistic model of steady flow distribution can be represented as $U(R)=0$, $R \sim N_r(\bar{R}, C_R)$, where N_r – *r* –dimensional normal probability distribution; *r* – dimensional of vector *R*. In the case of normal distribution of X, if we neglect the non-linear distortion of distribution $p[Y(X), \phi_{Y(X)}]$ (where $Y(X)$ – implicit function given by the flow distribution equations), the problem can be reduced to the determination of $\varphi_R = \{\bar{R}, C_R\}$ with the given function $\varphi_X = \{\bar{X}, C_X\}$ and under condition $U(R) = U(X,Y) = 0$. Moreover, the composition of X should provide solvability of equations $U(X,Y)=0$ with respect to Y, i.e. $\dim(Y) = \dim(U) = \mathrm{rank}(\partial U / \partial Y)$, where $\partial U / \partial Y$ – Jacobian matrix (of partial derivatives) under fixed boundary conditions X^* in the neighborhood of the solution point Y^*, $\dim(\cdot)$ – vector dimensional, $\mathrm{rank}(\cdot)$ – matrix rank.

Methodological approach. Let $\xi_X = (X - \bar{X})$ be a random deviation of possible realization of boundary conditions from its mathematical expectation $\bar{X}$. After linearizing function $Y(X)$ in the neighborhood of $\bar{X}$, we

obtain $Y \approx Y(\bar{X}) + (\partial Y / \partial X)\xi_X$, where $\partial Y / \partial X$ is derivative matrix at point $\bar{X}$. Since $E(Y) = \bar{Y}$ and $E(\xi_X) = 0$, where E is the operation of mathematical expectation, then $\bar{Y} = Y(\bar{X})$. Thus, the mathematical expectation of unknown operating parameters ($\bar{Y}$) is the function of flow distribution equations under boundary conditions $\bar{X}$. Correspondingly,

$$\bar{R} = \begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix} = \begin{pmatrix} \bar{X} \\ Y(\bar{X}) \end{pmatrix} \tag{2}$$

and $C_R = E\left[\begin{pmatrix} \xi_X \\ \xi_Y \end{pmatrix}\begin{pmatrix} \xi_X \\ \xi_Y \end{pmatrix}^{\mathrm{T}}\right] = \left[\begin{array}{c|c} C_X & C_{XY} \\ \hline C_{YX} & C_Y \end{array}\right]$, where $C_Y = E\left[\xi_Y \xi_Y^{\mathrm{T}}\right] \approx E\left[\frac{\partial Y}{\partial X}\xi_X \xi_X^{\mathrm{T}}\left(\frac{\partial Y}{\partial X}\right)^{\mathrm{T}}\right] = \frac{\partial Y}{\partial X} C_X \left(\frac{\partial Y}{\partial X}\right)^{\mathrm{T}}$,

$C_{XY} = C_{YX}^{\mathrm{T}} = E(\xi_X \xi_Y^{\mathrm{T}}) = E\left(\xi_X \xi_X^{\mathrm{T}}\left(\frac{\partial Y}{\partial X}\right)^{\mathrm{T}}\right) = C_X\left(\frac{\partial Y}{\partial X}\right)^{\mathrm{T}}$, $\xi_Y = (Y - \bar{Y})$.

Thus, *the general scheme for solving the problem of probabilistic calculation of hydraulic parameters* is reduced to the following: 1) to obtain vector $\bar{Y}$ by traditional methods for calculating the flow distribution with the given $\bar{X}$; 2) to determine matrix C_R, whose individual blocks are determined using the known matrix C_X and derivative matrix $\partial Y / \partial X$ at point $\bar{X}$.

Here two main questions arise: 1) based on what do we set the distribution parameters of boundary conditions ($\varphi_X = \{\bar{X}, C_X\}$); 2) what is the final form of relationships for the resultant covariance matrices in different variants of the division of R into X and Y, since in the traditional methods for the flow distribution calculation the derivatives $\partial Y / \partial X$ are not calculated in explicit form, which represents a separate problem.

Probabilistic description of consumer loads. A typical example of pipeline systems operating under the conditions of stochastic consumer loads is water supply systems. The approach applied to the probabilistic description of these stochastic conditions is based on the use of the queuing theory methods and on results of the studies [4, 5, etc.], which found their reflection in the regulatory documents [6]. According to these results, the probability of using plumbing units (p_{hr}) can be described by Erlang formulas, which demonstrate a discrete limit distribution of used channels, depending on the characteristics of the flow of requests and the performance of the queuing system.

The suggested technique for calculating the mathematical expectation of consumer flow rates ($\bar{q}_{hr}$) and their variances ($\sigma^2_{q,hr}$) consist in the following:

1. Knowing the number of plumbing units at the consumption node (N) and the probability of using them p_{hr} [6], we can calculate $m = \bar{m}_{hr}$ such that maximum value ($p_{\max}(m)$) acquires the probability

$$p(m) = \left(\frac{\left(N\, p_{hr}\right)^m}{m!}\right) \Big/ Z \;,\; m = 0, 1, \ldots, N \;, \tag{3}$$

where $Z = \sum_{k=0}^{N} \frac{\left(N\, p_{hr}\right)^k}{k!}$, m is the number of simultaneously used plumbing units; Np_{hr} is their usage rate.

2. We should determine the average hourly flow rate $\bar{q}_{hr} = \bar{m}_{hr}\, q_{0,h}$, where $q_{0,h} = q_{0,hr} / 1000$ – hourly water flow rate by one device, m^3/h; $\bar{q}_{hr}$ – can be interpreted as the mathematical expectation of flow rate at the consumption node; $q_{0,hr}$ – standardized value, l/h.

3. When approximating the discrete Erlang distribution by the continuous normal distribution, we should calculate the equivalent variance by formula $\sigma^2_{m,hr} = 1 / 2\pi p^2_{\max}(m)$.

4. The variance of the average hourly flow rate will be determined as $\sigma^2_{q,hr} = q^2_{0,h}\, \sigma^2_{m,hr}$.

Figure 1 presents a diagram of function (3), where N =270 and p_{hr} =0.023. The diagram shows that the maximum probability density function corresponds to $\bar{m}_{hr}$, whose average hourly flow rate is $\bar{q}_{hr}$.

General scheme of obtaining the covariance matrix consists of three stages: 1) to linearize system (1) at point $\bar{X}$; 2) to reduce linearized system $\frac{\partial U}{\partial R}\xi_R = 0$ to $\xi_Y = \frac{\partial Y}{\partial X}\xi_X$; 3) to obtain covariance matrix of the vector of unknown operating parameters C_R using the operation $E\left[\begin{pmatrix} \xi_X \\ \xi_Y \end{pmatrix}\begin{pmatrix} \xi_X \\ \xi_Y \end{pmatrix}^{\mathrm{T}}\right]$.

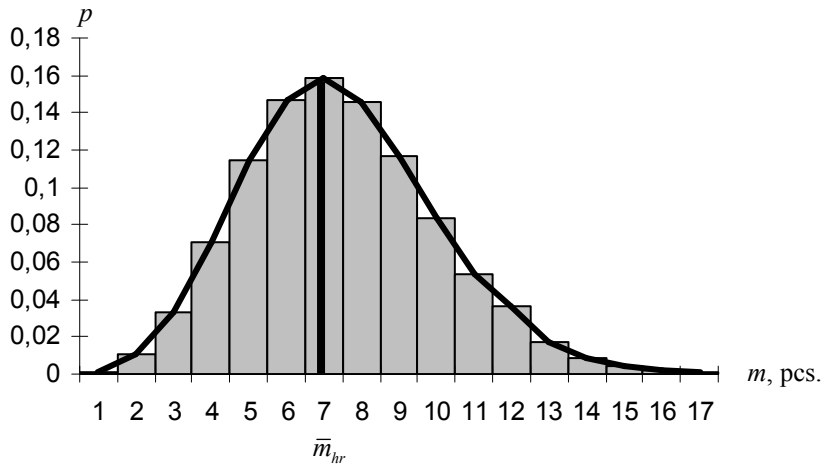


Fig. 1. Continuous approximation of Erlang distribution for the probability of simultaneously used devices for the case where N =270 and p_{hr} =0.023.

Thus, for the case, where $X = Q$, $Y = \begin{pmatrix} x \\ P \end{pmatrix}$, $P_m = \text{const}$, $\alpha = \text{const}$: $\dfrac{\partial U}{\partial R} = \left[\begin{array}{c|c} A & 0 \\ \hline f' & A^{\mathrm{T}} \end{array}\right]$;

$$\begin{pmatrix} \xi_x \\ \xi_P \end{pmatrix} = \left[\begin{array}{c} (f_x')^{-1} A^{\mathrm{T}} M^{-1} \\ \hline M^{-1} \end{array}\right] \xi_Q ;$$

$$C_R = \left[\begin{array}{c|c|c} C_Q & C_{Qx} & C_{QP} \\ \hline C_{xQ} & C_x & C_{xP} \\ \hline C_{PQ} & C_{Px} & C_P \end{array}\right] = \left[\begin{array}{c|c|c} C_Q & C_Q M^{-1} A (f_x')^{-1} & C_Q M^{-1} \\ \hline (f_x')^{-1} A^{\mathrm{T}} M^{-1} C_Q & (f_x')^{-1} A^{\mathrm{T}} M^{-1} C_Q M^{-1} A (f_x')^{-1} & (f_x')^{-1} A^{\mathrm{T}} M^{-1} C_Q M^{-1} \\ \hline M^{-1} C_Q & M^{-1} C_Q M^{-1} A (f_x')^{-1} & M^{-1} C_Q M^{-1} \end{array}\right],$$

where f_x' – diagonal matrix with elements $\partial f_i(x_i, \alpha_i) / \partial x_i$; C_Q – known covariance matrix of nodal flow rate; C_P, C_x – covariance matrix of nodal pressure and covariance matrix of flow rate in branches; $C_{Qx} = C_{xQ}^{\mathrm{T}}$ – covariance matrix of nodal flow rate and flow rate in branches; $C_{PQ} = C_{QP}^{\mathrm{T}}$ – covariance matrix of nodal pressure and flow rate; $C_{Px} = C_{xP}^{\mathrm{T}}$ – covariance matrix of nodal pressure and flow rate in branches. Thus, knowing $C_X = C_Q$, we can calculate C_R. No special requirements are imposed on matrix C_Q, however, in practice it is usually taken as a diagonal matrix from considerations of statistical independence of consumer loads. This means that $\operatorname{cov}(Q_j, Q_t) = \sigma_{Q_j}^2$ for $j = t$, and $\operatorname{cov}(Q_j, Q_t) = 0$ for $j \neq t$.

Covariance matrix for the general case of setting boundary conditions $X = (Q_X^{\mathrm{T}}, P_X^{\mathrm{T}}, \alpha_X^{\mathrm{T}})^{\mathrm{T}}$, where at each node we can set either the flow rate or the pressure, and each branch is characterized by $n_{\alpha,i}$-dimensional vector (e.g. $\alpha_i = \{s_i, H_i\}$, $n_{\alpha,i} = 2$) of hydraulic parameters, which is specified in the probabilistic form in full or partially [1, 2].

Divide the set of nodes in the design scheme into subsets of nodes with the given flow rate (J_Q) and pressure (J_P), and the set of branches into subsets of branches with hydraulic parameters given in the probabilistic (I_V) and deterministic (I_D) forms. We omit the conclusion and give the finite expressions for the covariance matrix of unknown operating parameters:

1) Covariance matrix of unknown nodal pressure

$$C_{PY} = \mathrm{E}\left[\xi_{PY},\ \xi_{PY}^{\mathrm{T}}\right] = \frac{\partial P_Y}{\partial Q_X} A_{QV} \frac{\partial x_V}{\partial \alpha_V} C_{\alpha V} \frac{\partial x_V}{\partial \alpha_V} A_{QV}^{\mathrm{T}} \left(\frac{\partial P_Y}{\partial Q_X}\right)^{\mathrm{T}} + \frac{\partial P_Y}{\partial Q_X} C_{QX} \left(\frac{\partial P_Y}{\partial Q_X}\right)^{\mathrm{T}} + \frac{\partial P_Y}{\partial P_X} C_{PX} \left(\frac{\partial P_Y}{\partial P_X}\right)^{\mathrm{T}};$$

2) Covariance matrix of flow rate in the branches with deterministically specified characteristics

$$C_{x,D} = \mathrm{E}\left[\xi_{x,D}, \xi_{x,D}^{\mathrm{T}}\right] = \frac{\partial x_D}{\partial P_Y} C_{PY} \left(\frac{\partial x_D}{\partial P_Y}\right)^{\mathrm{T}} + \frac{\partial x_D}{\partial P_X} C_{PX} \left(\frac{\partial x_D}{\partial P_X}\right)^{\mathrm{T}};$$

3) Covariance matrix of flow rate in the branches with probabilistically specified characteristics

$$C_{xV} = \mathrm{E}\left[\xi_{xV}, \xi_{xV}^{\mathrm{T}}\right] = \frac{\partial x_V}{\partial P_Y} C_{PY} \left(\frac{\partial x_V}{\partial P_Y}\right)^{\mathrm{T}} + \frac{\partial x_V}{\partial P_X} C_{PX} \left(\frac{\partial x_V}{\partial P_X}\right)^{\mathrm{T}} + \frac{\partial x_V}{\partial \alpha_V} C_{\alpha V} \left(\frac{\partial x_V}{\partial \alpha_V}\right)^{\mathrm{T}};$$

4) Covariance matrix of unknown nodal flow rates

$$C_{QY} \equiv \mathrm{E}\left[\xi_{QY}, \xi_{QY}^{\mathrm{T}}\right] = A_{PD} C_{xD} A_{PD}^{\mathrm{T}} + A_{PV} C_{xV} A_{PV}^{\mathrm{T}},$$

where A_{QD} – $(m_Q \times n_D)$-dimensional incidence matrix with elements a_{ji}, $j \in J_Q$, $i \in I_D$; A_{QV} – $(m_Q \times n_V)$-dimensional incidence matrix with elements a_{ji}, $j \in J_Q, i \in I_V$; A_{PD} – $(m_P \times n_D)$-dimensional incidence matrix with elements a_{ji}, $j \in J_P$, $i \in I_D$; A_{PV} – $(m_P \times n_V)$-dimensional incidence matrix with elements a_{ji}, $j \in J_P$, $i \in I_V$; $\frac{\partial x_D}{\partial P_Y} = \left(\frac{\partial f_{xD}}{\partial x_D}\right)^{-1} A_{QD}^{\mathrm{T}}$, $\frac{\partial x_D}{\partial P_X} = \left(\frac{\partial f_{xD}}{\partial x_D}\right)^{-1} A_{PD}^{\mathrm{T}}$, $\frac{\partial x_V}{\partial P_Y} = \left(\frac{\partial f_{xV}}{\partial x_V}\right)^{-1} A_{QV}^{\mathrm{T}}$, $\frac{\partial x_V}{\partial P_X} = \left(\frac{\partial f_{xV}}{\partial x_V}\right)^{-1} A_{PV}^{\mathrm{T}}$, $\frac{\partial x_V}{\partial \alpha_V} = \left(\frac{\partial f_{xV}}{\partial x_V}\right)^{-1} \frac{\partial f_{xV}}{\partial \alpha_V}$ – matrices of partial derivatives of the corresponding combinations of parameters, which implicitly depend on three matrices only: $\frac{\partial f_{xD}}{\partial x_D}$, $\frac{\partial f_{xV}}{\partial x_V}$ and $\frac{\partial f_{xV}}{\partial \alpha_V}$, whose structure is determined by the type of branch characteristics. Moreover, the first two of them are diagonal, and therefore, easily invertible.

Thus, based on the given relations, we can sequentially calculate the covariance matrices of all the operating parameters, if we know the covariance matrices of nodal flow rate set in the probabilistic form (C_{QX}), nodal pressure (C_{PX}), and hydraulic characteristics of branches ($C_{\alpha V}$).

Probabilistic calculation of dynamics of hydraulic operating parameters. Stochastic boundary conditions initiate the change in hydraulic operating parameters with time. As a result we face the problem of probabilistic modeling and analysis of operating parameter dynamics $R(t), 0 \le t \le T$ as a random process for the calculation period T.

Figure 2 presents the graphs of realization-frequency distribution of two hydraulic operating parameters (the nodal flow rate and the nodal pressure). The first parameter can be considered as a disturbance, the second – as a response. Figure 2a shows the graph of water flow rate frequencies for an individual residential building in the water supply system that is constructed based on the experimental data. Figure 2b shows the graph of pressure frequencies at the connection node of the reservoir in the water supply system in one of the Irkutsk districts that is obtained by processing the data of the dispatching department for 490 days.

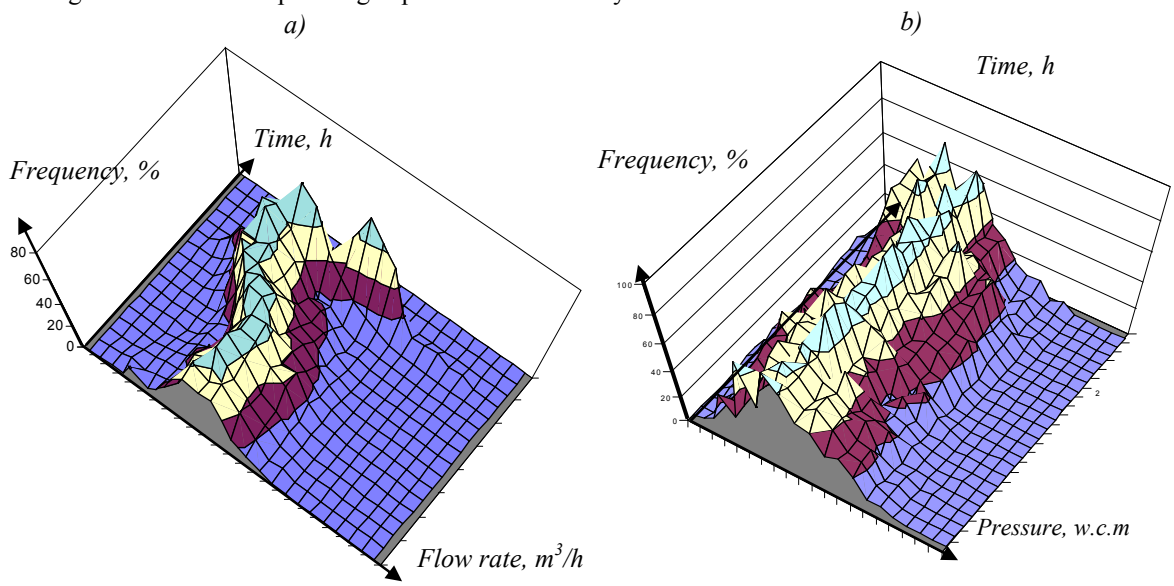


Fig. 2. Daily change in the frequency distribution of hydraulic operating parameters

a) For the nodal flow rate, b) For the nodal pressure

Analysis of both processes in Fig. 2 indicates that: 1) the frequency distribution at any cross-section of both processes is approximated by the normal (Gaussian) distribution satisfactorily enough; 2) the variance of every process (σ^2) is practically invariable. The root-mean-square deviation (σ) for daily water flow rate changes negligibly, i.e. within 10 per cent (Table 1), for pressure – within 7 per cent; 3) the mathematical expectation for both processes changes during a day (Fig. 3a); 4) the autocorrelation function stabilizes at the zero value (for the nodal value in Fig. 3b) fast enough.

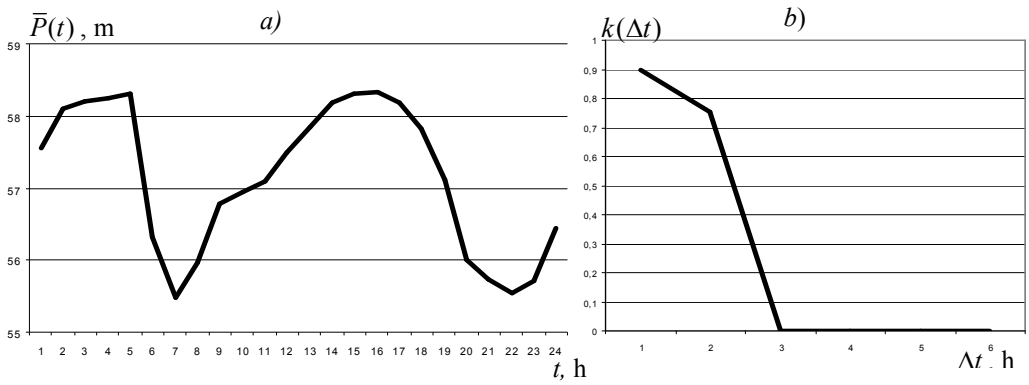


Fig. 3. Statistical characteristics of change in the nodal pressure as a random process
a) Dynamics of mathematical expectation;
b) Graph of the autocorrelation function of pressure in the reservoir

Table 1. Values of mathematical expectations and root-mean-square deviations of the nodal flow rate during day hours for the conditions in Fig. 2a.

Day hour	$\overline{Q}$, m^3/h	σ	$\frac{\sigma}{\overline{\sigma}}100\%$	Day hour	$\overline{Q}$, m^3/h	σ	$\frac{\sigma}{\overline{\sigma}}100\%$
1	4.60	2.14	3.11	13	11.97	2.25	8,41.
2	2.32	1.94	6.52	14	11.77	2.2	6.00
3	1.88	1.99	4.12	15	11.28	2.15	3.59
4	1.66	1.94	6.52	16	11.16	2.07	0.26
5	1.87	1.84	7.97	17	11.53	2.0	3.63
6	3.28	2.2	6.00	18	12.32	2.09	0.70
7	7.88	2.02	2.67	19	12.35	2.17	4.56
8	10.80	2.08	0.22	20	13.34	2.05	1.22
9	10.88	1.96	5.56	21	13.68	2.04	1.71
10	12.40	2.26	8.89	22	14.34	2.02	2.67
11	12.48	2.02	2.67	23	12.51	1.85	10.86
12	12.13	2.28	9.86	24	9.10	2.18	5.04

The hydraulic operating parameters vary in time in response to three main disturbing actions (boundary conditions): 1) random actions of regular character (consumer loads); 2) deterministic actions of regular character (control actions); 3) random actions of irregular character (fires, accidents). The second type of disturbances is taken into account algorithmically on the basis of the specified control rules. Analysis of the consequences of relatively rare disturbances of the third type is the subject of the reliability theory of pipeline systems and is not carried out here.

Dynamics of hydraulic operating parameters $R(t)$, $0 \le t \le T$ may be considered as a random process with the discrete time (a quasidynamic approach). At each time instant of the process the operating parameters obey the normal distribution. Variation of the operating parameters at the adjacent instants may be considered as insignificant and the flow distribution – as steady. Thus, the problem of probabilistic calculation of hydraulic operating parameter dynamics is reduced to the determination of $\overline{\mathbf{R}} = [\overline{R}(0)^{\mathrm{T}}, \overline{R}(1)^{\mathrm{T}}, ..., \overline{R}(T)^{\mathrm{T}}]^{\mathrm{T}}$ and $\mathbf{C_R} = E\left[\xi_R \xi_R^{\mathrm{T}}\right]$ based on the specified parameters $\overline{\mathbf{X}} = [\overline{X}(0)^{\mathrm{T}}, \overline{X}(1)^{\mathrm{T}}, ..., \overline{X}(T)^{\mathrm{T}}]^{\mathrm{T}}$, $\mathbf{C_X}$ and the conditions $A(t)x(t) = Q(t,P)$, $\overline{A}(t)^{\mathrm{T}} \overline{P}^{\mathrm{T}}(t) = y(t)$, $y(t) = f\left(x(t), \alpha(t)\right)$, $t = 0,...,T$. In this case the suggested analytical probabilistic models and the calculation methods can be applied to each calculation instant, which will sharply decrease computational efforts. The computing experiments in Table 1 have shown the decrease in running time by tens of times.

Table 2. Time required for probabilistic calculation by the Monte Carlo and analytical methods [7]

Number of scheme nodes and branches	Time for method		$t_{\text{Ì -Ê}} / t_{\text{âí àëèò.}}$
	analytical	Monte Carlo	
6 nodes and 8 branches	3.2 s	3 min	56.25
12 nodes and 19 branches	4.8 s	28.5 min	356.25
12 nodes and 29 branches	16.2 s	1.25 h	277.77

In some cases such as availability of reservoirs it is important to take account of the lagging factor of internal responses of pipeline systems, when the successive operating condition depends on the prehistory of conditions. Availability of reservoirs can be taken into account by using the additional dynamic relation $P_{j,k} = P_{j,k-1} + \rho g(\Delta t / F_j) Q_{k,j}$, where Δt – duration of the k-th condition; F_j – liquid surface area in the reservoir; j – index of the node with a reservoir; g – gravitational acceleration; ρ – liquid density. The reservoir operation can be modeled by insertion of a dummy branch connected to a dummy node with zero (or air) pressure. The hydraulic characteristic of such a branch has the form: $y_{i,k} = s_{i,k} x_{i,k} - H_{i,k}$, where $H_{i,k} = P_{j,k-1}$, $s_{i,k} = \rho g \Delta t / F_j$. Let H_k^f be a vector of dummy pressure rises in the branches that represent all the reservoirs. The covariance matrix of vector H_k^f that is used at the k-th calculation step will have the form: $C_{H_f}(t_k) = C_{PY}^*(t_{k-1})$, where $C_{PY}^*(t_{k-1})$ – block of covariance matrix C_{PY} that was calculated at the previous step and is attributed to the pressures at the nodes with reservoirs.

Calculation of probabilistic operating parameters of pipeline systems. The suggested approach to the calculation of statistical parameters of pipeline system operation offers an opportunity to obtain probabilistic estimates of virtually any operating parameters of pipeline systems depending on their operating conditions by the known formulas of the probability theory. For example the probability that any "nondegenerate" subset of operating parameters belongs to a given range at the time t_k will be determined by the formula

$$p_{Rk} = \frac{1}{\sqrt{(2\pi)^n |C_{Rk}|}} \int_{\underline{v}_n}^{\overline{v}_n} \ldots \int_{\underline{v}_1}^{\overline{v}_1} \exp\left\{ -\frac{1}{2} \left(R_k - \overline{R}_k \right)^{\mathrm{T}} \mathrm{C}_{Rk}^{-1} \left(R_k - \overline{R}_k \right) \right\} dR_1 \mathrm{K}\, dR_n , \tag{4}$$

where R_k – n-dimensional vector (subvector) of operating parameter values at the time instant t_k; $\overline{R}_k$ – n-dimensional vector of mathematical expectation R_k; C_{Rk} – ($n \times n$)-dimensional covariance matrix for R_k; p_{Rk} – probability that R_k belongs to a specified range [$\overline{v}$, $\underline{v}$]; $\overline{v} = [\overline{v}_1, \ldots, \overline{v}_n]^{\mathrm{T}}$ and $\underline{v} = [\underline{v}_1, \ldots, \underline{v}_n]^{\mathrm{T}}$ – vectors of upper and lower boundaries of the studied range, whose components can take infinite values to take account of one-sided intervals or their absence.

The assessment of probability that R_k belongs to a specified range [$\overline{v}_r$, $\underline{v}_r$] during period T will be determined by the formula

$$p_{RT} = \sum_{k=1}^{K} \left(p_{Rk} \Delta t_k \right) \Big/ \sum_{k=1}^{K} \Delta t_k = \sum_{k=1}^{K} \left(p_{Rk} \Delta t_k \right) \Big/ T , \tag{5}$$

where K – the number of calculated periods over period $T = \sum_{k=1}^{K} \Delta t_k$; Δt_k – duration of the k-th condition.

Equations (4) and (5) can be applied to estimate the operation of pipeline system, its fragments or individual components in a definite operating condition or over the period of time, for example in terms of the extent to which they are loaded, consumer demand is satisfied, or process constraints are met, etc.

Numerical example

Let us consider a numerical example of calculating the stochastics of the hydraulic operating parameters for the network presented in Fig.4. The network consists of 7 nodes and 11 branches of which: one node has a fixed pressure; two nodes have lumped loads; two nodes are nonfixed loads depending on pressure; one branch represents a pumping station with an increasing head H_0=21 m; one dummy branch simulates a reservoir (water level in the reservoir H^f=16.4 m); two dummy branches simulate nonfixed loads, their resistances are random values. Thus, this example illustrates the possibilities of the suggested approach in terms of the random composition of boundary conditions.

The input information specified in the probabilistic form is: $X = (Q_X^{\mathrm{T}}, P_X^{\mathrm{T}}, \alpha_X^{\mathrm{T}})^{\mathrm{T}} = (\overline{Q}_4, \overline{Q}_5, \overline{P}_7, \overline{s}_9, \overline{s}_{10})$ = (5.2, 1.8, 0, 0.30359, 1.2407); C_X – a diagonal matrix with nonzero elements (1.065, 0.3969, 0.0001, 0,059, 0.51564).

Resistances in the dummy branches 9 and 10, that simulate nonfixed flow rates at consumers are determined by the formula [1, 2] $\overline{s}_i = P_j^r / (Q_j^r)^2$, and variances – $\sigma_{s,i}^2 = \left(4 (P_j^r)^2 / (Q_j^r)^6 \right) \sigma_{Qr,j}^2$, where P_j^r, Q_j^r – design (required) pressures and flow rates for this consumer, j – index of the initial node of the i-th branch. Correspondingly in the example $Q_2^r = 7.7$, $\sigma_2^2 = 9.61$, $P_2^r = 18$, $Q_3^r = 7.11$, $\sigma_3^2 = 0.81$, $P_3^r = 12$.

Resistances in the branches that were specified deterministically are: $s_1 = 0.00257$, $s_2 = 0.8996$, $s_3 = 0.00408$, $s_4 = 0.095$, $s_5 = 0.67$, $s_6 = 0.067$, $s_7 = 0.0957$, $s_8 = 0.00646$, $s_{11} = 0.014$.

The calculation results for nodes are presented in Table 3 and for branches – in Table 4.

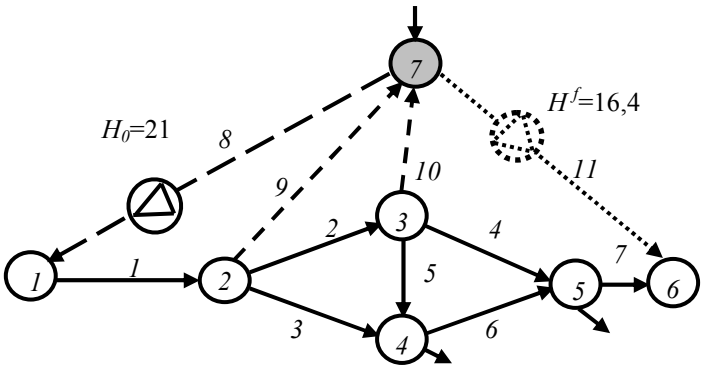


Fig.4. Example of the calculated scheme of the pipeline system for the general case of boundary conditions

Real section; dummy branch simulating nonfixed consumer loads;
dummy branch simulating reservoir; dummy branch simulating pumping station;
node with the specified nodal loads; node with the specified pressure.

Table 3. Calculation results for nodes

j	Parameters			
	P_j, *Mwc*	$\sigma_{P,j}$	Q_j, m^3/h	$\sigma_{Q,j}$
1	18.22	0.89	–	–
2	17.11	1.25	9.19	4.03
3	14.96	1.21	6.48	1.08
4	16.70	1.25	–	–
5	16.01	0.83	–	–
6	16.37	0.02	–	–
7	–	–	22.67	9.07

Table 4. Calculation results for branches

i	Parameters	
	x_i, m^3/h	$\sigma^2_{x,i}$
1	20.75	3.29
2	1.54	0.02
3	10.01	0.03
4	-3.32	0.59
5	-1.61	0.02
6	3.20	0.98
7	-1.92	2.21
8	20.75	3.29
9	9.19	4.03
10	6.48	1.08
11	1.82	0.66

Figure 5 presents a graphical interpretation of the calculated probability of providing consumers with a required flow rate. For example for the consumer at the second node $p(0 < \bar{Q}_2 < Q_2^r) \approx 0.3442$ or $p(Q_2^r < \bar{Q}_2 < +\infty) \approx 0.64446$, and at the third node $p(0 < \bar{Q}_3 < Q_3^r) \approx 0.71914$ or $p(Q_3^r < \bar{Q}_3 < +\infty) \approx 0.28083$, where Q^r is the required flow rate.

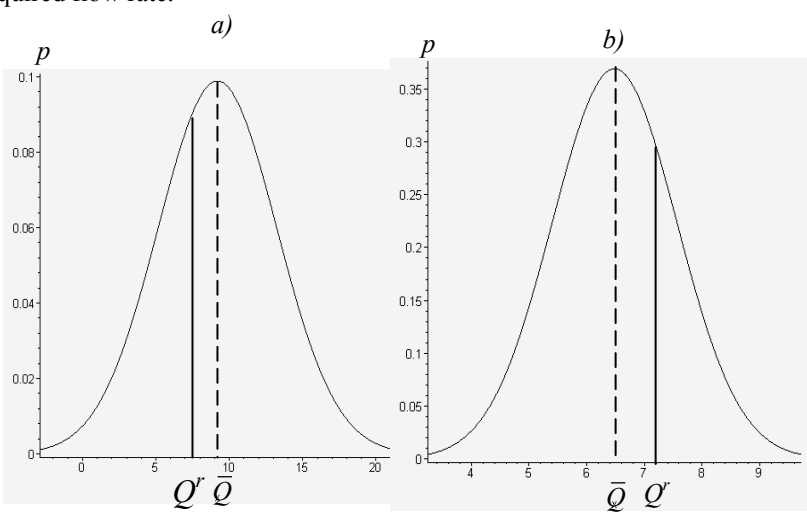


Figure 5. Illustration to the calculation of probability of providing consumer with a required flow rate: a) at node 2, b) at node 3.

$\bar{Q}$ – Calculated value of mathematical expectation of consumer flow rate considering its dependence on nodal pressure, Q^r – required value of consumer flow rate.

Conclusions.

1. The paper presents:
   - a technique for apriori calculation of statistical characteristics of a probabilistic process of the transported medium consumption as a queuing process;
   - a general scheme for probabilistic calculation of pipeline system hydraulic operating parameters. The calculation suggests determining statistical characteristics of the operating parameters by specified characteristics of boundary conditions and flow distribution model. It is shown that such a calculation is reduced to solving a traditional problem of flow distribution at the point of mathematical expectation of boundary conditions in combination with an additional procedure for calculating covariance matrices of operating parameters;
   - a technique for obtaining the analytical expressions for covariance matrices of operating parameters as well as the expressions for the general case of specifying boundary conditions;
   - a technique for probabilistic modeling of changes in the hydraulic operating parameters on the basis of developed analytical probabilistic flow distribution models. This technique provides a considerable reduction in computational efforts against the known methods of simulation modeling.
2. The suggested technique for modeling pipeline systems provides the possibility of obtaining probabilistic estimates of practically any pipeline system operating parameters that depend on operating conditions.
3. A numerical example of probabilistic calculation of the steady flow distribution in the pipeline system is given for the general case of boundary conditions. The example illustrates the suggested probabilistic approach.

References

1. Novitsky N.N., Vanteyeva O.V. Problems and methods of probabilistic modeling of hydraulic conditions of pipeline systems // Scientific technical bulletin of StPSU. – 2008. – No.1. – P. 68-75.
2. Novitsky N.N., Vanteyeva O.V. Modeling of stochastic flow distribution in hydraulic circuits // Izv RAN . Energetika. – 2011. – No.2. – P.145-154.
3. Merenkov A.P., Khasilev V.Y. Theory of hydraulic circuits. – M.Nauka, 1985. – 280 p.
4. Abramov N.N. Reliability of water supply systems M.Stroiizdat, 1979. – p.231.
5. Shopensky L.A. Study on the operating conditions of water pipelines in residential buildings: PhD dissertation thesis. M., 1968.
6. Internal sanitation of buildings (Construction Norms and Rules 2.04.01-85*) – M.: CITP Gosstroi Rossii, 2004. – 58 c.
7. Vanteyeva O.V. Probabilistic models and methods for analysis of pipeline system operating conditions. PhD dissertation. – Irkutsk, 2011. – 139 p.

An empirical study on the index of satisfaction of student allocation in the Portuguese undergraduate engineering courses

Raquel Oliveira[1], A. Manuela Gonçalves[2], and Rosa M. Vasconcelos[3]

[1]Faculty of Social Sciences, Catholic University of Portugal, Braga, Portugal (E-mail: roliveira@braga.ucp.pt)
[2]Department of Mathematics and Applications, CMAT - Centre of Mathematics of University of Minho, Portugal (E-mail: mneves@math.uminho.pt)
[3]Department of Textile Engineering, 2C2T - Centre for Textile Science and Technology, University of Minho, Portugal (E-mail: rosa@det.uminho.pt)

Abstract. This paper aims to describe and characterize student allocation in the Portuguese public higher education system, namely in the academic engineering programs. The application of multivariate methodologies for the evaluation of the students' satisfaction index detects natural clusters of academic programs and identifies some determinants of students' choices concerning higher education. These determinants may be used as explanatory variables of a model for the access to higher education engineering programs. The data used in this paper concerns the academic year of 2010/2011 and was provided by the Portuguese Ministry of Education.

Keywords: Students' index satisfaction, higher education, education policy, clusters analysis;

1 Introduction

The implementation of the Bologna Process in Portugal was led by the Portuguese Ministry of Science Technology and Higher Education (MSTHE), and it was a part of a process of reorganization and rationalization of the higher education system, OCDE[1]. The Bologna Declaration of June 19, 1999 - Joint Declaration of European Ministers of Education had, among its many objectives, the standardization of degrees conferred in the European space based on three cycles: Bachelor, Master, PhD. This process also intended the recognition and comparability of degrees, the promotion of students, teachers and researchers' mobility, in order to ensure teaching quality and the incorporation of the European dimension in higher education[2].

The MSTHE instructed the higher education institutions that they could restructure their study programs according to the Bologna principles, beginning in 2006/2007 and with a deadline extended to 2008/2009. This implementation brought about profound modifications in the Higher Education system, European Ministers of Education[3].This study aims to describe and characterize students' allocation in the academic engineering programs in the Portuguese public higher education system through the analysis of the satisfaction index of the academic programs offered by the Portuguese higher education institutions.
This document is organized as follows: in section 2 we give a description of the Portuguese higher education system with regard to organization and admission procedure; in section 3 we present the data set in consideration and the main results that characterize the available data. Section 4 concerns the statistics techniques used and the results; and finally in Section 5 are presented the conclusions and future work.

2 Portuguese higher education system

This section presents, in brief, the Higher Education system in Portugal, as well as the description of the application process to higher education.
Portugal has a binary higher education system, consisting of university and polytechnic education, each with distinct purposes that translate into specific curricular concepts, http://www.dges.mctes.pt [1][4].
University education, guided by a constant perspective of promoting research and knowledge creation, aims at ensuring a solid scientific and cultural preparation, at providing a technical training that qualifies for the exercise of professional and cultural activities and at promoting the development of design capabilities, innovation and critical analysis.
Polytechnic education, guided by a constant perspective of applied research and development aims at understanding and solving concrete problems, at providing a solid cultural and technical level, at developing the capacity for innovation, critical analysis scientific theoretical and practical in nature and its applications for the pursuit of professional activities.
As to their nature, higher education institutions may be public or private.
This study focuses on the publicly funded higher education system offering engineering study programs, since these programs include the majority of the candidates and they are also representative in terms of supply of the land area. The MSTHE (presently the MEC-Ministry of Education and Science), and more specifically the Department of Higher Education (Direção Geral do Ensino Superior-DGES), is in charge of the higher education sector and regulates access to the higher education system. Currently, the access to higher education

[1] As well as the military and police institutions, integrated in the National Network of Public Higher Education Institutions

is conditioned by system of *numerus clausus*, which defines the maximum number of students for each study program in both the public and the private sectors. This number is defined by each institution, in fixed dates, and is subject to approval of the MSTHE. *Numerus clausus* works as a restriction on the supply side of the system, affecting the size and composition of the tertiary education sector, Oliveira *et al*[5].

Access to higher education, for the public sector, is done through a national contest. The national competition is based on the students' revealed preferences in the application. Each student ranks a maximum of six study program/institution pairs, from the most preferred (the first one) to the least preferred (the last one) alternative. The ensuing nationwide competition allocates the candidates based on their grade point average and the stated ranking of preferences, Oliveira *et al*[5].

3 Methodology

The used data is available online, on a web site of the DGES[6] designed to disseminate the results of the candidates allocation in the national competition. The data was collected for the academic year of 2010/11, and the following variables are available: overall demand for each program (total number of students listing the pair institution/program among their preferences, irrespective of their ranking), as well as the number of students who have selected each program as their first choice, second choice and so forth (up to a maximum of six choices); number of vacancies available at each program in the first stage of the application process; classification of the classification of the last student allocated, number of applicants allocated in the six choices, program size (first cycle or integrated master), and the index satisfaction (ratio of the number of students allocated in the first choice and the number of vacancies available at each program).
This data set comprises 14 universities (U) and 20 polytechnics institutes (PI), from the public education sector, offering 275 academic engineering programs of which 58 are integrated masters and 217 are first cycles.

4 Data analysis and results

The used data set comprises 34 Institutions of Higher Education (IHE), universities and polytechnics institutes. The engineering courses were organized according to the National Classification of Areas of Education and Training (Classificação Nacional de Áreas de Educação e Formação-CNAEF)[7], as is shown in Figure 1.
Figure 2 displays the satisfaction index considering de type of institution.

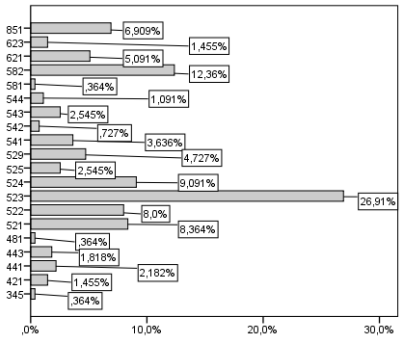


Fig. 1. Areas of Education and Training[2]

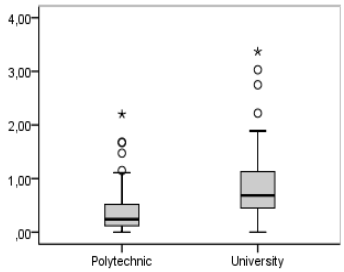


Fig. 2. Satisfaction index by type of institution

Taking into account the distribution of satisfaction index by the Areas of Education and Training of the academic engineering programs, was considered the satisfaction index mean, the results obtained are shown in Figure 3.

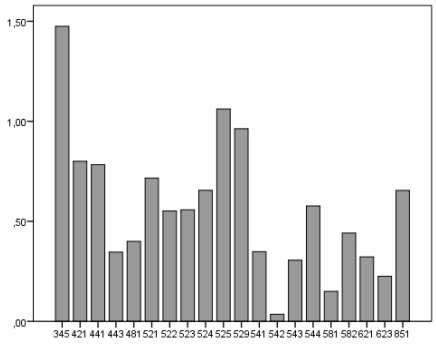


Fig. 3. Distribution of satisfaction index mean by Areas of Education and Training

A few Mann-Whitney and Kruskal-Wallis tests were performed in order to analyze if the satisfaction index depends on the type of institution, on the

[2] The Areas of Education and Training are the following: Management and Administration (345), Biology and Biochemistry (421), Physics (441), Earth Sciences (443), Computer Science (481), Metallurgy and Metalworking (521), Power and Energy (522), Electronics and Automation (523), Chemical Process Technology (524), Construction and Repair of Motor Vehicles (525), Engineering and Similar Techniques (529), Food Industries (541), Industries of the Textile, Clothing, Footwear and Leather (542), Materials (Manufacturing of Wood, Cork, Paper, Plastic, Glass and others) (543), Extractive Industries (544), Architecture and Urbanism (581), Construction and Civil Engineering (582), Agricultural and Animal Production (621), Forestry and Hunting (623), Technology of Environmental Protection (851).

program size (first cycle or integrated master) or on the Areas of Education and Training, Conover[8].
The results of the Mann-Whitney test indicates that the satisfaction index depends on the type of the HIS (test statistics= -8.367; p-value=0.00) and on the program size (test statistics= -8.069; p-value=0.00). A similar situation happens when comparing the Areas of Education and Training. Kruskal-Wallis results show that the satisfaction index also depends on the Areas of Education and Training (test statistics= 34.852; p-value=0.015).
These analyses emphasize the idea that the index of allocation satisfaction is highly related with all variables considered.
A hierarchical agglomerative cluster analysis was performed on the data satisfaction index by using the variables Areas of Education and Training and academic engineering programs, Gore[9]. Euclidean distance was used. The final result of the obtained groups was discussed according to the Ward linkage method, Barnet[10].
The obtained dendrogram and the clusters representation are shown in Figure 4.

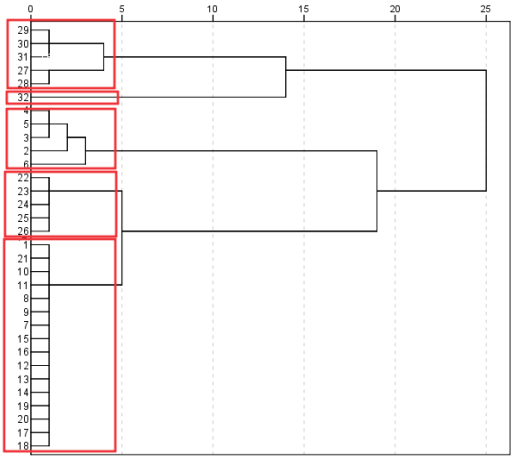


Fig. 4. Dendrogram according to Ward Method

According to the dendrogram analysis, was decided to form the clusters at a cut distance of d=4.7, thus obtaining five well-differenced clusters defined in Table 1. The resulting dendrogram has a cophenetic correlation coefficient of 0.87 (correlation coefficient between the distance matrix and the "cophenetic matrix"), which validates the clustering procedure.

Table1. Cluster identification[3]

Cluster1	27,28,28,30,31
Cluster2	32
Cluster3	2,3,4,5,6
Cluster4	22,23,24,25,26
Cluster5	1,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21

Taking into consideration the index satisfaction averages within each cluster, they are ranked in this order: Cluster 2, Cluster 5, Cluster 1, Cluster 3 and Cluster 4. The results confirm previous knowledge about the ratio existent between the number of vacancies and the number of students allocated in the first choice – satisfaction index, i.e. the cluster better classified is the one corresponding to the academic programs with fewer candidates.

Conclusions and future work

This document contributes to a better understanding of student allocation and the respective satisfaction index in the Portuguese undergraduate engineering courses. This empirical study confirms that index satisfaction is an important indicator of the determinants of students' choices concerning higher education. Although, this issue deserves a much more detailed and careful analysis and

[3]Management and Industrial Engineering (1), Biological Engineering (2), Physics Engineering (3), Geographic Engineering (4), Geological Engineering (5), Systems Engineering (6), Mechanical Engineering (7), Metallurgical and Materials Engineering (8), Industrial Engineering (9), Electromechanical Engineering/Energy Engineering/Renewable Energy Engineering/Energy and Environment Engineering (10), Electrical Engineering/Electrical and Electronics Engineering (11), Computer Engineering/Network Engineering and Systems/Electronics Maritime Systems Engineering (12), Electronics and Telecommunications Engineering/Electrical Engineering and Telecommunications/Electrical and Computer Engineering/Electrical Engineering/Electronics and Computers (13), Telematics and Computer Engineering/Computer Science and Telecommunications/Telecommunications Engineering and Computer/Network Engineering Communications/Communications Engineering (14), Mechatronics Engineering/Macro and Nanotechnologies Engineering/Computer Engineering and Medical Instrumentation (15), Metrology and Instrumentation Engineering/Automation, Control and Instrumentation Engineering (16), Chemical Engineering/Chemistry and Biochemistry/Chemistry and Biological (17), Biomedical Engineering/Bioengineering/Biotechnology (18), Aeronautical Engineering/Aerospace Engineering/Automotive Engineering/Automotive Mechanics Engineering (19), Engineering and Naval Architecture (20), Engineering and Industrial Management/Engineering Sciences/Civil Eng + Eng Electr Eng Informatics and Telec + Design + Interactive Media (21), Food Engineering /Biological and Food (22), Textile Engineering (23), Materials Engineering (24), Polymer Engineering (25), Mining and Geological Engineering/Science Engineering/Mining and Geoenvironment Engineering/Geotechnical and Geoenvironment Engineering (26), Topographic engineering (27), Civil Engineering/Civil Engineering Sciences (28), Agronomic Engineering/Engineering and Agro-Livestock (29), Zootechnical Engineering/Production Animal Engineering (30), Forestry and Natural Resources/Forestry (31), Environmental Engineering/Environmental Engineering and Management (32)

should be the subject of further research, namely extend the study applying other multivariate methodologies with more variables available.

References

1. OECD, Organization for economic co-operation and Development (2006). *Reviews of National policies for education: Tertiary education in Portugal. Examiner's Report*. Available at http://www.dges.mctes.pt/NR/rdonlyres/8B016D34-DAAB-4B50-ADBB-25AE105AEE88/2564/Backgoundreport.pdf.
2. http://www.europa.eu
3. European Ministers of Education. *The Bologna Declaration*, 1999.
4. http://www.dges.mctes.pt
5. Raquel Oliveira, A. Manuela Gonçalves and Rosa M. Vasconcelos *"A statistical model for the demand of undergraduate engineering courses in Portugal: A first study of Bologna Process impact"*, AIP Conference Proceedings, American Institute of Physics (Ed.) DOI: 10.1063/1.4756505, NUMERICAL ANALYSIS AND APPLIED MATHEMATICS ICNAAM 2012: International Conference on Numerical Analysis and Applied Mathematics, Kos, Greece, September 19-25, 1479, 1, 1724-1727, 2012.
6. http://www.dges.mctes.pt/DGES/pt/Estudantes/Acesso
7. http://www.dados.gov.pt/PT/CatalogoDados/Dados.aspx?name=ClassNacionaldeareasdeeducacaoeformacao
8. W.J. Conover. *Practical Nonparametric Statistics*. Third Edition, John Wiley & Sons: New York, 269-300, 1999.
9. P.A. Gore. *Cluster analysis*. In Handbook of applied multivariate statistics and mathematical modeling, ed. H.E.A. Tinsley and S.D. Brown, 297–321. San Diego, CA: Academic Press, 2000.
10. V. Barnett. *Interpreting Multivariate Data*. Sheffield: John Wiley & Sons, 1981.

Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013)
International Conference, Mataró (Barcelona), Spain 25 - 28 June 2013

Comparisons of multistate models with discrete-time pure-birth process for recurrent events and uncertain parameters

Eva Maria Ortega[2], Encarnacion Ortega-Quiles[3], and Jose Alonso[4]

[1] Residencia Universitaria Oblatas, Murcia, Spain
(E-mail: emariaop@gmail.com)
[2] Colegio Hnas. Carmelitas, Elche, Spain
(E-mail: salvenc@hotmail.com)
[3] Hospital Virgen Caridad, Cartagena, Murcia, Spain
(E-mail: salvenc@hotmail.com)

Abstract. This paper presents a multistate model with recurrent events that are modeled by a discrete-time pure birth process. The probability distribution of the multistate model is based on random sums with support on N and with geometric summands. Exact distributions are given in Belzunce et al. (2009) for independent inter-arrival times. These models are analyzed by stochastic orderings in the case of independent geometric, arbitrarily distributed, and positively correlated inter-arrival times. Some multivariate extensions are described. Motivations are provided in reliability, economics, biology and demography, in particular, the discrete-time pure birth shock model by Belzunce et al. (2009), where the lifetime of a system subjected to shocks (failures) arriving randomly is analyzed. Also, we provide results for some qualitative properties of the kind 'new better than used' of the multistate model assuming independent and identically distributed summands, that are illustrated.
60E15, 62P05.
Keywords: multistate models, discrete-time pure birth process, recurrent events, unobserved heterogeneity, stochastic orderings, demography.

1 Introduction and preliminaries

The multistate models have become a central tool in population biology, medicine and demography for estimating some parameters under incomplete observation or imperfect detection of individuals. Frequentist probability theory, counting processes, maximum likelihood estimation, bayesian methods, Markov chain Monte Carlo simulation are several of the approaches that have been used in the literature to study movements between states that determine the lifetime (age) of an individual, the population size of microorganisms growing or the migration transitions (see Courgeau and Lelivre (1992), Commenges (1999), Rondeau et al. (2003), among others). A multistate model is defined as a stochastic process, which at any time point remains in one of a set of discrete states. The illness-death model is the most commonly applied multistate model in population biology. For such kind of models, there are three states: healthy, illness and death. The individuals start out healthy and from this state the transitions may be to the illness or to the death. Ill individuals may die or become healthy again. Similarly, most of multistate models have three types of

states, which are determined by environmental conditions: the initial state (or the state where the individuals may enter the experimental study), the states where the individuals are studied from their transitions to move to other states, and the states where the individuals are not interesting for the experimental study since such condition is reached. Many multistate modeling concepts are connected with the Markov chain theory, where the previous states can be interpreted as initial, transient and absorbing. Theoretically, the multistate models extend the univariate survival model in two ways: parallel events (where one individual may experience multiple simultaneous events, or where several individuals that may experience an event are grouped in a cluster), and recurrent events (where a single individual may experience the same event several times). In this paper we are interested in multistate models for recurrent events data (see a review in Cook and Lawless (2007)), for instance, the fertility histories of women where the state corresponds to the number of children that a woman has given birth to at any time (or age), and from each birth she moves to the next state. There are three classical event history models: the independent increment model of Andersen and Gill (1982), the marginal model by Wei et al. (1989) and the conditional model of Prentice et al. (1981). For the marginal and conditional models, each occurrence of the event is modeled as a separate event, but for the independent increment model all the events of one individual are identical. For such event history models, a matter of interest is the choice of the time scale, as well as the risk-interval.

Some of the classical and the recent literature deals with incorporation of unobserved heterogeneity in the models for recurrent events, that arises, for example, when not all the individuals have an event before they are lost to follow-up or the study ends. Some authors have incorporated the unobserved heterogeneity by frailty modeling (see Vaupel et al. (1979), Nielsen et al. (1992), Oakes (1992) and Hougaard (2000), among others). On the other hand, modeling and analyzing the number of occurrences of the events by a stochastic process constitutes a matter of interest. For the Markov models, the transition rate only depends on the state where an individual is in, but neither on the time that an individual has been in that state nor on any other events that occurred before entering that state. Several extensions have addressed how to extend Markov models to let the transition rates depend on some features on the events occurred before entering the state, the order or occurrence of the particular state, recurrent events effects, by introducing specific duration distributions, or specific covariate effects (see e.g., Nielsen et al. (1992)).

In this paper, we introduce the discrete-time pure birth process for describing the number of recurrent events, that does not assume the Markov property, and allows that if the event occurred j times, then the probability that the recurrent event will occur depends on j. The main objective of the paper is to analyze a multistate model with recurrent events that are modeled by a discrete-time pure birth process. Its probability distribution is based on random sums with support N and with geometric summands. Specifically, let $\{U_j | j \geq 1\}$ be a sequence of independent random variables, that have the geometric distribution, with parameter α_j, with $0 < \alpha_j < 1$, for all $j \in N \cup \{0\}$ (that is, $\Pr[U_{j+1} = k] = \alpha_j \beta_j^k$, for $k \in N \cup \{0\}$, where $\beta_j = 1 - \alpha_j$, for any

$j \in N$), and let $\{M_j | j \geq 1\}$ be a sequence of positive integer-valued random variables defined on the same probability space, and furthermore, assume that the random variables M_j, $j \geq 1$ are independent of all the independent random variables U_j, $j \geq 1$. The random sums defined by $T_{M_n} = \sum_{j=1}^{M_n} U_j$ for $n = 1, ...,$ and $T_0 = 0$, have been investigated in applied probability for a long time (see e.g, Feller (1971)), with special applications in reliability theory to model the duration of systems with defective components. For instance, the sum of n geometric random variables represents the number of good items inspected before the n-th defective is observed, when the inspection is performed on an assembly line and the chance of a defective item is given initially by $\beta_1 = 1 - \alpha_1$, and subsequently (after adjustments) by $\beta_j = 1 - \alpha_j$, $j = 2, 3, ...$ Other applications of random sums and convolutions of geometric random variables can be found in queueing theory, biology, medicine, insurance and finance (see e.g., Rolski et al. (1999) and Cook and Lawless (2007)).

In the general multistate model, the random sum T_{M_n} represents the number of transitions made by an individual along his/her life. Some extensions of the model are described for incorporating unobserved heterogeneity by random event probabilities. Some motivations are provided in demography. The origin model is based on the discrete-time pure birth shock model by *Belzunce, Ortega and Ruiz (2009) Ageing properties of a discrete-time failure and repair model. IEEE Transactions Reliability, 58, 161-171*, to study the lifetime of a system (measured by the number of tasks executed by a system) which is submitted to shocks (failures) arriving randomly, and the lifetime is a random sum with geometric summands, describing the inter-arrival times of the shock arrival process. The multistate model aforementioned is analyzed by stochastic orderings in the case of independent inter-arrival times. Some of these results can be generalized in three directions: to inter-arrival times having arbitrary distribution (not necessarily geometric), to correlated inter-arrival times and to multivariate populations. The results in the case of independent inter-arrival times can be found with alternative proofs in *Belzunce, Ortega and Ruiz (2001). A note on stochastic comparisons in the discrete pure birth shock model. Technical Report, N. I-2001-17.* The results in the case of correlated inter-arrival times via random environments can be found with alternative proofs in *Escudero, Ortega and Alonso (2010). Variability comparisons for some mixture models with stochastic environments in biosciences and engineering. Stochastic Environmental Research and Risk Assessment 24, 199-209.* Some similar results to those in the case of arbitrarily distributed number of recurrent events can be found in Ortega, Alonso, Ortega (2013). Stochastic comparisons of parametric families in stochastic epidemics. *Mathematical Biosciences, doi: 10.1016/j.mbs.2012.12.006.* Several exact and numerical bounds of transformations of the probability distributions of the multistate model can be derived from the earlier stochastic comparisons. These bounds provide a statistical knowledge to select a better scenario via variations of the recurrent event distribution. In addition, we provide results for some qualitative properties of the kind 'new better than used' of the multistate model assuming independent and identically distributed summands. The organization of the paper is as follows. In Section 2 we introduce the multistate model with discrete-time pure

birth process for recurrent events. Several extensions of the origin model and some motivations in demography are provided. In Section 3 we study stochastic ordering of the multistate model with discrete-time pure birth process for recurrent events, based on results for random sums with geometric summands, and some ageing notions of the multistate model in a particular case. From now on, $X \sim \mathcal{G}(\alpha)$ denotes a geometric random variable, with parameter α, with $0 < \alpha < 1$; and *iid* is used as a shorthand for independent and identically distributed, 'increasing' and 'decreasing' stand for non-decreasing and non-increasing, respectively, and we will assume that the expectation is finite, whenever it is used. In general, we will consider a non-negative integer-valued random variable X, with survival function $\overline{P}_k = Pr[X \geq k]$, for $k \in N \cup \{0\}$, and probability mass function $p_k = \overline{P}_k - \overline{P}_{k+1}$, $k \in N$, with $p_0 = 0$ and notice that the previous definition for the discrete survival function follows the general lines for probability distributions in Roy (1997). For the definitions on stochastic ordering concepts, their implications, as well as, their main properties we will refer to Shaked and Shanthikumar (2007), and for the ageing notions we will refer to Barlow and Proschan (1981) and Lai and Xie (2006).

2 A multistate model with recurrent events

In this section, we introduce the multistate model with a discrete-time pure birth process for recurrent events, which is motivated by Belzunce et al. (2009).

2.1 Mathematical description

Consider an individual that makes sequentially transitions between the states labelled by $1, 2, ..., \omega$, and is subjected to recurrent events, which influence on the stochastic process that describes the movement of the individual. Assume that the events occur randomly in time, according to a discrete-time process $R = \{R(m), m = 1, 2, ...\}$, where $R(m)$ represents the number of events until the $m - th$ transition, for $m \in N$. We notice that the number of transitions made by the individual is considered here as the timescale. The occurrence of the recurrent events has influence time on the transitions that is negligible. The events arrival process R can be described in terms of a sequence of independent geometric interarrival times U_1, U_2, ..., where U_{j+1} is a geometric random variable, with parameter α_j, with $0 < \alpha_j < 1$. This means that if j events have occurred, then the following recurrent event will occur with probability α_j, for $j = 1, 2, ...$ Let now $\overline{P}_k$ denote the probability that the individual makes another transition having experienced the first k events, for $k \in N \cup \{0\}$. Assuming that $1 = \overline{P}_0 \geq \overline{P}_1 \geq \overline{P}_2 \geq ...$ then $\overline{P}_k$, can be considered as the survival function at $k \in N \cup \{0\}$ of an integer-valued positive random variable M, involved with making another transition when the individual has experienced a number of recurrent events M. Then, the random variable $X = \sum_{j=1}^{M} U_j$, describes the number of transitions made by the individual along his/her lifetime. Assuming that U_1, U_2, ... and M are independent, and if we denote by $T_n = \sum_{j=1}^{n} U_j$, for any $n \in N$, then $X = T_M$, is a mixture of

the family $\{T_n | n \in N\}$ with mixing distribution given by M. Throughout this paper, we call T_M *the multistate model with a discrete-time pure birth process for recurrent events.*

2.2 Some examples of multistate models in demography

The fertility histories of women are processes that the demographers have dealt with for a long time. Consider the life course of a woman through her fertility age $A = \{15, 16, ..., \omega\}$ and the multistate model for which each state is defined by her age and each birth is a recurrent event such that from each birth she moves to the next state. If we model the births according to a discrete-time pure birth process $R = \{R(m), m = 1, 2, ...\}$, where $R(m)$ represents the number of children that a woman has given birth to at age m, for $m \in A$, having geometric inter-arrival times $U_1(\alpha_0)$, $U_2(\alpha_1)$, ..., where $0 < \alpha_j < 1$ represents the probability of a birth given that the woman had j children. If $\overline{P}_k$ denotes the probability that the woman had k live children, for $k \in N \cup \{0\}$, for an integer-valued positive random variable M, then $X = \sum_{j=1}^{M} U_j$, describes the age of the woman. Conditioning on a given number of children, X is a convolution of geometric random variables. The bounds for the age of the woman can be calculated by straighforward calculation. Migration and changes in marital status are other transitions that can be studied by using the multistate models above.

2.3 Probabilistic properties of the model

The properties of the probability distribution of T_M are studied in Belzunce et al. (2009) assuming that T_M has support on N and that M and U_1, U_2,..., are independent random variables. This representation will be used in the results. We recall some distributional properties. For $M = k$, then $T_k = \sum_{j=1}^{k} U_j$, $k \in N$, and $\{R(m) \leq k\}$ if and only if $\{T_{k+1} \geq m\}$. We will use the notation

$$z_k(m) = \Pr\left[R(m) = k\right] = \Pr\left[T_k < m \leq T_{k+1}\right].$$

The function $z_k(m)$ is a totally positive in $(k, m) \in N \times N \cup \{0\}$ (see Karlin (1968)), and a recursive equation for computing $z_k(m)$ is given in Lemma 6 in Belzunce et al. (2009). Assuming that T_M has support on N and that M and U_1, U_2,..., are independent random variables, an exact distribution in terms of $z_k(m)$ can be stated, for $m \in N$, the survival function for T_M is given by

$$\overline{H}(m) = \Pr[T_M \geq m] = \sum_{k=0}^{\infty} z_k(m)\overline{P}_k. \tag{1}$$

Some extensions of the earlier multistate model can be derived in several ways. To introduce the heterogeneity, first we consider that the parameters α_j are random for any $j \in N$ (see Escudero et al. (2010)). Then the random variable that describes the number of transitions made by the individual along his/her lifetime is a random sum by $X = T_M(\alpha_1, ..., \alpha_m, ...) = \sum_{j=1}^{M} \mathcal{G}(\alpha_j)$, and conditioning on $M = m$ it is a mixture of $\{T_m = \sum_{j=1}^{m} \mathcal{G}(\alpha_j) | (\alpha_1, ..., \alpha_m) \in (0, 1)^m\}$ with mixing distribution given by $(\alpha_1, ..., \alpha_m)$. Extensions of the multistate model in the multivariate case will be discussed in Section 3.

3 The results

In this section we provide results on stochastic comparisons of multistate models with discrete-time process for recurrent events. In the following results let M, N be positive integer-valued random variables and $U_1(\alpha_1)$, $U_2(\alpha_2)$,... be a sequence of independent or positively correlated random variables, that are independent of M and of N. Consider the random sums $T_M = \sum_{j=1}^{M} U_j(\alpha_j)$ and $T_N = \sum_{j=1}^{N} U_j(\alpha_j)$ to describe multistate models with a counting process for recurrent events, under two scenarios determined by the distribution of the number of recurrent events M and N, respectively, experienced before making a new transition.

3.1 Some theoretical results

As a starting point, we give a discrete version of a classical result for absolutely continuous random variables.

Theorem 1. *Let $U_1 \sim \mathcal{G}(\alpha_1)$, $U_2 \sim \mathcal{G}(\alpha_2)$,... be a sequence of independent geometric random variables. Denote $\beta_n = 1 - \alpha_n$ and $T_n = \sum_{j=1}^{n} U_j$, for any $n \in N$. The following assertions hold:*
(i) if $\beta_i \leq \beta_{i+1}$, for any $i \in N$ then $U_i \leq_{LR} U_{i+1}$, for any $i \in N$.
(ii) $T_n \leq_{LR} T_{n+1}$, for any $n \in N$.

3.2 The case of independent inter-arrival times.

Next, we obtain results for several stochastic orderings. Some related results can be found in Ross and Schechner (1984), Kochar (1990), Pellerey (1993), and Ortega (2009), among others.

Theorem 2. *Let $U_1 \sim \mathcal{G}(\alpha_1)$, $U_2 \sim \mathcal{G}(\alpha_2)$,... be a sequence of independent geometric random variables, that are independent of M and of N. Consider the random sums $T_M = \sum_{j=1}^{M} U_j(\alpha_j)$ and $T_N = \sum_{j=1}^{N} U_j(\alpha_j)$.*
i) If $M \leq_{LR} N$, then $T_M \leq_{LR} T_N$.
ii) If $M \leq_{FR} N$, then $T_M \leq_{FR} T_N$.
iii) If $M \leq_{RHR} N$, then $T_M \leq_{RHR} T_N$.
iv) If $M \leq_{ST} N$, then $T_M \leq_{ST} T_N$.
v) If $\beta_i \leq \beta_{i+1}$, for any $i \in N$ and $M \leq_{ICX} N$, then $T_M \leq_{ICX} T_N$.
vi) If $\beta_i \leq \beta_{i+1}$, for any $i \in N$; and $M \leq_{Lt} N$, then $T_M \leq_{Lt} T_N$.
vii) If $\left(\sum_{j=k}^{\infty} \frac{\beta_j}{\alpha_j} \overline{P}_j\right) / \left(\sum_{j=k}^{\infty} \frac{\beta_j}{\alpha_j} \overline{Q}_j\right)$ is decreasing in $k = 0, 1, ...$, $T_M \leq_{MR} T_N$.

The previous results provide technical conditions to be checked for ranking multistate models. Also, these conditions enable us to construct analytical bounds for the multistate models based on parametric families used to describe the survival for recurrent events.

3.3 The case of arbitrarily distributed number of recurrent events

In this subsection we deal with the case of independent inter-arrival times having arbitrary distribution (not necessarily geometric) and having the ILR property, for which the main assertions in Theorem 2 also hold.

Theorem 3. *Consider the random sums* $T_M = \sum_{j=1}^{M} U_j(\alpha_j)$ *and* $T_N = \sum_{j=1}^{N} U_j(\alpha_j)$, *where the summands are independent random variables with an arbitrary distribution.*
i) If U_1, U_2,... *have the* ILR *property, if* $M \leq_{LR} N$, *then* $T_M \leq_{LR} T_N$.
ii) If U_1, U_2,... *have the* IFR *property, if* $M \leq_{LR} N$, *then* $T_M \leq_{FR,RHR} T_N$.
iii) If U_1, U_2,... *have the* IFR *property, if* $M \leq_{MR} N$, *then* $T_M \leq_{MR} T_N$.

For analyzing the effect of unobserved heterogeneity given by random parameters, we have a random sum of mixed geometric random variables that are independent of the number of summands. The stochastic comparisons of mixed geometric random variables are studied in Ortega, Alonso and Ortega (2013).

3.4 The case of correlated inter-arrival times

Analogously, for arrival times being increasing in the ST order, the assertion in Theorem 2 holds too, and the same happens for the CX, and Lt orders. Similar results for the LR and FR orders can be given.

Theorem 4. *Let* $U_1 \sim \mathcal{G}(\alpha_1)$, $U_2 \sim \mathcal{G}(\alpha_2)$,... *be a sequence of geometric random variables. Consider* $T_M = \sum_{j=1}^{M} U_j(\alpha_j)$, $T_N = \sum_{j=1}^{N} U_j(\alpha_j)$.
i) If $U_j \leq_{CX} U_{j+1}$ *for any* $j = 1, 2, ...$ *and* $M \leq_{CX} N$, *then* $T_M \leq_{CX} T_N$.
ii) If $U_j \leq_{ST} U_{j+1}$ *for any* $j = 1, 2, ...$ *and* $M \leq_{ICX} N$, *then* $T_M \leq_{ICX} T_N$.
iii) If $U_j \leq_{Lt} U_{j+1}$ *for any* $j = 1, 2, ...$ *and* $M \leq_{Lt} N$, *then* $T_M \leq_{Lt} T_N$.

When the correlated inter-arrival times are dependent via random environments the ICX order is studied in Escudero et al. (2010). In addition, we observe that the results in Escudero et al. (2010) can be used to compare two multistate models with discrete-time pure birth process for recurrent events, with different arrival processes (i.e., different parameters for the inter-arrival time distributions) and with the same distribution of the number of experienced recurrent events. They extend the results in the literature, of stochastic comparisons for convolutions of geometric random variables, from some related majorization type orders between two parameter vectors of the geometric summands (see e.g., Boland et al. (1994)).

3.5 The case of multivariate multistate models

The earlier setting can be extended to the multivariate case in several ways. A first extension is determined by probabilities of recurrent events that depend on the state of the process apart of the number of previous events. Given that j events have occurred, then a new recurrent event happens with probability $\alpha_{j,k}$, where k represents that the process is in the state k. A second extension

comes from considering random vectors of environmental or exposure parameters for the components of the models, as in Escudero et al. (2010). Third, by taking random vectors of multistate models and using multivariate stochastic orderings, for which similar results as in Belzunce et al. (2006) can be derived. The results in this issue are omitted.

3.6 The case of discrete-time Poisson process for recurrent events

Consider a sequence of *iid* geometric random variables where $U_j \sim \mathcal{G}(\alpha)$, $j \in N$, with $0 < \alpha < 1$. Then $T_k = \sum_{j=1}^{k} U_j$ has the negative binomial distribution with parameters (k, α), with $k \in N$. Unconditioning T_M is a random sum with *iid* geometric summands, and assume that $\Pr[T_M = 0] = 0$, i.e., T_M has on values over N. For $\beta = 1 - \alpha$, the survival function of T_M is

$$\overline{H}(m) = \sum_{k=0}^{\infty} \alpha^k \beta^m k + m - 1 k \overline{P}_k, m \in N. \tag{2}$$

Next, we study the ageing properties NBU, $NBU(2)$, and $NBUC$ for the multistate model when the probability of occurring another recurrent event does not depend on the previous events, that is, in the case when there are identically distributed inter-arrival times. The dual notions arise analogously.

Theorem 5. *Consider the random sum $T_M = \sum_{j=1}^{M} U_j$ given by (2), where the summands are iid geometric random variables, that are independent of M.*
i) If M is NBU, then T_M is NBU.
ii) If M is $NBU(2)$, then T_M is $NBU(2)$.
iii) If M is $NBUC$, then T_M is $NBUC$.

The NWU notion is studied in Theorem 4.1 in Willmot and Cai (2001), and also in Li et al. (2006) and Belzunce et al. (2009) for several kind of discrete random sums, under other assumptions. Using some inequalities for these ageing notions as can be seen in Barlow and Proschan (1981), some numerical upper and lower bounds for the number of transitions of an individual can be calculated. Statistical inference can be used to test these ageing notions (see Lai and Xie (2006)). From the migration movements in demography, the previous results mean that the number of migrations to be completed by an individual in his/her life are larger, in some stochastic sense, than the number of migrations that remain to be completed by individuals who have carry out a given amount of movements. These bounds can be used to develop social and economical policies concerning with possible massive migration movements.

Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013)
International Conference, Mataró (Barcelona), Spain 25 - 28 June 2013

Optimal Insurance with Investment and Loans

Darya Ostrovskaya[1]

Department of Probability Theory, Lomonosov Moscow State University, Russia
(E-mail: ostrovskaya.dar@gmail.com)

Abstract. Insurance plays the leading role in compensation for losses, and its necessity increases together with the development of the economy. Investment activity is also one of the main factors for effective operating of the insurance company. The paper under consideration is devoted to the search of the optimal investment strategy for an insurance company. Several discrete models are investigated by means of dynamic programming and Bellman equation.

1 Description of the model with equal rates for investment and borrowing

We consider the discrete-time model connected to the operating of some insurance company.
Let x be initial capital of the insurance company. In a unit interval the company earns some premia c and has to pay possible claims η_i. We let $\{\eta_i\}_{i=1}^n$ be an iid sequence with distribution function $F(t)$ and density function $f(t)$. Since claims are nonnegative, let $F(t) = 0$ and $f(t) = 0$ for $t < 0$ and suppose that $F(t)$ increases in $[0; \infty)$.
In the beginning of every period after collecting premia the company is supposed to be able to invest or to borrow some value y with some interest rate r. In the end of the period company gets claims and if it does not have enough capital to pay the claims, then it borrows lacking money immediately at a higher rate q ($q > r$). The aim of the company is to find the optimal value of investment (or borrowing) y_n^* for every $n = 1, 2, 3, ...$ in order to minimize expected costs for n periods. If y_n^* obtained is positive, then the company should borrow y_n^*, otherwise it should invest $-y_n^*$.

1.1 One-period strategy

Let us discuss optimal strategy for a company, which operates for one period only Let ξ be expected costs of the company with initial capital x, then

$$\xi = \begin{cases} yr, & \eta \le x + c + y; \\ yr + q(\eta - (x + c + y)), & \eta > x + c + y. \end{cases}$$

Denote by $P(x, y, \eta)$ capital of the company with initial capital x in the end of the first period, provided that the value y was borrowed (or $-y$ was invested).

$$P(x, y, \eta) = \begin{cases} x + c - \eta - yr, & \eta \le x + c + y; \\ -y(r + 1) - (1 + q)(\eta - (x + c + y)), & \eta > x + c + y. \end{cases}$$

Let us find the expectation of the costs for one period:

$$G(x,y) = E\xi = yr + \int_{x+c+y}^{\infty} q(s-(x+c+y))f(s)\,ds.$$

Having computed the derivative of this expression

$$G'_y(x,y) = r - q + qF(x+c+y),$$

we equate it to zero, and find the optimal solution:

$$F(x+c+y) = \frac{q-r}{q};$$

$$y_1^*(x) = F^{-1}\left(\frac{q-r}{q}\right) - x - c.$$

The second derivative of expected costs

$$G''_y(x,y) = qF'(x+c+y) = qF'\left(F^{-1}\left(\frac{q-r}{q}\right)\right) > 0,$$

is positive since $F(t)$ is increasing function. Therefore, the function of expected costs reaches its minimum in the obtained point $y_1^*(x)$

1.2 Two-period strategy

Let $f_n(x)$ be expectation of costs for n periods provided that value of y_i^* was chosen optimally in each period.
According to Bellman optimality principle we derive following equation:

$$f_n(x) = \min_y \left[G(x,y) + Ef_{n-1}\left(P(x,y,\eta)\right)\right].$$

Consider this equation for $n = 2$:

$$f_2(x) = \min_y \left[G(x,y) + Ef_1\left(P(x,y,\eta)\right)\right].$$

$f_1(x)$ may be found since we know one-period optimal solution:

$$f_1(x) = G(x, y_1^*(x)) = r\left(F^{-1}\left(\frac{q-r}{q}\right) - x - c\right) +$$

$$+ \int\limits_{F^{-1}\left(\frac{q-r}{q}\right)}^{\infty} q\left(s - F^{-1}\left(\frac{q-r}{q}\right)\right) f(s)\,ds.$$

Hence $f_1'(x) = -r$;

$$\left[f_1\left(P(x,y,\eta)\right)\right]'_y = f_1'\left(P(x,y,\eta)\right) \cdot P'_y(x,y)$$

$$P'_y(x,y,\eta) = \begin{cases} -r, & \eta \le x+c+y; \\ -(r+1)+(1+q), & \eta > x+c+y. \end{cases}$$

Having computed the derivative of this expression

$$[G(x,y) + Ef_1(P(x,y,\eta))]'_y = (r-q)(1+r) + (1+r)qF(x+c+y).$$

we equate it to zero, and find the optimal solution:

$$F(x+c+y) = \frac{(q-r)(1+r)}{q(1+r)}.$$

$$y_2^*(x) = F^{-1}\left(\frac{q-r}{q}\right) - x - c = y_1^*(x).$$

Again the second derivative of expected costs

$$[G(x,y) + Ef_1(P(x,y,\eta))]''_y = (r+1)qF'(x+c+y_2^*) = (r+1)qF'\left(F^{-1}\left(\frac{q-r}{q}\right)\right) > 0$$

is positive. Thus, the function of expected costs reaches its minimum in the obtained point $y_2^*(x)$

1.3 n-period strategy

Proposition 1 *$y_n^*(x) = F^{-1}\left(\frac{q-r}{q}\right) - x - c$ for every $n = 1, 2, 3, \ldots$.*

Proof is by induction. Suppose that for all $n \le k-1$ $y_n^*(x) = F^{-1}\left(\frac{q-r}{q}\right) - x - c$ and $f'_{n-2}(x) = 1-(1+r)^{n-2}$. We will prove that it holds for $n = k$. The recurrence equation will be as follows:

$$f_k(x) = \min_y [G(x,y) + Ef_{k-1}(P(x,y,\eta))];$$

$$[f_{k-1}(P(x,y,\eta))]'_y = f'_{k-1}(P(x,y,\eta)) \cdot P'_y(x,y,\eta);$$

$$f_{k-1}(x) = G(x, y_{k-1}^*(x)) + Ef_{k-2}\left(P(x, y_{k-1}^*(x), \eta)\right);$$

$$f'_{k-1}(x) = G'_x(x, y_{k-1}^*(x)) + G'_y(x, y_{k-1}^*(x)) \cdot {y_{k-1}^*}'(x) + E[f_{k-2}(P(x, y_{k-1}^*(x), \eta))]'_x;$$

$$G'_y(x, y_{k-1}^*(x)) = G'_y(x, y_1^*(x)) = 0,$$

since $y_1^*(x)$ is a solution of the equation $G'_y(x,y) = 0$.

$$G'_x(x, y_{k-1}^*(x)) = G'_x(x, y_1^*(x)) = -q(1 - F(x+c+y_1^*(x))) = -r;$$

$$\left[f_{k-2}(P(x, y_{k-1}^*(x), \eta))\right]'_x = f'_{k-2}(P(x, y_{k-1}^*(x), \eta)) \cdot$$
$$\cdot \left(P'_x(x, y_{k-1}^*(x), \eta) + P'_y(x, y_{k-1}^*(x), \eta) \cdot {y_{k-1}^*}'(x)\right).$$

Let us find the derivatives of the capital in the end of the first period:

$$P'_x(x, y_{k-1}^*(x)) = P'_x(x, y_1^*(x), \eta) = \begin{cases} 1, & \eta \le F^{-1}\left(\frac{q-r}{q}\right); \\ 1+q, & \eta > F^{-1}\left(\frac{q-r}{q}\right). \end{cases}$$

$$P'_y(x, y^*_{k-1}(x), \eta) = P'_y(x, y^*_1(x), \eta) = \begin{cases} -r, & \eta \le F^{-1}\left(\frac{q-r}{q}\right); \\ q-r, & \eta > F^{-1}\left(\frac{q-r}{q}\right). \end{cases}$$

By the inductive assumption, ${y^*_{k-1}}'(x) = {y^*_1}'(x) = -1$.
Therefore, we find

$$E\left[f_{k-2}(P(x, y^*_{k-1}(x), \eta))\right]'_x = \left(1 - (1+r)^{k-2}\right) \cdot$$

$$\cdot \left(\int\limits_0^{F^{-1}\left(\frac{q-r}{q}\right)} (1+r) f(s)\, ds + \int\limits_{F^{-1}\left(\frac{q-r}{q}\right)}^{\infty} (1+q+r-q) f(s)\, ds \right) =$$

$$= \left(1 - (1+r)^{k-2}\right) \cdot \left((1+r)\frac{q-r}{q} + (1+r)\left(1 - \frac{q-r}{q}\right)\right) =$$

$$= \left(1 - (1+r)^{k-2}\right) \cdot (1+r);$$

$$f'_{k-1}(x) = -r + \left(1 - (1+r)^{k-2}\right) \cdot (1+r) = 1 - (1+r)^{k-1};$$

Using all the expressions above we may find:

$$\left[G(x,y) + Ef_{k-1}\left(P(x,y,\eta)\right)\right]'_y = (r-q)(1+r)^{k-1} + qF(x+c+y)(1+r)^{k-1}.$$

By equating this derivative to zero and solving the following equation

$$F(x+c+y) = \frac{q-r}{q}.$$

we obtain $y^*_k = F^{-1}\left(\frac{q-r}{q}\right) - x - c$.
The second derivative:

$$\left[G(x,y) + Ef_{k-1}\left(P(x,y,\eta)\right)\right]''_y = q(1+r)^{k-1} \cdot F'(x+c+y) > 0,$$

is positive since $F(t)$ is increasing function. Thus $\mathrm{y}^*_k isaminimumpoint. Henceinductivestepisdoneandthe Propo$

2 Description of the model with different rates for investment and borrowing

In this section we assume that rates for borrowing and investment differ, namely r is rate for borrowing, v is rate for investment $(r > v)$

2.1 One-period strategy

In this model the capital of the insurance company in the end of the first period is:

$$P(x,y,\eta) = \begin{cases} x+c-\eta-yr, & \eta \le x+c+y, & y > 0; \\ x+c-\eta-yv, & \eta \le x+c+y, & y \le 0; \\ -y(r+1) - (1+q)(\eta - (x+c+y)), & \eta > x+c+y, & y > 0; \\ -y(v+1) - (1+q)(\eta - (x+c+y)), & \eta > x+c+y, & y \le 0. \end{cases} =$$

$$= \begin{cases} x+c-\eta-y(I(y>0)r+I(y\le 0)v), & \eta \le x+c+y; \\ -y(I(y>0)r+I(y\le 0)v+1)-(1+q)(\eta-(x+c+y)), & \eta > x+c+y. \end{cases}$$

The costs are as follows:

$$\xi = \begin{cases} yr, & \eta \le x+c+y, \quad y>0; \\ yv, & \eta \le x+c+y, \quad y\le 0; \\ yr+q(\eta-(x+c+y)), & \eta > x+c+y, \quad y>0; \\ yv+q(\eta-(x+c+y)), & \eta > x+c+y, \quad y\le 0. \end{cases} =$$

$$= \begin{cases} y(I(y>0)r+I(y\le 0)v), & \eta \le x+c+y; \\ y(I(y>0)r+I(y\le 0)v)+q(\eta-(x+c+y)), & \eta > x+c+y. \end{cases}$$

Let us find the expectation of the costs:

$$G(x,y) = E\xi = \int_0^{x+c+y} y\,(I(y>0)r+I(y\le 0)v)\,f(s)\,ds +$$

$$+ \int_{x+c+y}^{\infty} [y\,(I(y>0)r+I(y\le 0)v)+q(s-(x+c+y))]\,f(s)\,ds =$$

$$= y\,(I(y>0)r+I(y\le 0)v) + \int_{x+c+y}^{\infty} q(s-(x+c+y))f(s)\,ds.$$

In order to find extremums we need to find the derivative with respect to y:

$$G'_y(x,y) = (I(y>0)r+I(y\le 0)v)+q\,(F(x+c+y)-1)\,;$$

$$G''_y(x,y) = qF'(x+c+y) > 0,$$

Hence $G(x,y)$ is a convex function. Let us consider two cases:
1) $y>0$. Then, having equated derivative to zero:

$$r-q+qF(x+c+y)=0,$$

we obtain

$$y_1^* = F^{-1}\left(\frac{q-r}{q}\right) - x - c.$$

$$x < F^{-1}\left(\frac{q-r}{q}\right) - c.$$

2) $y\le 0$.

$$v-q+qF(x+c+y)=0;$$

$$y_1^* = F^{-1}\left(\frac{q-v}{q}\right) - x - c.$$

$$x \ge F^{-1}\left(\frac{q-v}{q}\right) - c.$$

Taking into account that $r > v$, notice that $F^{-1}\left(\frac{q-v}{q}\right) - c > F^{-1}\left(\frac{q-r}{q}\right) - c$. Then we need to obtain optimal value y, when $F^{-1}\left(\frac{q-v}{q}\right) - c > x \geq F^{-1}\left(\frac{q-r}{q}\right) - c$. With such initial capital the function of expected costs is decreasing for $y < 0$ and increasing for $y \geq 0$, hence function of expected costs reaches minimum at $y_1^* = 0$.
Summarizing the results we may write:

$$y_1^*(x) = \begin{cases} F^{-1}\left(\frac{q-r}{q}\right) - x - c, & x < F^{-1}\left(\frac{q-r}{q}\right) - c; \\ F^{-1}\left(\frac{q-v}{q}\right) - x - c, & x \geq F^{-1}\left(\frac{q-v}{q}\right) - c; \\ 0, & F^{-1}\left(\frac{q-v}{q}\right) - c > x \geq F^{-1}\left(\frac{q-r}{q}\right) - c. \end{cases}$$

2.2 Two-period strategy

In this subsection I would like to provide some graphs illustrating the behaviour of the derivative in case of exponential distribution function $F(t)$ depending on the values of premium c and initial capital x. The derivative on every interval is computed. There may be three qualitatively different situations.

1. The graph of the derivative intersects negative anxle. Thus, to reach minimal costs the company is to invest some value.

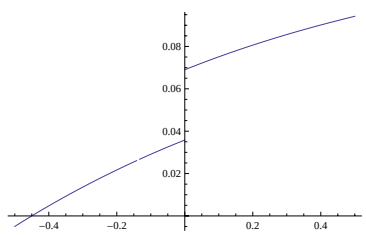


Fig. 1.

2. The graph of the derivative intersects positive anxle. Thus, to reach minimal costs company is to borrow some value.

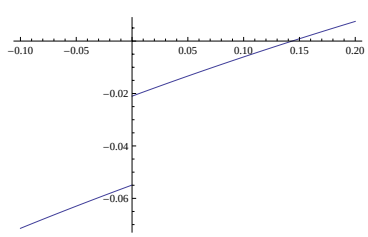


Fig. 2.

3. The graph of the derivative does not intersect anxle anywhere. In this case function of expected costs is decreasing while $y < 0$ and increasing while $y > 0$. Therefore minimum of the function of expected costs is at point $y = 0$.

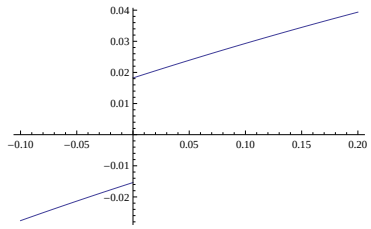


Fig. 3.

References

1.R. Bellman. *Dynamic Programming*, 1957

2.Zhongfei Li, Ken Seng Tan, Hailiang Yang. *Multiperiod Optimal Investment-Consumption Strategies with Mortality Risk and Environment Uncertainty*, 2006

Estimating child mortality from information on previous birth: data from a Portuguese birth cohort.

Neir A. Paes[1]; Cristina Teixeira[2]; Henrique Barros[3]

[1] Department of Statistics of the Federal University of Paraíba, Cidade Universitária, João Pessoa, PB, Brazil (E-mail: **antunes@de.ufpb.br**)

[2,3] Institute of Public Health, University of Porto, Porto, Portugal and Department of Clinical Epidemiology, Predictive Medicine and Public Health, University of Porto Medical School, Porto, Portugal (E-mail: **cristina.teixeira@ipb.pt; hbarros@med.up.pt**).

Abstract: A whole range of techniques have been developed for estimating infant and child mortality from the information routinely recorded in maternity registers on age and reproductive history of delivering mothers. Among them, the best-known and most widely applied indirect technique was developed by William Brass. Using the baseline information from a birth cohort (Generation XXI) assembled in the period 2005-2006 in the metropolitan area of Porto, North of Portugal, we addressed in this paper the following objectives: (1) to estimate the child mortality rates based on information about previous births and their survivorship among women recruited for this birth cohort; (2) to compare these indirect estimates with direct values retrieved from Portuguese vital statistics; (3) to analyse the potential of this birth cohort to generate plausible estimates of life-table indicators. We retrieved data on mother's age, previous live births prior to the current one, and number of surviving and deceased children from a group of multiparous women (n=3521). The data was divided into seven 5-year groups by maternal age and survival and death probabilities were computed for each group. Through the Brass method, we obtained estimates of probability of dying before attaining certain exact childhood ages, $q(x)$, by using the multipliers $k(i)$ as proposed by Trussel. Then, a logit life-table system was used to derive life-table indicators. Accordingly, probabilities of dying between birth and 2, 3, 5 and 10 years were respectively: 4.1; 6.7; 9.4 and 13.1 per 1,000 children ever born, which were alocatted in time-period. These indirect estimates compared with the direct ones obtained from Portuguese vital statistics revealed that they were very similar. The life expectancy at birth was 77.6 years for both

sexes, and the implied infant mortality was 4.0 per 1000 live births for Grand Porto during the period 2005-2006. The use of indirect method to analyze the potential of the Generation XXI cohort data in provide relevant information on reproductive issues, not available in the vital statistics, seems to be an important and effective tool, and promissory for analyzing the follow-up studies of this cohort held in 2009 and 2012.

Keywords: Child mortality, Brass' Method, Preceding Birth Technique.

1 Introduction

The 5th stage of demographic transition, which many call postmodern demographic regime, already occurs in some of the more developed countries of the world where death rates, although low, begin to exceed birth rates, which results in a negative crude rate of natural increase [1]. Portugal went through all stages of the Demographic Transition Model. According to the Portuguese vital statistics [2], since 2007 the number of deaths (n=103,512) has exceeded the number of live births (n=102,492). In the recent years Portugal went into the 5th phase; the fertility rate is around 1.3 and infant mortality rate reached 3.4 per 1000 live births which are among the lowest rates in Europe [3].

Tracing the evolution of child health indicators is essential for evaluating the availability and quality of health care in a community. Infant and child mortality [4,5] have been used for this purpose. The accurate estimation of the probabilities of dying before certain age, particularly under-five years old, are largely used by international agencies to monitor development progress [6] and, is one of the principal input parameters used to develop estimates of life expectancy at birth and other summary indicators of mortality [7,8].

In the face of dramatic reduction in absolute level of infant and child mortality in developed countries, the challenge is to document disparities across groups concerning such deaths. The evaluation of potential efficacious interventions to reduce adverse neonatal and childhood outcomes should take into account a set of characteristics, namely demographic, social and obstetric factors in order to

perceive the decline of disparities rather than the decrease of crude mortality rates. The estimates of demographic and health indicators available in Portugal are based on census data and vital statistic registration [2]. From the latter the infant and child mortality rates can be directly obtained. However, Portuguese vital statistics has no information about a set of demographic characteristics such as parity, social factors as education, adverse maternal behaviours and obstetric complications; medical records are usually more detailed than vital statistics concerning such information.

A whole range of indirect techniques have been developed for estimating infant and child mortality based on the information retrieved from medical records of women delivering a child, or based on sample surveys, on age and reproductive history of delivering mothers [7-9]. These indirect estimation techniques could be useful to provide childhood mortality rates when a set of maternal characteristics is not available in vital statistics. They also offer potential estimates for continuous follow-up studies over time.

One of the largest birth cohort ever assembled in Portugal was established between 2005 and 2006 in the Porto Metropolitan Area, North of Portugal (Generation XXI). We addressed in this paper the following purposes: (1) to estimate the child mortality rates based on information about previous births and their survivorship among women recruited for this birth cohort; (2) to compare these indirect estimates with direct values retrieved from Portuguese vital statistics; (3) to analyse the potential of this birth cohort to generate plausible estimates of life-table indicators.

2 Methods

In Portugal, nearly all deliveries occur within hospitals and 90% of them occur in public hospitals free of charge for all childbearing women and their offspring. The participants of the present study were recruited in five public hospitals level III, while assembling a birth cohort in Porto Metropolitan Area, in the north of Portugal (Generation XXI). Between April 2005 and August 2006, 70% of all

pregnant women delivered at those five public hospitals were invited as participants on the basis of "first come first served" and only 8% of those invited refused to participate. This approach allowed a representative sample. The final sample comprised 8495 women who delivered live infants (>24 weeks). Information on social and demographic characteristics, obstetric and gynaecological history, lifestyles and current pregnancy events was obtained using a structured questionnaire. Individual interviews were performed 24 to 72 hours after delivery by trained interviewers. Information on pregnancy complications, delivery circumstances and data on newborn characteristics were abstracted from patient medical records.

From all mothers enrolled in this cohort only multiparous women with at least a previous child before the current delivery (n=3520) were included in the present analysis. We retrieved data on mother's age, previous live births prior to the current one and number of deceased children.

In this work we used the method developed by William Brass [7], the best-known and most widely applied indirect technique for estimating child mortality rates. Accordingly, women were stratified by five-year age groups from 15-19 to 45-49. The procedure converts proportions dead children ever born, *D(i),* reported by women's age group into estimates of the probability of dying before attaining certain exact childhood age, $q(x=1, 2, 3, 5, 10, 15$ and 20 years old), by using the multipliers *k(i)* as proposed by Palloni-Heligman, assuming that Far Eastern family in the United Nations model life tables system is an adequate representation of the pattern of mortality of Porto [7].

We computed also the reference time-period, $t(x)$, which represents an estimate of the numbers of years before the survey date to which the child mortality estimates, $q(x)$, refer. The values of the coefficients to estimate $t(x)$ were retrieved from the United Nations Report [7].

We assumed that the pattern of fertility by age of the women and the childhood mortality have remained without important changes during the recent past. Indeed, between 2000 and 2006 the total fertility rate varied from 1.36 and 1.41 live births per woman and infant mortality rate varied between 5.5 and 3.6 infant

deaths per 1000 live births. We assumed also a similar pattern of fertility by age of the multiparous in the Generation XXI and the Portuguese women as a whole. We obtained the direct estimates of child mortality from Portuguese vital statistics [2], taking into account the reference period and we compared the values of direct and indirect estimates.

Once $q(x)$ is estimated, its complement $l(x)$, the probability of surviving from birth to exact age x, is readily obtained as $l(x) = 1.0 - q(x)$, which was converted in logit function, $Y(x)$. According to the logit life-table system as proposed by Brass [7], the Portugal (2009-2011) life table was used as a standard model to derive the adjusted life table to Porto for ages under 10 [2].

3 Results and Discussion

Table 1 shows the basic data and the main results on the application of the Brass's method according to five-year age group of mother. In Portugal, increasing numbers of women are delaying childbearing into their thirties and early forties, making the consequences of older maternal age for the infant an important public health concern. Grand multiparity, by contrast, is now exceedingly rare in this country. There were 3521 multiparous women in the present analysis; only 24% (n=829) of those women reported to have 2 or more children previously to the current birth. The mean age of women was 32 years (standard deviation ±5.26). The total number of children ever born was 4651; 1.2% (n=55) of those children died.

Proportions deceased are quite instable in age group 15-19 due to the very small numbers of births. On the other hand, for the last age group, no dead was reported, and the proportion of the age group 40-44 is out of line probably due to the age composition of the sample. Therefore, mortality estimates based on the reports of women for these age groups were disregarded. Proportions deceased were also instable in age group 25-29 and 35-39, which demanded some adjustments. Since there is no evidence that the proportion of death

provided by the age group 20-24 and 30-34 are unreliable, the obtained estimates for the adjacent age groups (25-29 and 35-39) was performed by linear interpolation. As presented in Table 1, the indirect estimates of the probability of dying (mortality rates) between birth and age 2, 3, 5 and 10 years old were 4.0, 5.7, 7.3 and 9.6 per 1000 children ever born, respectively. Assuming that Porto has a little lower mortality rates compared with the country as whole, and considering the reference time-period to which child mortality $q(x)$ refer, in general the mortality rates presented similar values when indirect and direct estimates of child mortality are compared.

Calculation of logit transformation of the estimated survivorship probabilities, $l(x)$, and the corresponding logit tranformation of the standard life-table used (Portugal 2009-2010) showed a reasonable degree of coincidence between one set and the other for multiparous womem. Notably deviant points were those associated with the $l(5)$ and $l(10)$ estimate. On the basis of these observations, the level indicated to match with the standard life-table was $q(2)$, which correspond to a more recent period. In this way, the adjusted pattern was obtained by assuming an value for a level of $\alpha = 0.050$ and standard $\beta = 0.979$ according to the Brass's logit life table system.

Table 1. Data required to compute child mortality according to the Brass' method and indirect and direct estimates of mortality rates under 2, 3, 5 and 10 years old to Porto Metropolitan Area, Portugal.

Women's Age Group (years)	Average parity per women	Number of children ever born	Proportion of children dead	Index x	Generation XXI		Portuguese vital statistics	
					Prob. of dying	Reference time-period	Prob. of dying	Reference time-period
15 – 19	1,0000	35	0.0286	$q(1)$	-	-	-	-
20 – 24	1,1682	382	0.0052	$q(2)$	0,0040	2001	0,0057	2000/01
25 – 29	1,2247	894	0.0179	$q(3)$	0,0057	1999	0,0060	2000/01
30 – 34	1,2592	1681	0.0071	$q(5)$	0,0073	1997	0,0073	1999/00
35 – 39	1,4651	1279	0.0149	$q(10)$	0.0096	1997	0.0860	1999/00
40 – 44	1.7438	354	0.0141	$q(15)$	-	-	-	-
45 – 49	2,3636	26	0.0000	$q(20)$	-	-	-	-

The construction of a life table based on the estimated level (α) and pattern (α) derived from the aplication of the Brass's logit life table system, implied in a life expectancy at birth of 77.7 years and, an infant mortality of 3.7 per 1000 live births for both sexes for Grand Porto during the period 2005-2006. These life table estimates seem very plausible, since the official estimate of life expectancy at birth for the Grand Porto in the year 2005 using direct vital registration data was 78.0, and infant mortality of 3.7 per 1000 live births [2].

Conclusions

As other European countries, Portugal witnessed a dramatic decline in crude rates of child mortality [3]. The challenge nowadays is to document disparities in child mortality rates. A set of characteristics should be take into account in order to perceive the decline of these disparities. However, important demographic, social and obstetric factors are not available in the vital statistics. Thus indirect estimation of child mortality based on information about previous births retrieved from medical records of women, could be a useful tool.

We evaluated the appropriateness of the Brass' method to estimate the probability of dying of dying before attaining certain exact childhood ages based on information about previous births, and their survivorship collected in the north of Portugal from a group of multiparous women, when registering a current birth. Accordingly, we observed that the obtained indirect estimates were generally consistent with direct estimates for child mortality rate under 10 years. But don't add substantial insight beyond direct estimates when vital statistics is available. However, the use of indirect method to analyze the potential of the Generation XXI cohort data in provide relevant information on reproductive issues, not available in the vital statistics, seems to be an important and effective tool, and promissory for analyzing the follow-up studies of this cohort held in 2009 and 2012.

References

1. Rosset E. [The fifth phase of demographic transition: population regression]. *Studia Demogr.* 1980(61-62):3-27.
2. Instituto Nacional de Estatística. Available at:. http://ine.pt/. Accessed May 20, 2011.
3. World Health Organization Regional Office for Europe. European health for all database (HFA-DB). Available at. http://data.euro.who.int/hfadb/. Accessed March 14, 2012
4. EURO-PERISTAT project, with SCPE, EUROCAT, EURONEOSTAT. European Perinatal Health Report 2008. Available at:. http://www.europeristat.com. Accessed October 01, 2009.
5. Pritchard C, Williams R. Poverty and child (0-14 years) mortality in the USA and other Western countries as an indicator of "how well a country meets the needs of its children" (UNICEF). *Int J Adolesc Med Health.* 2011;23(3):251-255.
6. UNICEF. The State of the World's Children 2011. Adolescence An Age of Opportunity. 2011. Available at. 2011. http://www.unicef.org/sowc2011/pdfs/SOWC-2011-Main-Report_EN_02092011.pdf.
7. United Nations. Indirect Techniques for Demographic Estimation. Manual X. A collaboration of the Population Division of the Department of International Economic and Social Affairs of the United Nations Secretariat with the Committee on Population and Demography of the National Research Council, United States National Academy of Sciences. 1983. Available at. http://www.un.org/esa/population/publications/Manual_X/Manual_X.htm. Accessed March 21, 2013.
8. Wilmoth J, Zureick S, Canudas-Romo V, Inoue M, Sawyer C. A flexible two-dimensional mortality model for use in indirect estimation. *Popul Stud (Camb).* Mar 2012;66(1):1-28.
9. Hill AG, Aguirre A. Childhood Mortality Estimates using the Preceding Birth Technique: Some Applications and Extensions. *Population Studies.* 1990;44(2):317-340.

Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013)
International Conference, Mataró (Barcelona), Spain 25 - 28 June 2013

Optimal quantization and transmission for ECG signals using genetic algorithms

Alvaro Pastor-Sanchez[1], Enrique Alexandre[1], Sancho Salcedo-Sanz[1], Daniel Gallo-Marazuela[1], and Lucas Cuadra[1]

Department of Signal Theory and Communications, University of Alcala. 28805 Alcala de Henares, Spain
(E-mail: enrique.alexandre@uah.es)

Abstract. In this paper we explore a method that, based on a particular implementation of a genetic algorithm (GA), aims to optimize the quantization of Electrocardiogram (ECG) signals to be transmitted over communication channels with low bit rate. In particular, the goal of the proposed GA is to compute the optimum number of quantization bits that minimizes the Percentage Root Mean-square-difference (PRD) between the quantized and original samples, constrained to the maximum number of available bits and the maximum distortion allowed. In this effort, we propose a novel GA-based ECG encoding approach to represent the candidate solutions or individuals, and those evolutionary algorithms that finally lead to the optimum solution. The experiments point out that, although in its preliminary steps, the method could be useful to quantize ECG signals to be transmitted over channels suffering from low bit rate.
Keywords: ECG signals, Quantization, Genetic Algorithm.

1 Introduction

Electrocardiogram (ECG) data compression is very often required because of the huge amount of information generated by this monitoring systems when applied to a patient over a long period of time. In order to store all these data, or to transmit them (for example, in telemedicine applications [1,2]), it is very important to achieve high compression ratios while maintaining an acceptable quality of the reconstructed signal in the sense that it does not interfere with the diagnostic process, or, in order words, in the effort of avoiding misdiagnosis, what is of the utmost importance.

Regarding this, there are several different approaches in the specialized literature focused on the compression of ECG data. There are important contributions, such as, for instance, [3], [4], [5] or [6], which are particular cases of either time-domain, parametric or transform-domain techniques. In [7], a different approach, consisting in the use of the JPEG2000 image compression standard to code ECG signals, has been studied.

In this paper we explore the feasibility of using Genetic Algorithms (GAs) to ECG data compression in the sense that our GA, constrained to a number of limitations (the number of available bits versus the quality of the synthesized signal), search for the optimum number of bits to quantize ECG samples.

GAs are optimization and search methods which, inspired by the principles of Genetics and Natural Selection [8], exhibit useful properties for solving

problems that, otherwise, would be intractable. Among these advantages, it is worth mentioning that GAs deal with large number of variables, provide a global solution for multilocal extrema problems, optimize functions with continuous or discrete variables, optimize variables with extremely complex cost surfaces, and do not require derivative information. In the effort of making this paper stand by itself, Section 2 focuses on summarizing the GA fundamentals that will assist us in better explaining our approach to quantize the ECG signals.

Although it will explain throughout the paper in a more detailed way, the system we propose divides the ECG signal into frames (or blocks), and computes the number of bits to properly quantize the bits of any frame. Just in this respect, Section 3 describes the proposed GA, which aims at computing the optimal number of bits to quantize the ECG that minimize the Percentage Root Mean-square-difference (PRD) of the quantized signal (when compared to the original one), constrained to the maximum number of available bits and the maximum distortion allowed.

Section 4 focuses on the experiments we have carried out in the effort of designing the GA, and discusses some preliminary results which seem to point out that the proposed GA-based approach, after some improvements, could be a feasible tool to optimally quantize ECG signals.

Finally, Section 5 summarizes our approach and the main conclusions we have found.

2 Genetic Algorithms in a nutshell

Genetic Algorithms [8] are one the most representative strategies belonging to the wide field of "Evolutionary Computation" (EC) approaches, which are inspired by the principles of Genetics and Natural Selection. In turns, EC belongs to the broader research area of "Soft-Computing" (SC). It is a key part of Artificial Intelligence (AI), and many of its methods also belong to the area of knowledge called "Natural Computing" (NC). The wider term Natural Computing refers to algorithms inspired by the way Nature solves extremely complex problems. It draws inspiration from Evolution (leading to Evolutionary Computation), Physics (Simulated Annealing), social living being networks (social insects, what leads to Ant Colony Optimization (ACO) algorithm [9], and Swarm Intelligence [10]), Neural Networks (human brain metaphor inspires Artificial Neural Networks [11,12]), Immune Systems (leading to Immunocomputing), the simulation of an orchestra composition (Harmony Search algorithm [13]), and so on.

Evolutionary Computation algorithms have been widely used for solving combinatorial optimization problems, which work primarily in intrinsically discrete search spaces. Put it very simple, these algorithms are based on an encoding of candidate solutions of the problem by using strings of numbers, and on evolving the population of candidate solutions by applying a series of evolutionary operators.

In particular, a GA is based on three key facts: 1) Encoding the candidate solution; 2) Generate an initial population of candidate solutions, and 3) Applying evolution mechanisms.

1. **Encoding candidate solutions**
   In Nature, all of the genetic information which encodes and causes the external characteristics of a living organism (or "individual") is called genotype. Any particular characteristic produced by a piece of this genetic information is encoded by a gene, the chromosome being the set of these genes. Each gene is located at a particular position on the chromosome and can have different values, called allele. Note that this strategy can be considered as transforming the real search space into another in which working is much easier. From a mathematical point of view, if F is the set containing all the candidate solutions, and G is the set of chromosomes that codifies them, this representation is equivalent of defining a bijection $\zeta : F \to G$, so that any solution is represented by an unique chromosome. Thus, roughly speaking, the terms chromosome and individual are interchangeable.
2. **Generating an initial population of candidate solutions**
   The size of the initial populations of individuals is a crucial issue for GA performance. On the one hand, a large population could cause more genetic diversity (and thus, a higher search space) and, consequently, suffer from slower convergence. On the other hand, with a very small population, only a reduced part of the search space is explored, thus increasing the risk of prematurely converging to a local extreme. Once the adequate size of the population has been explored, the second question remaining is related to the initial values its individuals should have. Usually, the values of the chromosomes representing this population are randomly initialized.
3. **Applying evolutionary mechanisms**
   In Nature, the random creation of novel genetic information may lead to the ability to survive. The better an individual is suited to an environment, the higher its probability of survival. This is the so-called *survival of the fittest* and the longer the individual's life is, the higher its chances of having descendants. In this procreation process, the parent chromosomes are combined (recombination) to provide a novel chromosome. Sporadically, and because of unavoidable errors in copying genetic information or external factors (for instance, radiation), mutations (random variations) occur. The consequence is the creation of a generation of living beings with some novel characteristic which are slightly different from those of their progenitors. If the new attribute makes the offspring better suited to the varying environment, the probabilities of survival and of having descendants also increase. Part of the offspring could inherit the modified genes and the corresponding external characteristic. In this way, the population of individuals evolves and for a number of generations the described process results in the creation of individuals better adapted to the environment, and in the extinction of those worst suited.

With these concepts in mind, we can now better understand the fundamentals of the standard genetic algorithm, proposed in [14]. This algorithm uses

operators of selection, crossover and mutation, binary encoding and selection by means of the roulette wheel method.

As motivated before, the genetic algorithm is based on a number of evolution operators, which will be detailed below, implemented in a loop process. The algorithm starts with the initialization of individuals, usually at random, and the calculation of "fitness" values associated with each individual (that is, to what extent each candidate solution solve the problem). It then enters a loop in which evolution operators are applied, until a certain stopping condition, usually, when, after a number of generations, no improvements are observed in the results. These operator work as follows:

- A selection operator aims at selecting those individuals (population components) that will be part of the population for the next generation. In the standard implementation of the algorithm, each individual has a probability of survival for the next generation proportional to its associated fitness value (objective function to optimize). This selection procedure is called the roulette method. Although the method of roulette is considered the standard algorithm, there are other well known selection methods such as the method of probabilistic tournament (widely used by its good results), the ranking selection, and so on.
- A crossover operator, whose goal consists in generating novel individuals from existing pairs of individuals. In the standard implementation, individuals are paired at random, and crossed (by exchanging parts of the binary string) with a probability called crossover probability that is usually around 60% (that is, 60% of pairs of individuals are crossed in each generation). Each pair then leads to another pair of individual offsprings, replacing parents in the next generation. There are different types of crossover methods depending on whether the parents are crossed exchanging parts in one, two or more points in the binary string.
- Mutation operator. It aims to mimic the following fact in Nature: the chromosomes (which contain genes that encode the physical characteristics of an individual −genotype−) can undergo random changes called mutations. They may be due to external causes (eg. radiation) or internal (a simple failure to copy the material). These mutations can generate individuals with novel external physical characteristics (phenotype) that may allow them (or not) adapt to the changing environment. If advantageous, the feature can be spread with a certain probability to later generations. In a Genetic Algorithm, the mutation operator generates a new individual from an existing one. This process is performed by changing certain bits chosen from 0 to 1 and vice-versa, with very low probability (usually the mutation probability for a given individual is about 1%). It differs from the previous one in that the bits change occurs within the same individual and not with another of its generation.

For non-binary implementations of algorithms, the operators of crossover and selection can be maintained as we have defined for the standard algorithm, and only the mutation operator would change, which should suit the implemented encoding selected for the specific problem.

All these concepts will assist us in explaining our approach to the problem at hand. This is just the goal of the next section.

3 The proposed GA-based quantization algorithm

Put it very simple, the ECG data stream to be quantized is divided into blocks or frames, each block containing a number of N samples. In this problem, the number of samples per block, N, and the maximum number of bits for coding the ECG signal, n_{max}, are two input parameters. For instance, the value used for n_{max} could be limited by the uplink data bit rate of the communication system in a telemedicine application.

The goal of the proposed GA-based algorithm is to find, constrained to the aforementioned limitations, the number of bits to represent the samples of each N-length block, n_i, by minimizing the overall Percentage Root Mean-square-difference (PRD) between the quantized signal and the original one. The ECG signals belong to the public MIT ECG database. This database has been divided into a training subset and a test subset to properly design the GA and test to what extent the algorithm works accurately.

The following paragraphs emphasize the key points in the design of our GA-based approach.

3.1 Encoding the candidate solution

A proper way to encode the candidate solutions or individuals is like the one illustrated by the following example. Let imagine we have only 5 blocks. The chromosome of a candidate solution could be, for example, 2 3 0 1 4, which means that each sample of the first block will be coded with 2 bits, any sample of the second block will be coded with 3 bits, and so on. To what extent this potential solution is good enough will be evaluated by the selection operator, as will be shown later on.

Once we have decided the more appropriate way of representing the candidate solution, the second key point is the design of the initial population of individuals to be evolved.

3.2 Generating the initial population of candidate solutions

The size of the initial population has been chosen to be 300 individuals, a good balance between a larger population (which could lead to slow convergence) and a smaller population (which could lead to converge to a local extreme).

The population for the problem at hand has been initialized by dividing the number of available bits, n_{max}, by all the blocks of the input signal. This division has been made randomly, ensuring that the total number of bits is equal or less than the maximum n_{max}.

Once the initial population has been initialized, the algorithm starts with the calculation of fitness values associated with each individual, and enters a loop in which evolution operators are applied until it reaches 600 generations, a stopping condition that, as will be shown in the results, keeps the PRD stable.

3.3 Applying evolutionary mechanisms

The crossover operator used is a one-point crossover operator. It randomly chooses a point on the two parent chromosomes and, by interchanging those genes before and after the point, creates the offsprings. The probability that the crossover operator is applied to each individual is called crossover probability, $p_c < 1$. In this respect, it is important to note that not all individuals are selected for crossover. After a number of experiments we have chosen $p_c = 0.9$.

The Gaussian mutation operator implemented works as follows: it selects a gene with a mutation probability p_m, adds an unit Gaussian noise (with zero mean and a standard deviation of σ chosen 0.1 times the length of the dynamic range [8,15]), and rounds it to the nearest integer. Its purpose is to maintain diversity within the population and inhibit premature convergence to local minima. Note that not all the offspring chromosomes are mutated: the probability that the mutation operator is applied to each chromosome is usually $p_m \ll 1$. Empirically, it has been observed that a value of $p_m = 0.02$ reaches good results.

Since both crossover and mutation may lead to a total number of bits higher than the maximum allowed, it is necessary to include a correction operator. This operator looks for, among all individuals, those whose number of bits is above the limit. In these cases, the number of bits is randomly reduced in steps of 1 bit until the problem is fixed.

Finally, regarding the selection, a ranking selection operator has been selected since it has been found it reaches good results.

4 Preliminary results

In this section, we present some preliminary results in the effort of checking the utility of the proposed algorithm.

Figure 1 shows the evolution of the fitness value in terms of PRD as a function of the number of generations. We have represented two PRD curves: the one corresponding to the best fitness, and another for the case in which an average of 3 bits is fixed for the genetic algorithm. As it can be observed, the convergence is fast, obtaining the best result in only around 100 generations.

While Figure 1 has shown the fast convergence of the proposed algorithm, Figure 2 represents the fitness value in terms of PRD as a function of the average number of bits per sample. As it can be observed, a fast decrease of the number of bits needed to quantify the signal is obtained when increasing the number of bits.

5 Conclusions

In this work we have explored the feasibility of using a particular implementation of genetic algorithms to optimize the quantization of electrocardiogram signals by minimizing the Percentage Root Mean-square-difference between the quantized and the original signal, constrained to the maximum number of available bits and the maximum distortion allowed. After explaning the concepts

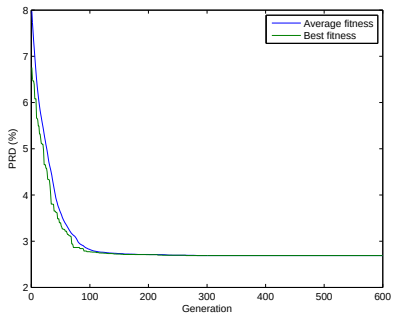


Fig. 1. Percentage Root Mean-square-difference (PRD) values (%) of the quantized signal obtained as a function of the number of generations. The blue line corresponds to the PRD results when an average of 3 bits is fixed for the genetic algorithm, while the red line corresponds to the best fitness.

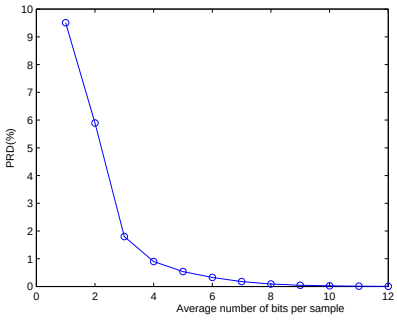


Fig. 2. Percentage Root Mean-square-difference (PRD) values (%) of the quantized signal obtained as a function of the mean number of bits per sample.

the approach is based on, we have proposed: 1) a novel encoding approach to represent the candidate chromosomes (a string of integer that labels the number of bits for each sample in a frame); and 2) some implementations of evolutionary algorithms (selection, crossover, mutation, and correction −which keeps the values of genes within the allowed dynamic margin−).

The work presented is a preliminary approach to the problem, but it is enough to show that the application of evolutionary algorithms to dynamic bit allocation problems is a promising research topic. The algorithm has proved to converge in a low number of generations, which can be reduced even more if information from one block is used to initialize the next one.

Next works will continue exploring this idea, introducing also the block length as a variable for the genetic algorithm, in the effort of improving the results.

Acknowledgements

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

References

1.R.S.H. Istepanian and A.A. Petrosian, Optimal zonal wavelet-based ECG data compression for a mobile telecardiology system, *IEEE Transactions on information technology in biomedicine*, 4(3) pp. 200-211, 2000.
2.A.I. Hernandez, F. Mora, M. Villegas, G. Passariello and G. Carrault, Real-time ECG transmission via Internet for nonclinical applications, *IEEE Transactions on information technology in biomedicine*, 5(3) pp. 253-257, 2001.
3.R.J. Coggins and M.A. Jabri, A low-complexity intracardiac electrogram compression algorithm, *IEEE Transactions on Biomedical Engineering*, 46(1), 1999.
4.G. Antoniol and Tonella, EEG data compression techniques, *IEEE Transactions on Biomedical Engineering*, 44(2), 1997.
5.G. Nave and A. Cohen, ECG compression using long-term prediction, *IEEE Transactions on Biomedical Engineering*, 40(9), 1993.
6.R.S.H. Istepanian, L.J. Hadjileontiadis and S.M. Panas, ECG data compression using wavelets and higher order statistics methods, *IEEE Transactions on information technology in biomedicine*, 5(2) pp. 108-115, 2001.
7.Ali Bilgin, Michael W. Marcellin and Maria I. Altbach, Compression of electrocardiogram signals using JPEG2000, *IEEE Transactions on Consumer Electronics*, 49(4) pp. 833-840, 2003.
8.A. E. Eiben and J. E. Smith, Introduction to evolutionary computing, Springer-Verlag, 2003.
9.M. Dorigo and T. Stützle, Ant Colony Optimization: Overview and Recent Advances. M. Gendreau and Y. Potvin, editors, Handbook of Metaheuristics, 2nd edition. Vol. 146 in International Series in Operations Research and Management Science, pp. 227-263. Springer, Verlag, New York, 2010.
10.R. Eberhart and Y. Shi, Particle swarm optimization: developments, applications and resources, In *Proc. IEEE Congress on Evolutionary Computation*, 2001.
11.C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer 2007
12.R O. Duda, P. E. Hart , and D G. Stork, *Pattern Classification*, Wiley and Sons, 2000.
13.Z. W. Geem, J. Hoon Kim and G. V. Loganathan, A New Heuristic Optimization Algorithm: Harmony Search, *Simulation*, 76(2) (2001) 60-68.
14.D. E. Goldberg, *Genetic algorithms in search, optimization and machine learning*, Reading, MA: Addison-Wesley, 1989.
15.P.S. Andrews, An Investigation into Mutation Operators for Particle Swarm Optimization, In *Proc. 2006 IEEE Congress on Evolutionary Computation*, 2006.

Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013)
International Conference, Mataró (Barcelona), Spain 25 - 28 June 2013

Multistage Non-Homogeneous Markov Chain Modeling of the Non-Homogeneous Genetic Algorithm and Convergence Results.

André G. C. Pereira[1] and Viviane S. M. Campos[2]

[1] Mathematics Department, Universidade Federal do Rio Grande do Norte, Natal/RN, Brazil
(E-mail: andre@ccet.ufrn.br)

[2] Mathematics Department, Universidade Federal do Rio Grande do Norte, Natal/RN, Brazil
(E-mail: viviane@ccet.ufrn.br)

Abstract. In this work, the results presented in the paper titled *Multistage Markov Chain Modeling of the Genetic Algorithm and Convergence Results* are extended. The size of the neighborhood in the crossover stage of the Algorithm 1, is permitted to vary throughout the evolution of the algorithm and instead of a fixed mutation probability, in mutation stage of the Algorithm 2, a sequence of mutation probabilities is introduced . Finally, numerical simulations are developed and comparisons between the original algorithms and the new versions of them are presented.

Keywords: Simulated annealing, Global optimization, Genetic Algorithms..

1 Introduction

The Canonical Genetic Algorithm (CGA), introduced in Holland[6], is a computational tool that describes the natural genetic evolutionary process of a population that undergoes three stages: selection, crossover (mating) and mutation. In the CGA, a population of N individuals or chromossomes, $(u_1, u_2, ..., u_N)$, is considered. An evaluation function $f : E \to (0, \infty)$ assigns to each individual u_i a fitness value $0 < f(u_i) < \infty$. In the selection stage, the actual population will be resampled, individuals with higher fitness are more likely to be selected and those with low fitness tends to be eliminated (elitist selection). Following the natural evolutionary process, biological reproduction (crossover) and eventual mutation occur. In the crossover stage, individuals are independently chosen for crossover with a prescribed probability p_c. Mutation also operates independently on each individual with a prescribed probability p_m. In order to be easier for implementation, each individual is represented by a binary vector of lenght l, where l depends on the desired precision. For more details as well as implementation procedures see, for example, Campos *et al.*[2], de Andrade *et al.*[3], Goldberg[5] .

In optimization context, CGAs are used to solve problems of the type $\max\{f(x), x \in E\}$ with the objective function satisfying $0 < f(x) < \infty$. The individuals represent the feasible solutions and the selection stage preserves with higher probability the best fitted/searched points. In the crossover stage,

neighboring points are searched, allowing a refined comparison in the surrounding. In the mutation stage, random points, possibly away from the preserved ones, are visited and constitute a strategy to avoid being trapped in local optimum points.

In Dorea *et al.*[4], the convergence of the genetic algorithm was proved under certain hypothesis. In particular, the Algorithm 2, which has a superindividual, was proved to converges almost surely to a set of all populations that have an optimum point as one of its elements. All algorithms presented in Dorea *et al.*[4] were modeled by an homogeneous Markov chain because theirs transition probabilities did not change through time.

The non-homogeneous Genetic Algorithms (NHGA) was introduced in Campos *et al.*[1] as an attempt of improving the eficiency of CGA, by allowing the mutation and crossover probabilities to vary under certain hypothesis. The elitist genetic algorithm (EGA), which was introduced in Rudolph[8], it was a modification in the CGA that solved the problem of eficiency of the CGAs. A non-homogeneous version of the EGA, called elitist non-homogeneous genetic algorithm (ENHGA), was introduced in Rojas Cruz and Pereira[7] in order to improve the eficiency of the EGA. Other attemptings to improve the eficiency of the CGA, without changing the mutation and crossover probabilities can be seen in Dorea *et al.*[4], some numerical comparisons between ENHGA and EGA, can be seen in Campos *et al.*[2], and the proper way of running the ENHGA can be seen in de Andrade *et al.*[3].

In this paper two different versions of each of the algorithms presented in Dorea *et al.*[4] are proposed. All algorithms presented in this paper have a superindividual as explained in Dorea *et al.*[4]. Theorical convergence results of these algorithms and numerical comparisons among these algorithms are also presented.

This paper is divided into 4 sections. In section 2, the algorithms presented in Dorea *et al.*[4], the results that guarantee their convergences as well as some other results that will be used in the rest of the paper are presented. In section 3, the changes in the algorithms presented in section 2 are introduced and convergence results are obtained. In section 4, numerical comparisons between the original algorithm and its versions are made.

2 Preliminaries

Summarizing what was said in the previous section, let $f : E \to (0, \infty)$ be a function, and genetic algorithms were set in order to find the following point

$$x^* = argmax_x\{f(x), x \in E\},$$

where E is a discretization of the domain of the function f. To proceed the following steps of the the algorithm, such points are represented as binary vectors of length l, where l depends on the desired precision. A population of size N is considered, let $Z = \{(u_1, u_2, ..., u_N); u_i \in E, i = 1, 2, ..., N\}$ be the set of all populations of size N. Z is the state space of the Markov chain that is used to prove the convergence of the algorithm, see Campos *et al.*[1], Dorea *et al.*[4],

Rojas Cruz and Pereira[7] and Rudolph[8].

In Dorea *et al.*[4], in both algorithms presented, the roulette scheme is used in the selection stage. In algorithm 1, it is proposed a different scheme for the crossover stage. In that approach, neighborhoods of the actual points of the population are used to select the next points of the population. The mutation stage is performed in a traditional way, that is, the population is considered as a binary vector and each entry can be changed independently with a fixed probability p_m . By their stages, those algorithms can be summarised in the following way:

- **Algorithm 1:** In the selection stage it is used the roulette scheme, in the crossover stage it is used neighborhoods scheme, and in the mutation scheme it is used the traditional method where each gene can be changed independently with a fix probability p_m.
- **Algorithm 2:** In the selection stage it is used the roulette scheme, there is no crossover stage, and in the mutation scheme it is used the traditional method.

The evolution of the ENHGA is different from the evolution of the EGA just by the update of the values of parameters p_m and p_c. Thus, the elitist algorithm can be summarized in the following sketch:

a) Choose randomly an initial population having N elements, each one being represented by a binary vector of length l, and create one more position, the $N+1$-th entry of the population vector, which will keep the best element from the N previous elements.

b) Repeat
   1. perform selection with the first N elements
   2. perform crossover with the first N elements
   3. perform mutation with the first N elements
   4. If the best element from this new population is better than that of the $N+1$-th position, change the $N+1$-th position by this better element, otherwise, keep the $N+1$-th position unchanged
   5. perform p_c and p_m changes, as previously planned.

c) until some stopping criterion applies.

Denote this new state space by $\tilde{Z}$.

In Rojas Cruz and Pereira[7], it is shown that the ENHGA is a non-homogeneous Markov chain, with a finite state space $\tilde{Z}$, whose transition matrices are given by $P_n = SC_nM_n, \forall n \in I\!N$, where S, C_n, M_n are transition matrices which represent the selection, crossover and mutation stages respectively. Here the M_n is composed by the third and fourth steps described in sketch just presented. In the same paper it is shown that there is a sequence $\{\alpha_n\}_{n\in I\!N}$ such that

$$\inf_{i\in\tilde{Z}, j\in\tilde{Z}^*} P_n(i,j) \geq \alpha_n,$$

where $\tilde{Z}^* \subset \tilde{Z}$, which contains all populations that have the optimum point as one of its points. The following results were obtained.

Corollary 1: Let $\{X_n\}_{n\in I\!N}$ be the Markov chain which models the elitist non-homogeneous genetic algorithm, if the sequence above is such that $\sum_{k\geq 1}\alpha_k = \infty$ then

$$P(\lim_{n\to\infty} X_n \in \tilde{Z}^*) = 1. \tag{1}$$

A more simple condition, to run in simulations, that guarantee the above result is

Corollary 2: Let $\{X_n\}_{n\in I\!N}$ be the Markov chain which models the elitist non-homogeneous genetic algorithm, if the mutation probabilities $\{p_m(n)\}_{n\in I\!N}$ are such that $p_m(n) > \gamma > 0$ for all $n \in I\!N$ or $\sum_n p_m(n)^l = \infty$ then (1) holds.

3 The New Versions of the Algorithms 1 and 2

In this section, two new versions for each algorithm 1 and 2 of Dorea *et al.*[4] as well as its almost sure convergence are presented. All algorithms presented in this paper have a superindividual as explained in Dorea *et al.*[4], or in other words, the algorithms are elitists as explained in Rojas and Pereira[7].

Consider these two versions of Algorithm 1 (with a superindividual) of Dorea *et al.*[4]

- **Algorithm 1-1:** In this algorithm we perform selection and mutation as in Algorithm 1, but in crossover stage the neighborhood size is allowed to vary in a predefined way or randomly.
- **Algorithm 1-2:** In this algorithm we perform selection as in Algorithm 1, crossover as in Algorithm 1-1 and in the mutation stage, the mutation probabilities are allowed to vary throughout time. The mutation probabilities $\{p_m^{(n)}\}_{n\in I\!N}$ are such that
  $p_m^{(n)} > \gamma > 0$ for all $n \in I\!N$ or $\sum_n p_m(n)^l = \infty$.

Note that, while the algorithms 1 and 2, in their superindividual versions are EGAs, Algorithms 1-1 and 1-2, are ENHGAs, see Rojas and Pereira[7]. Thus, we have, as a consequence of corollary 2, that

Corollary 3: In a ENHGA such that its selection stage is defined as in Algorithm 1, the size of the neighborhoods are allowed to vary throughout the evolution of the algorithm and the mutation probabilities $\{p_m(n)\}_{n\in I\!N}$ are such that $p_m(n) > \gamma > 0$ for all $n \in I\!N$ or $\sum_n p_m(n)^l = \infty$ then (1) holds.

This show that algorithm 1-1 and 1-2 converge almost surely to a subset of all populations that have as one of its points, the optimum point.

Consider these two versions of Algorithm 2 of Dorea *et al.*[4]

- **Algorithm 2-1:** In this algorithm we perform selection and crossover as in Algorithm 2, but in mutation stage, the mutation probabilities $\{p_m^{(n)}\}_{n\in I\!N}$ are such that
  $p_m^{(n)} > \gamma > 0$ for all $n \in I\!N$ or $\sum_n p_m(n)^l = \infty$.

- **Algorithm 2-2:** In this algorithm the selection is performed as in Algorithm 2, crossover as in Algorithm 1-1 and in the mutation stage, a simulated annealing (SA) algorithm without a cooling schedule is performed as in example 2 of Dorea *et al.*[4]. Note that, in Dorea *et al.*[4] it is used a matrix $Q = (Q_{xy})_{x,y \in E}$ to generate the next points, and this matrix does not change through time.

Obs: The acceptance probability used in the SA is of the form (in case of maximization)

$$a_n(X_i, Y_i) = \begin{cases} 1, & \text{if } f(Y_i) \geq f(X_i) \\ e^{-\frac{f(X_i)-f(Y_i)}{T_n}}, & \text{if } f(Y_i) < f(X_i) \end{cases} i = 1, 2, ..., N$$

where $\{T_n\}_{n \in I\!N}$ is the cooling schedule. Here $a_n(x, y) = a(x, y), \forall n \in I\!N$, because the algorithm 2-2 does not have a cooling schedule, that is, $T_n = c, \forall n \in I\!N$.

When applying the SA in each entry of the population vector, in the mutation stage, $M_n(X, Y)$ is obtained as a product of terms of the form $A_i(X, Y) = Q_{X_i Y_i} a_n(X_i, Y_i)$ when $X_i \neq Y_i$ and $B_j(X, Y) = 1 - \sum_{i \neq j} Q_{X_i Y_i} a_n(X_i, Y_i)$ when $X_j = Y_j$, where X_i and Y_i are the i-th coordinate of X, the actual population vector, and the population which was generated, Y, respectively, that is, they are the i-th individuo in the actual and generated population respectively.

Theorem 1: In an ENHGA such that its selection and crossover are performed as in algorithm 1-1 and a SA is used in the mutation stage. If the SA is such that

$$Q(x, y), a(x, y) > 0, \forall x, y \in E$$

then (1) holds.

Proof: It was shown in Campos *et al.*[1] that $P_n = SC_nM_n$, where S is a stochastic matrix that describes the selection stage,C_n is a stochastic matrix that describes the crossover stage at time n and $M_n = M$ is a stochastic matrix that describes the mutation stage at time n (in that paper the selection matrix did not change as time goes by). The SA used in mutation stage is such that

$$\inf_{A \in \tilde{Z}, B \in \tilde{Z}^*} M(A, B) \geq \gamma > 0,$$

for some $\gamma \in I\!R$. That happens because Q can generate any point of the space with a positive probability, $a_n(x, y) = a(x, y) > 0, \forall x, y \in E$ and $B \in \tilde{Z}^*$ (that is essential). Observe that such constant does not exist in the case where $A \in \tilde{Z}^*$ and $B \notin \tilde{Z}^*$. Thus, for $X, Y \in \tilde{Z}$ we can write

$$P_n(X, Y) = \sum_{A, B \in \tilde{Z}} S_n(X, A) C_n(A, B) M(B, Y)$$

If $Y \in \tilde{Z}^*$,

$$P_n(X, Y) \geq \gamma \sum_{A, B \in \tilde{Z}} S_n(X, A) C_n(A, B) = \gamma > 0$$

So,

$$\inf_{X \in \tilde{Z}, Y \in \tilde{Z}^*} P_n(X,Y) \geq \gamma = \delta_n > 0.$$

Hence,

$$\sum_{n=1}^{\infty} \delta_n = \infty$$

and by corollary 1, (1) holds.

4 Numerical comparisons

In this section we present numerical comparisons among algorithms 1, 1-1, 1-2 and among algorithms 2, 2-1,2-2. For these comparisons, the following two-variable functions are used:

- 1. $f : [-2,1] \times [-2,1] \to I\!R$ defined by $f(x,y) = 6 + x^2 - 3cos(2\pi x) + y^2 - 3cos(2\pi y)$.
- 2. $f : [-\frac{1280}{63}, \frac{1240}{63}] \times [-\frac{1280}{63}, \frac{1240}{63}] \to I\!R$ defined by $f(x,y) = .5 - \frac{sin^2(\sqrt{x^2+y^2})-0.5}{(1+.001*(x^2+y^2))^2}$.

These functions were chosen for having many local minimum and maximum points, which turns the convergence of optimization algorithms hard.

In the following simulations, each algorithm was run 100 times and in each repetition 1000 steps of the algorithm were performed. The graphics illustrate the results of such simulations. If (x,y) is a point of the graphic, y represents the number of repetitions in which the optimum point was found in x steps. Each interval that compounds the domain was partitioned into 2^6 points. Thus, a net of 2^{12} feasible solutions was created. Each algorithm was simulated for two different population size, namely: 5 and 15 points.

It was used $p_m = 0.01$ for those algorithms that used a fixed mutation probability. For those algorithms that uses a varying mutation probability, it was used a linear function to do this changes in the value of p_m. It begins with $p_m = 0.5$ and dimishes linearly to its final value $p_m = 0.001$ after 1000 steps.

It was used a neighborhood of size one for those algorithms which has a fixed neighborhood scheme.For those algorithms that use a varying neighborhood scheme, a linear funtion was used to change the size of the neighborhood in each step. The size of the neighborhood begins begger than half of the number of genes of the population and it diminishes linearly to its final value, which is the same of those algorithms that use a fixed neighborhood scheme.

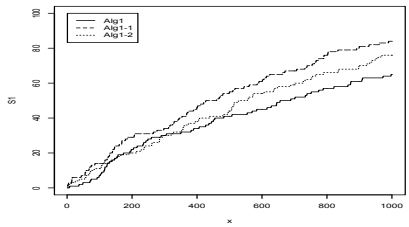


Fig. 1. Function 1, population size=5

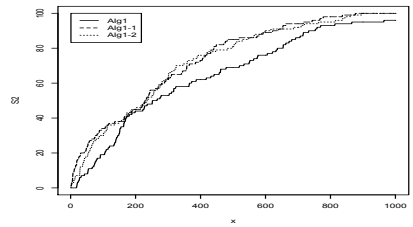


Fig. 2. Function 1, population size=15

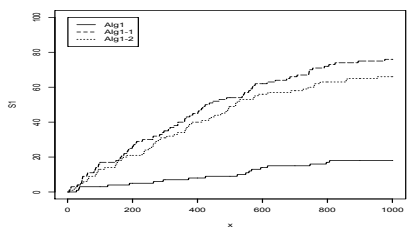


Fig. 3. Function 2, population size=5

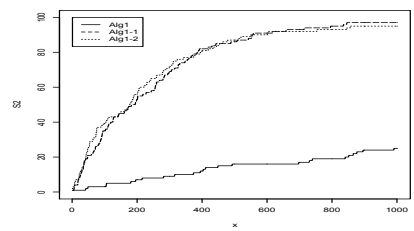


Fig. 4. Function 2, population size=15

To perform the simulations of the algorithm 2, the distribution used to generate the next point needs to be specified . In this paper a simmetric (in relation to the actual point) distribution is used. A distribution where the closer the point is from the actual point the higher is the probability of it to be chosen and its expression is given by

$$Q(x,y) = \frac{e^{-|x-y|}}{\sum_{z\in E} e^{-|x-z|}}$$

and in the acceptance probability the values of $T_n = c, \forall n \in I\!N$ that was used is $c = 10$.

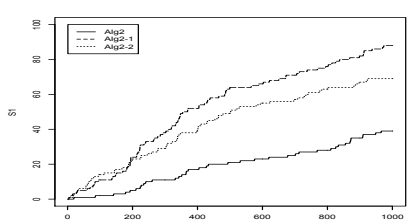


Fig. 5. Function 1, population size=5

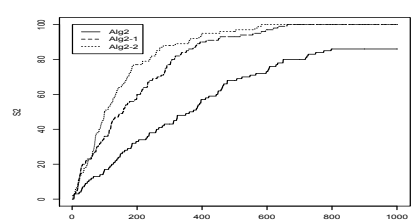


Fig. 6. Function 1, population size=15

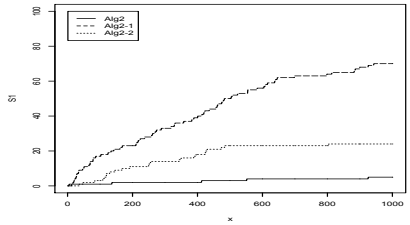


Fig. 7. Function 2, population size=5

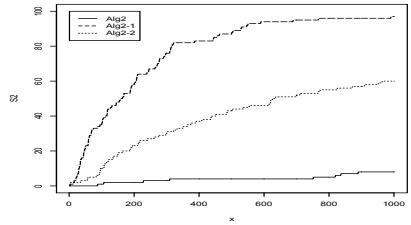


Fig. 8. Function 2, population size=15

Conclusions Changes in the algorithms 1 and 2 of [4] were proposed and sufficient conditions to guarantee its convergence almost surely to a population that has an optimum point in it were given. Examples in the previous section showed that it is possible to find a way of varying the parameters in order to obtain an improvement on the time that is spent by the algorithm for finding its optimal point. These simulations were done with the belief that higher probabilities should be used at the beginning so that the algorithm is able to search a solution in a larger region initially, and as time goes by, those probabilities should get smaller in order to do a more intense search near the population that evolves until that moment. That is just one manner to proceed, a problem that is still open is the following: How to vary the values of the parameters in order to obtain the best performance of the algorithms?

References

1.V.S.M. Campos, A.G.C. Pereira and J.A. Rojas Cruz. Modeling the Genetic Algorithm by a Non-Homogeneous Markov Chain: Weak and Strong Ergodicity. *Theory of Probability and its Applications*, v.57,Issue 1,185–192, 2012.

2.V.S.M. Campos, A.G.C. Pereira, L.A. Carlos and I.A.S. de Assis. Algoritmo Gentico por cadena de Markov homognea versus no-homognea : Un estudio comparativo. *Journal of the Chilean Institute of Operations Research*, vol 1, **1**, 2012.

3.B.B. de Andrade, A.G.C. Pereira and V.S.M. Campos. On the Convergence of Non-homogeneous Genetic Algorithms and Experimentation with Varying Mutation Rates, (submitted).

4.C.C.Yu Dorea, J.A. Guerra Jnior, R. Morgado, A.G.C. Pereira. Multistage Markov Chain Modeling of the Genetic Algorithm and Convergence Results. *Numerical Functional Analysis and Optimization*, v. 31, 164–171, 2010.

5.D.E. Goldberg. *Genetic Algorithms in Search, Optimizations and Machine Learning*, Assison Wesley, Teading, MA, 1989.

6.J.H. Holland.*Adaptation in natural and artificial systems*. Ann Arbor: The University of Michigan Press, 1975.

7.J.A. Rojas Cruz and A.G.C. Pereira. The Elitist Non-homogeneous Genetic Algorithm: Almost sure convergence. *Statistics and Probability Letters* (submitted).

8.G. Rudolph.Convergence Analysis of Canonical Genetic Algorithms. *IEEE Transactions on Neural Networks*, **5**,96–101, 1994.

Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013)
International Conference, Mataró (Barcelona), Spain 25 - 28 June 2013

EXAMINING THE FACTORS OF SUCCESS OF THE FOOTBALL TEAMS IN 2011-2012 SUPER LEAGUE SEASON IN TURKEY VIA PLS PATH MODELING

Esra Polat[1], Semra Turkan[1], and Suleyman Gunay[1]

[1]Hacettepe University, Faculty of Science, Department of Statistics, 06800-Beytepe, Ankara, Turkey (E-mails: espolat@hacettepe.edu.tr, sturkan@hacettepe.edu.tr, sgunay@hacettepe.edu.tr)

Partial Least Squares (PLS) is a family of multivariate data analysis methods that allows for analyzing blocks of variables over a set of observations. The variety of PLS methods depends on the number of blocks of variables and the relationships among blocks. The common denominator of all PLS methods is the fact that they are based on an iterative algorithm in which the parameters are calculated by a series of least squares regressions. Due to the introduction of PLS, one of the main Structural Equation Modeling (SEM) techniques, by Wold in 1966, it has been received with considerable interesting among researchers. In most cases, the PLS approach to SEM is usually named as PLS Path Modeling (PLS-PM). PLS-PM is a multivariate data analysis methodology, which provides a framework for analyzing multiple relationships between a set of blocks of variables. It is supposed that the relationships among the blocks are established taking into account previous knowledge (theory) of the phenomenon under analysis. The purpose of this study is to examine the success of 18 Turkish national football teams in 2011-2012 Super League Season by using PLS-PM.

Key Words: Football Teams, Attack, Defense, Success, PLS, PLS-PM.

1. Introduction

Partial Least Squares (PLS) is a family of multivariate data analysis methods that allows for analyzing blocks of variables over a set of observations. The variety of PLS methods depends on the number of blocks of variables and the relationships among blocks. The common denominator of all PLS methods is the fact that they are based on an iterative algorithm in which the parameters are calculated by a series of Least Squares (LS) regressions. The term Partial is due to the fact that the iterative procedure involves separating the parameters instead of estimating them simultaneously (Sanchez [2]).

PLS Path Modeling (PLS-PM) is the PLS approach to Structural Equation Modeling (SEM). Although both terms are often used interchangeably, the term PLS-PM is preferred over SEM around the small but growing PLS community. PLS-PM is a multivariate data analysis methodology that provides a framework for analyzing multiple relationships between a set of blocks of variables. It is supposed that the relationships among the blocks are established taking into account previous knowledge (theory) of the phenomenon under analysis. Moreover, each block of variables is assumed to play the role of a theoretical concept represented by in the form of a latent (unobserved) variable (LV). (Sanchez [1]; Sanchez [2]).

The purpose of this study is to examine the success of 18 Turkish national football teams in 2011-2012 Super League Season by using PLS-PM. For this purpose, following the study of Sanchez (2013), a model is proposed in which the overall success of the football teams depends on the quality of the attack as well as on the quality of the defense made by them. There are three LVs which are defined as attack, defense and success. The number of won matches at home and the number of won matches away variables are taken as indicators of success, the number of goals scores at home and the number of goals scores away variables are taken as indicators of attack and the number of goals conceded at home and the number of goals conceded away are taken as indicators of defense.

2. PLS Path Modeling

Before we get involved into the details of the PLS-PM, we need to talk about a couple of things closely related to it: a) Latent Variables, b) PLS Path Model (Structural Model, Measurement Model, Weight Relations) c) Path Diagrams (Sanchez [1]; Sanchez [2]).

2.1. Latent Variables: Measuring the Immeasurable

One of the first concepts we must mention is that of LVs. The most common definition of a LV is a variable that is unobserved. More precisely, a LV is a variable that could not be measured directly. Examples of LVs abound in social sciences (e.g., psychology, sociology, economy, and politics) such as intelligence, motivation, satisfaction, socioeconomic status, economic development, social crisis, etc. (Sanchez [2]). In statistics, LVs are

widely used in several data analysis and modeling techniques with applications in many fields of knowledge. Within the literature related to LVs we can find synonymous terms like: theoretical concepts, hypothetical variables, constructs, composites, factors and intangibles (Sanchez [1]; Sanchez [2]).

In the context of PLS-PM, a LV will be obtained as a linear combination of a set of observed variables, which are also called as manifest variables (MVs), indicators or items. We assume that MVs contain information that reflect or indicate one aspect of the construct; hence we use the information contained in MVs to obtain an approximate representation of the LV. LVs can be measured in two ways: (1) through their consequences or effects reflected on their MVs and (2) through different MVs that are assumed to cause the LVs. In the first case, called reflective way, MVs are considered as being caused by the LVs. The second case is known as formative way because the MVs are forming the LVs The main difference between the reflective and formative ways has to do with the causal-effect relationships between the MVs and the LVs (Sanchez [1]; Sanchez [2]).

Let's assume that we have p variables measured on n observations (i.e. individuals, cases, samples), and that the variables can be divided in j blocks. X is the data set containing the n observations and p variables that means a matrix of dimension nxp. X can be divided in j (mutually exclusive) blocks $X_1, X_2, \ldots, X_j$. Each block X_j has K variables: $X_{j1}, X_{j2}, \ldots, X_{jK}$. Each block X_j is assumed to be associated with a LV denoted as LVj. Keep in mind that LVj is just an abstract representation (i.e. unobserved). The estimation of a LV, also known as score, is denoted by $L\hat{V}_j = Y_j$ (Sanchez [1]).

2.2. PLS Path Model

In this section, the formalization of a PLS Path Model and its specifications are mentioned in detail.

2.2.1. Structural Model

Firstly, it could be talked about the specifications of the structural part in a PLS Path Model. A structural model, which is also called an inner model, relating some dependent LVs to other LVs (Tenenhaus *et al.* [3]).There are three things that is needed to consider about the inner relationships: Linear relationships, recursive models, regression specification.

Linear Relationships: The first aspect of an inner model is that the structural relationships are treated as linear relationships. The structural relations are expressed in mathematical notation as in Eq. (1). The subscript i of LV_i refers to all the LVs that are supposed to predict LV_j. The coeffcients β_{ji} are the path coeffients and they represent the "strength and direction" of the relations between the response LV_j and the predictors LV_i. β_0 is just the intercept term, and the error_j term accounts for the residuals (Sanchez [1]).

$$LV_j = \beta_0 + \sum_{i \to j} \beta_{ji} LV_i + \text{error}_j \tag{1}$$

Recursive Models: The second aspect, which is needed to be aware of, is that the system of equations must be a recursive system. This means that the paths formed by the arrows of the inner model could not form a loop (Sanchez [1]).

Regression Specification: The third aspect about the inner relationships is something called predictor specification which is just a fancy term to express a linear regression idea. The idea behind this specification is that the linear relationships are conceived from a standard regression perspective as shown in Eq. (2) (Sanchez [1]).

$$E(LV_j | LV_i) = \beta_{0i} + \sum_{i \to j} \beta_{ji} LV_i \tag{2}$$

2.2.2. The Measurement Model

The measurement or outer model is the part of a model that has to do with the relationships between a LV and its block of MVs. The relevant aspect about the outer model is that there are two main measurement options: reflective blocks and formative blocks (Sanchez [1]; Tenenhaus *et al.* [3]).

Reflective Way: The most common type of measurement is the reflective mode. In this case, the LV is considered as the cause of the MVs. That's why it's called reflective because the MVs are "reflecting" the LV. Just like in the inner model, the outer model relationships are also considered to be linear. The outer model for reflective way is expressed in mathematical notation as in Eq. (3). The coefficients λ_{jk} are called loadings, λ_0 is the intercept term, and the error terms account for the residuals (Sanchez [1]; Tenenhaus *et al.* [3]).

$$X_{jk} = \lambda_{0jk} + \lambda_{jk} LV_j + error_{jk} \tag{3}$$

Formative Way: The other type of measurement is the formative mode. In this case, the MVs are considered to be the cause of the LV. That's why it's called formative because the MVs are "forming" the LV. The outer model for formative way is expressed in mathematical notation as in Eq. (4) (Sanchez [1]; Tenenhaus *et al.* [3]).).

$$LV_j = \lambda_{0j} + \lambda_{jk} X_{jk} + error_j \tag{4}$$

2.2.3. The Weight Relations

There's one last thing belonging to the specifications of a PLS Path Model that it is needed to mention: the weight relations. The weight relations is used to bridge the gap between the virtual LVs and the material LVs. In PLS-PM, LVs are estimated as a linear combination of their MVs. Furthermore, an estimated LV_j is called a score, denoted as Y_j as in Eq. (5). Since LVs are calculated as a weighted sum of their MVs analogously to what is done in principal component analysis, PLS-PM is referred to as a component-based approach (Sanchez [1]).

$$L\hat{V}_j = Y_j = \sum_k w_{jk} X_{jk} \tag{5}$$

2.3. Drawing Models: Arrow (Path) Diagrams

The term path modeling is a very generic term used to designate a set of different statistical techniques that seek to explain the relationships among multiple variables. The reason for the term path modeling is due to the graphical representation of structural relationships by drawing a picture of the model. Pictures of path models are called path diagrams, which are drawn according to well established conventions of terminology and symbols: Variables can be of any kind, i.e. MVs, LVs or residual variables (error terms). MVs are enclosed in boxes, LVs are enclosed in circles/ellipses and error terms are maintained unclosed. Relationships also can be of three types: causal links meaning that variable A causes variable B; correlation links indicating simply correlation between two variables A and B without implying causality; or the affection of a error term ε to some variable A. Causal relationships are assumed to be linear, and are represented by straight single-headed arrows, correlations are represented by curved two-headed arrows, and residual affection by straight lines. In addition, variables may be grouped in two classes: those that are not caused by any other variables in the diagram, and those that are caused by one or more variables. The first class of variables is called exogenous or independent variables. The second class is known as endogenous or dependent variables. The convention is to use Greek letters for the LVs, and Italic letters for the MVs. Exogenous LVs are usually represented by the Greek letter $\xi(x_i)$, while endogenous LVs are represented by $\eta(\text{eta})$. Path diagrams are very helpful because they allow for the visualization of the relations and, in terms of a causal model, its graphical display makes it possible to understand the conceptualization of the model. The visual notations of these variables and relationships in path diagrams are shown in Fig. 1 (Sanchez [2]).

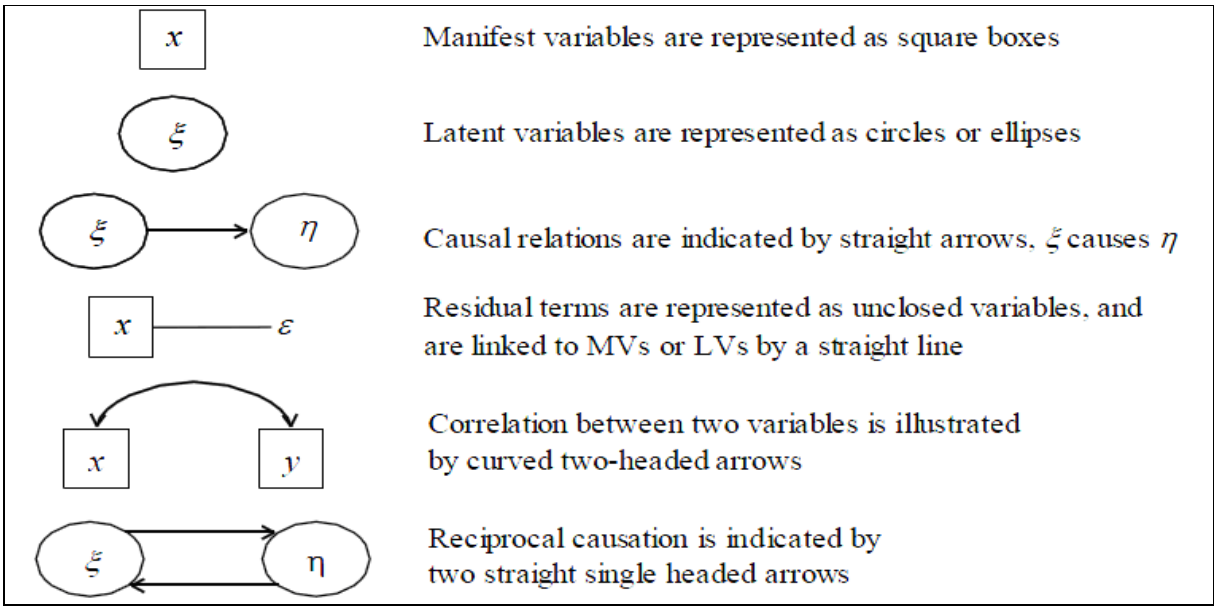


Fig. 1. Path diagram notations (Sanchez, 2009)

PLS Path Modeling follows a sequential procedure that can be divided in three major stages; Stage 1: Get the weights to compute LV scores; Stage 2: Estimating the path coefficients (inner model); Stage 3: Obtaining the loadings (outer model).

3. Application

The simple theory for examining the success of 18 Turkish national football teams in 2011-2012 Super League Season by using PLS-PM involves two hypotheses. In one of them, it is supposed that if a team improves its attack, it should be more successful, therefore, wins more matches. The other hypothesis is that if a team improves its defense, it should also be more successful or at least it should avoid losing matches. This theory could also be expressed in a more abstract form like this: $\text{Success} = f(\text{Attack}, \text{Defense})$. This is simply a conceptual way to say that Success is a function of Attack and Defense. But we could go further by specifying a linear function and expressing this theory with an equation like this one: $\text{Success} = b_1\text{Attack} + b_2\text{Defense}$.

3.1. Measuring Success, Attack and Defense

We have used the model proposed by Sanchez (2013) in which the Overall Success depends on the Quality of the Attack as well as on the Quality of the Defense. These are our three LVs. Now we need to establish a set of indicators for each of the three LVs. Following his study, we are going to take **GSH** and **GSA** variables as indicators of Attack; **GCH** and **GCA** variables as indicators of Defense; **WMH** and **WMA** variables as indicators of Success. These variables are defined as below:

• **GSH:** Number of goals scores at home
• **GSA:** Number of goals scores away

• **GCH:** Number of goals conceded at home
• **GCA:** Number of goals conceded away

• **WMH:** Number of won matches at home
• **WMA:** Number of won matches away

The data including these six variables, of 18 Turkish national football teams in 2011-2012 Super League Season, is given in Table 1.

Football Teams	GSH	GSA	GCH	GCA	WMH	WMA
Galatasaray	38	31	14	10	13	10
Fenerbahçe	39	22	12	22	14	6
Trabzonspor	36	24	22	17	10	5
Beşiktaş	29	21	18	21	10	5
Eskişehirspor	18	24	16	25	8	6
Büyükşehir bld	31	7	15	34	9	5
Sivaspor	25	32	20	34	7	6
Bursaspor	20	24	12	23	8	5
Gençlerbirliği	32	17	16	32	9	4
Gaziantepspor	24	15	13	20	9	4
Kayseri	28	14	19	20	8	5
Kardemir	30	14	21	35	11	2
Mersin	15	19	24	21	5	7
Ordu	17	11	13	21	7	3
Antalya	19	13	18	24	7	3
Samsun	24	12	24	23	6	3
Manisa	16	15	27	25	5	3
Ankaragücü	7	15	36	41	1	1

Table 1. The data for 18 Turkish national football teams in 2011-2012 Super League Season
(Data source: http://www.iddaliyim.org/gecmis/turkiye_superlig_gecmis/2011-2012_sup_lig_sonuc.php)

Firstly, it must be mentioned that to assess the quality of a reflective block it is need to understand the key ideas behind a reflective measurement model: it is supposed that reflective indicators are measuring the same underlying LV, hence they are reflections of the construct. On one hand, reflective indicators need to have strong mutual association. Shortly, they will be highly correlated. On the other hand, reflective indicators need to get along with its LV; they must show sings of membership and belonging to one and only one LV: they need to be loyal to its construct. If one indicator loads higher on another construct, this could be evidence of treason. **Basically, three aspects of reflective measures must be evaluated:**

1. Unidimensionality of the indicators: In PLS-PM there are three main indices to check unidimensionality:
a) Calculate the Cronbach's alpha**, b)** Calculate the Dillon-Goldstein's rho **c)** Check the first eigenvalue of the indicators' correlation matrix. The R output of the three indices of unidimensionality is given in Table 2. In Table 2, the first column shows the type of measurement. In our example, all the blocks are reflective. The second column indicates the number of MVs in each block (2 in our example). The third column contains the Cronbach's alpha, the fourth column is the Dillon-Goldstein's rho, the fifth and sixth columns are the first and second eignevalues, respectively. Dillon-Goldstein's rho is considered to be a better indicator than the Cronbach's alpha because it takes into account to which extent the LV explains the block of indicators. Hence, here we used this index to assess the unidimensionality. As a rule of thumb, a block is considered as unidimensional when Dillon-Goldstein's rho is larger than 0.7. As seen from Table 2, the values of Dillon-Goldstein's rho's for attack, defense and success blocks are larger than 0.7. So, these blocks are considered as unidimensional.

	Type.measure	MVs	C.alpha	DG.rho	eig.1st	eig.2nd
Attack	Reflective	2	0.4483647	0.7838111	1.288963	0.7110373
Defense	Reflective	2	0.6227777	0.8413180	1.452198	0.5478016
Success	Reflective	2	0.6940999	0.8673403	1.531511	0.4684893

Table 2. The unidimensionality metrics for each block of indicators

2. Check that indicators are well explained by its LV: Communalities represent the amount of variability explained by a LV. The loadings are correlations between a LV and its indicators. In turn, communalities are squared correlations. The R output of communalities are given in Table 3. Each element in the list is a table (a matrix) with four columns. The first column contains the outer weights. The second column are the loadings (correlations). Loadings greater than 0.7 are acceptable. In Table 3, all loadings are greater than 0.7.

Communalities are just squared loadings. They represent the amount of variablity explained by a LV. A loading grater than 0.7 means that more than $0.7^2 \approx 50\%$ of the variablity in an indicator is captured by its latent construct. As seen from third column in Table 3, the communalities of **GSA** and **GCA** are not greater than 0.7. It means that more than $0.7^2 \approx 50\%$ of the variablity in an indicator is not captured by its latent construct. But these variables are keeped in the model since loading values of these variables are greater than 0.7.

```
> data1_pls$outer.mod
$Attack
     weights std.loads communal redundan
GSH   0.7275    0.8741   0.7641        0
GSA   0.5073    0.7176   0.5149        0

$Defense
     weights std.loads communal redundan
GCH   0.6315    0.8760   0.7674        0
GCA   0.5408    0.8263   0.6828        0

$Success
     weights std.loads communal redundan
WMH   0.6122    0.8937   0.7987   0.7190
WMA   0.5296    0.8550   0.7311   0.6582
```

Table 3. The loadings and communalities

3. Assess the degree to which a given construct is different from other constructs: Besides checking the loadings of the indicators with their own LVs, the so-called cross-loadings must also be checked. That is, the loadings of an indicator with the rest of LVs. The reason for doing so is that it is needed to be sure that there are not traitor indicators. The cross-loadings are given in Table 4. It is needed to look at the list of results as if it was a super matrix. The way to read the cross-loadings is by looking at this super matrix block by block paying attention to the sections in the diagonal. These sections are the loadings of each block with its construct. A given loading in one of these sections must be greater than any other loading in its row (Sanchez [1]). For example, let's consider the first section that corresponds to the first column in the Attack block. GSH has a loading value of 0.8741. This value must be greater than any other value in that first row. The cross-loadings of **GSH** with Defense is -0.5248; the cross-loading of **GSH** with Success is 0.8127. Clearly, 0.8741 is greater than -0.5248 and 0.8127.

With the cross-loadings it is evaluated the extent to which a given construct differentiates from the others. The whole idea is to verify that the shared variance between a construct and its indicators is larger than the shared variance with other constructs. In other words, no indicator should load higher on another construct than it does on the construct it intends to measure. Otherwise, it is a traitor indicator. If an indicator loads higher with other constructs than the one it is intended to measure, its appropriateness might be considered because it is not clear which construct or constructs it is actually reflecting (Sanchez [1]).

```
> data1_pls$outer.cor
$Attack
     Attack Defense Success
GSH 0.8741 -0.5248  0.8127
GSA 0.7176 -0.2725  0.5660

$Defense
     Attack Defense Success
GCH -0.470  0.8760 -0.6974
GCA -0.413  0.8263 -0.5969

$Success
     Attack Defense Success
WMH 0.8262 -0.7076  0.8937
WMA 0.7035 -0.6230  0.8550
```

Table 4. The cross-loadings

Secondly, **<u>the quality of the structural model is evaluated by examining three indices or quality metrics:</u>** **a)** the R^2 determination coefficients, **b)** the redundancy index, **c)** the Goodness-of-Fit (GoF). These metrics are given in Table 5. As seen from Table 5, the inner model seems to be fine since the value of R-square, $R^2 = 0.90$,

is higher. It means that the 90% of variance in the Success explained by Attack and Defense. "Av.Redun" represents the percentage of the variance in the endogenous block that is predicted from the indepedent LVs associated to the endogenous LV. High redundancy means high ability to predict. In our model, the average redundancy for Success represents that Attack and Defense predict 68% of the variability of Success indicator. The third quality metric of model is GoF. GoF can be used as a global criterion that helps us to evaluate the performance of the model in both the inner and the outer models. Basically, GoF assess the overall prediction performance of the model. In our model, the value of GoF is 0.79 could be interpreted as if the prediction power of the model is of 79%. The naive rule of thumb is: the higher, the better. Acceptable "good" values within the PLS-PM community are GoF >0.7, as the case in our example.

```
> data1_pls$inner.sum
        LV.Type Measure MVs  R.square  Av.Commu  Av.Redun       AVE
Attack   Exogen   Rflct   2 0.0000000 0.6395011 0.0000000 0.6395011
Defense  Exogen   Rflct   2 0.0000000 0.7250808 0.0000000 0.7250808
Success Endogen   Rflct   2 0.9002366 0.7649069 0.6885972 0.7649069
```

Table 5. Quality Metrics of structural model

As PLS-PM is a soft modeling approach with no distributional assumptions, it is possible to estimate the significance of the parameters based on cross-validation methods like jack-knife and bootstrap (Trinchera and Russolillo [4]). The estimating the precision of the PLS parameter estimates is obtained by using bootstrapping which is a non-parametric approach. The bootstrap procedure is the following: M samples are created in order to obtain M estimates for each parameter in the PLS model. Each sample is obtained by sampling with replacement from the original data set, with sample size equal to the number of cases in the original data set (Sanchez [1]). The results of bootstrapping are given in Table 6. In table 6, for each of the displayed results, the bootstrap confidence interval (95%) provided by the percentiles 0.025 and 0.975 should be examined. . This is especially important for the path coefficients. When bootstrap intervals for the path coeffcients contain the zero, it means that these coefficients are not significant at a 5% confidence level. As seen from Table 6, the bootstrap confidence intervals for the Attack and Defense indicators do not contain zero. Hence, the path coefficients of these indicators are significant at a 5% confidence level.

```
BOOTSTRAP VALIDATION
weights
     Original  Mean.Boot  Std.Error  perc.025  perc.975
GSH     0.728      0.737      0.166    0.5500     0.997
GSA     0.507      0.455      0.170   -0.0584     0.626
GCH     0.631      0.644      0.181    0.3656     0.996
GCA     0.541      0.484      0.265   -0.1604     0.865
WMH     0.612      0.633      0.142    0.5154     0.926
WMA     0.530      0.479      0.179   -0.2325     0.637

loadings
     Original  Mean.Boot  Std.Error  perc.025  perc.975
GSH     0.874      0.882      0.133     0.789     0.976
GSA     0.718      0.640      0.273    -0.281     0.885
GCH     0.876      0.856      0.162     0.368     0.996
GCA     0.826      0.720      0.338    -0.255     0.952
WMH     0.894      0.899      0.130     0.782     0.973
WMA     0.855      0.776      0.298    -0.424     0.965

paths
                   Original  Mean.Boot  Std.Error  perc.025  perc.975
Attack->Success       0.66      0.655     0.1066     0.429     0.847
Defense->Success     -0.42     -0.425     0.0933    -0.614    -0.214

rsq
         Original  Mean.Boot  Std.Error  perc.025  perc.975
Success       0.9      0.908     0.0518     0.788     0.978

total.efs
                   Original  Mean.Boot  Std.Error  perc.025  perc.975
Attack->Defense       0.00      0.000     0.0000     0.000     0.000
Attack->Success       0.66      0.655     0.1066     0.429     0.847
Defense->Success     -0.42     -0.425     0.0933    -0.614    -0.214
```

Table 6. Results of Bootstrapping

In addition to expressing our model in text and mathematical format, we can also display our model in a graphical format using a path diagram as shown in Fig. 2. As mentioned before, these diagrams help us to

represent the relationships stated in our models in a visual way. In our example, the following diagram depicts the relation Success depending on the quality of the Attack as well as on the quality of the Defense. In brief, the model parameter estimates of PLS-PM of our example are shown in Fig. 2.

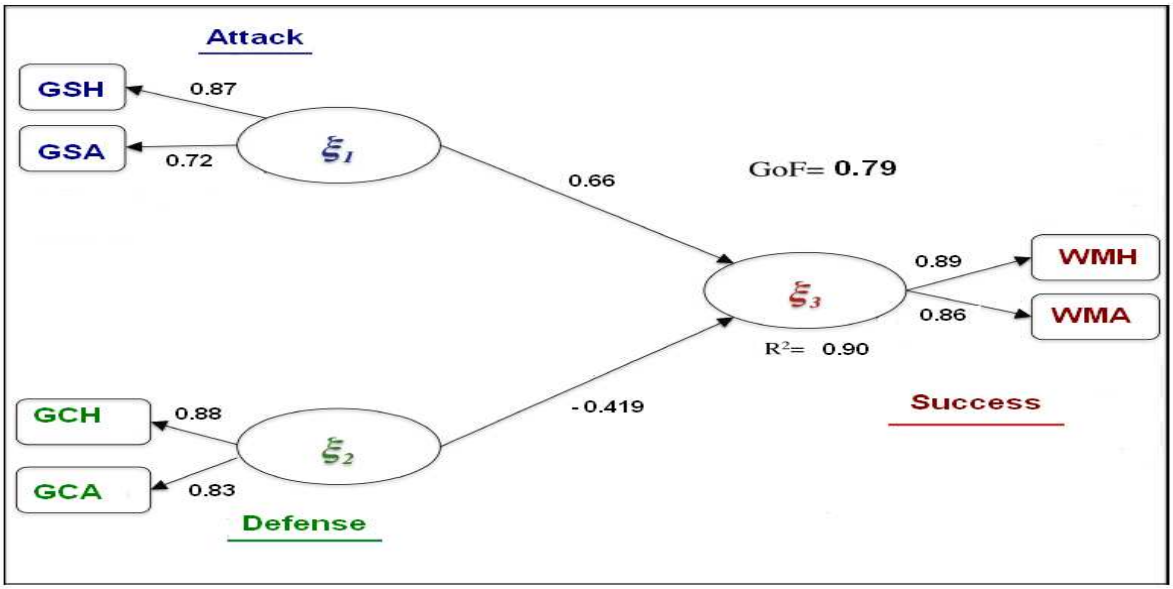


Fig. 2. PLS-PM analysis of Turkish football teams

As seen from Fig. 2, it is possible to examine the relations between Attack, Defense and Success via path coefficients: the two LVs Attack and Defense impact in opposite sense on the response Success. However, Success largely depends on Attack rather than on Defense. The higher attack causes the higher success but the higher defense causes the lower success. Because the higher defense means that the high values of GCH and GCA. Namely, teams conceded a lot of goals. Hence, the negative path coefficient related to defense refers to if the number of conceded goals of a team is increased, the success of the team is decreased. GSH, GSA, GCH, GCA, WMH, WMA are positively correlated to the LVs Attack, Defense, Success respectively.

Conclusions

In this study, we examine the success of 18 Turkish football teams using PLS Path Model. The relationship between LVs Attack, Defense and Success are examined. The LV Defense has a negative effect on the Success, however, the LV Attack is possitive effect on the Success. The Success of the Turkish teams largely depends on Attack. The better the attack, the more number of GSH and GSA, however, the better defense, the less number of GCH and GCA. Proposed a model in which the Overall Success depends on the Quality of the Attack as well as on the Quality of the Defense. Although GCH and GCA have to do with defense, they are measuring "lack" of defense. If a team has high values of GCH and GCA, they conceded a lot of goals, hence having a poor Defense quality. Briefly, GCH and GCA are pointing in the opposite direction. Hence, the better Quality of the Defense means that a team has low values of GCH and GCA.

References

1. G. Sanchez. *PLS Path Modeling with R*, http://www.gastonsanchez.com/PLS_Path_Modeling_with_R.pdf, 2013.
2. G. Sanchez. *Understanding Partial Least Squares Path Modeling (An introduction with R)*, Academic Paper, Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, March, 2009.
3. M. Tenenhaus , V. Esposito Vinzi, Y.M. Chatelinc , C. Lauro. PLS path modeling. *Computational Statistics & Data Analysis*, 48, 159-205, 2005.
4. L. Trinchera and G. Russolillo. *On the use of Structural Equation Models and PLS Path Modeling to build composite indicators*, Working paper n. 30, September, 2010, http://www.unimc.it/sviluppoeconomico/wpaper/wpaper00030/filePaper .

Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013)
International Conference, Mataró (Barcelona), Spain 25 - 28 June 2013

Confidence Intervals for Dynamic Concentration/Inequality Indices of Economic Systems Modeled by Birth-Death Processes

Philippe Regnault[1]

Laboratoire de Mathématiques de Reims, UFR Sciences Exactes et Naturelles,
Moulin de la Housse, BP1039, 51687 REIMS cedex 2, France
(E-mail: philippe.regnault@univ-reims.fr)

Abstract. In this paper, we are interested in estimating a dynamic index of the concentration of wealth shared by a population of economic agents – Theil's entropy. At any time, each agent produces an amount of wealth, independent of other agents. According to this wealth production, a wealth class (among predefined classes) is associated to each agent. We suppose that wealth class allocation is governed by a continuous-time birth-death process so that the repartition of total wealth's production among classes at any time is an explicit function of birth and death rates. Then, by considering Theil's entropy of wealth distribution among all classes at any time yields dynamic concentration index of wealth. If moreover, the birth-death process is ergodic, then Theil's entropy associated to the asymptotic repartition of wealth measures the concentration of wealth when agents' population reaches its equilibrium state.

When the class allocation process is observed at equidistant discrete times, it is necessary to estimate Theil's entropy index. In the same fashion as in Regnault [10], we prove the strong consistency and asymptotic normality of plug-in estimators of dynamic and asymptotic Theil's entropy, built from empirical estimators of the transition matrix of the discretized class allocation process. Confidence intervals for dynamical Theil's entropy are then derived.
Keywords: Population Dynamic, birth-death Process, Theil's Entropy, Reward, Estimation.

1 Introduction

Both inequality and wealth concentration dynamics within an economic system evolving over time have been extensively studied in the last years. In particular, D'Amico and Di Biase [1] recently proposed and studied stochastic versions of usual wealth concentration indices. They consider economic systems involving independent agents whose income evolves over time according to some discrete-time semi-Markov model. These agents are partitioned into a few number of economic classes according to their income or some other economic, social or demographic feature; concentration wealth among classes is then measured dynamically by considering its Theil's entropy at any time – the so-called dynamic Theil's entropy. The model has been successfully implemented in D'Amico *et al.* [2] for simulating an artificial economic system with immigration, and also in D'Amico *et al.* [3], applied to studying wealth concentration in several European countries whose citizen are partitioned according to their age.

The present paper aims at adapting the framework proposed by D'Amico and Di Biase [1] to estimating wealth concentration of dynamic economic systems involving a large number of independent agents, where agents' income class allocation is modeled by a continuous-time birth-death process. These assumptions are mainly motivated by the need of a simple and flexible model to be estimated and implemented for real applications. The birth-death model is also supported by previous studies; see Quah [9] and D'Amico *et al.* [3] where only transitions from one class to its neighboring classes are shown to be statistically significant. Moreover, income dynamics is usually modeled by a discrete time process because measurements of income of individuals occur at some fixed and known times, for example each year or fractions of year. On the contrary, agents' income can change at any time. One major problem is how to estimate the income inequalities from discrete observations of agents' income class allocation continuous-time process. We define and show good asymptotic properties of plug-in estimators of dynamic Theil's entropy function built from discrete observations of one or several agents' income. We then derive from these properties confidence intervals for Theil's entropy.

The paper is organized as follows: Section 2 presents the stochastic model. Section 3 introduces the dynamic version of Theil's entropy and its asymptotic value as time goes to infinity. Finally, Section 4 defines and establishes asymptotic properties of plug-in estimators of both dynamic and asymptotic Theil's entropy; confidence intervals for Theil's entropy are then derived from these properties.

2 The stochastic model

Let an economic system be composed of N economic agents, typically individuals. Each agent $i \in \{1, \ldots, N\}$ produces a quantity $X_i(t)$ of wealth – their income, changing randomly with time t so that $\mathbf{X}_i := (X_i(t))_t$ is a positive valued stochastic process.

Assume that the system evolves in a non competitive way so that wealth production of the agents do not depend on each others, that is the stochastic processes $\mathbf{X}_i$, $i \in \{1, \ldots, N\}$ are independent. The agents of the system are classified according to their income and allocated into one of K mutually exclusive and ordered classes

$$E = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K\},$$

such that agent i belongs to $\mathcal{C}_k$ at time t if its income $X_i(t)$ belongs to $[w_k, w_{k+1}[$, where $w_1 = 0 < w_1 < \cdots < w_K < w_{K+1} = \infty$ are fixed thresholds. Then, let us introduce for any agent i, the process $\mathbf{C}_i := (C_i(t))_{t \in \mathbb{R}_+}$ that maps income $X_i(t)$ of agent i at time t into its corresponding class $\mathcal{C}_k \in E$.

The main assumptions of the present study are the following.

Assumptions 1 *1. the number of agents N tends to infinity;*

2. the processes $\mathbf{C}_i = (C_i(t))_{t \in \mathbb{R}+}$, $i \in \mathbb{N}^*$ are independent and identically distributed birth-death processes taking values in $E = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ with common initial distribution ν;

3. *given* $k \in \{1, \ldots, K\}$, *the conditional expectation* $W_k := \mathbb{E}(X_1(t)|C_1(t) = \mathcal{C}_k)$ *is constant over time* t.

The dynamic of the process $\mathbf{C}_i$ is completely described by its generator which is a tridiagonal matrix, say

$$\mathbf{A} := \begin{pmatrix} -\lambda_1 & \lambda_1 & & 0 \\ \mu_2 & -\lambda_2 - \mu_2 & \mu_2 & \\ & \ddots & \ddots & \ddots \\ 0 & & \mu_K & -\mu_K \end{pmatrix}. \tag{1}$$

The coefficients λ_k and μ_k are called the birth and death rates.

Finally, let us define for $\eta > 0$ and $i \in \mathbb{N}$, the sequence $\mathbf{D}_i := (C_i(\eta m))_{m \in \mathbb{N}}$ of income classes of agent i at times ηm, $m \in \mathbb{N}$ – say discretized class allocation process. It is well known that $\mathbf{D}_i$ is an homogeneous irreducible and aperiodic Markov chain with transition matrix

$$\mathbf{P}_\eta := \exp(\eta \mathbf{A});$$

see for example Bladt and Sørensen [6]. Thus, the coefficients of the matrix $\mathbf{P}_\eta$ are probabilities that an agent's income has moved from one income class to another after η units of time.

Note that η must be closed to the mean jump time of $\mathbf{C}_1$ for $\mathbf{D}_1$ to catch relevant information about $\mathbf{C}_1$; see Bladt and Sørensen [6] and references therein for more details.

Example 1 *Service providers such as retail brands constitute economic systems involving a large number of agents – the customers. Such providers offer different levels of service according to the budget (or income)* $\mathbf{X}_i$ *of customers; for example, a retail brand can sell several competitive products for a common use with a large range of prices. Customers are then partitioned into exclusive economic classes* $\mathcal{C}_1, \ldots, \mathcal{C}_K$ *at any time, according to the service level (or product) they choose. Any customer allocated to the service class* $\mathcal{C}_k$ *contributes to the provider's turnover by* W_k *units of money per unit of time, where* W_k *is the price of type* k *service.*

3 Dynamic Theil's entropy

We are interested here in quantifying the inequality of repartition of the total wealth into the different classes $\mathcal{C}_k \in E$. For this purpose, let us define the instantaneous share of produced income due to class $\mathcal{C}_k$ at time t by

$$p_k(t) := \frac{\sum_{i=1}^N X_i(t) \mathbb{1}_{\{C_i(t) = \mathcal{C}_k\}}}{\sum_{i=1}^N X_i(t)}.$$

One of the most used inequality measure is Theil's entropy proposed by Theil [13]. It derives from the mathematical theory of communication founded by Shannon [12]. As stated in Athanasopoulos and Vahid [4] or Cowell [7], it is

more sensitive to transfers of wealth in the lower tail than in the upper tail. It is also additively decomposable: for any significant grouping of agents population, Theil's entropy of the whole population can be additively decomposed to inequality between subgroups and an appropriately weighted average within each group.

A stochastic version of Theil's entropy was recently proposed in [1] by considering the process $(\mathbb{T}(t))_{t\in\mathbb{R}_+}$, where

$$\mathbb{T}(t) := \sum_{k=1}^{K} p_k(t) \log\big(Kp_k(t)\big),$$

with the convention $0\log 0 = 0$.

Assumptions 1.1 and 1.2 and the law of large numbers jointly imply that the instantaneous share of produced income of class C_k at time t is

$$\begin{aligned} p_k(t) &= \frac{\left(\int_{w_k}^{w_{k+1}} x dP_{X_1(t)}(x)\right)\mathbb{P}(C_1(t)=\mathcal{C}_k)}{\sum_{k=1}^{K}\left(\int_{w_k}^{w_{k+1}} x dP_{X_1(t)}(x)\right)\mathbb{P}(C_1(t)=\mathcal{C}_k)} \\ &= \frac{\mathbb{E}(\mathbf{X}_1(t)|C_1(t)=\mathcal{C}_k)\mathbb{P}(C_1(t)=\mathcal{C}_k)}{\sum_{k=1}^{K}\mathbb{E}(\mathbf{X}_1(t)|C_1(t)=\mathcal{C}_k)\mathbb{P}(C_1(t)=\mathcal{C}_k)}. \end{aligned}$$

Hence, the multivariate process $\mathbf{p}(t) = (p_1(t),\ldots,p_K(t))_{t\in\mathbb{R}_+}$ becomes a deterministic function of time. Assumption 1.3 yields

$$p_k(t) = \frac{W_k\mathbb{P}(C_1(t)=\mathcal{C}_k)}{\sum_{k=1}^{K} W_k\mathbb{P}(C_1(t)=\mathcal{C}_k)}.$$

Theil's entropy $\mathbb{T}(t)$ of the system at time t thus becomes an explicit function of the probability distribution of $C_1(t)$, and hence, $\mathbb{T} = (\mathbb{T}(t))_{t\in\mathbb{R}_+}$ becomes a deterministic function of time t, explicitly determined by the generator $\mathbf{A}$ and the initial distribution ν of the allocation class process $\mathbf{C}_1$. Precisely,

$$\mathbb{T}(t) = \sum_{k=1}^{K} \frac{W_k(\nu\exp(t\mathbf{A}))_k}{\sum_{j=1}^{K} W_j(\nu\exp(t\mathbf{A}))_j} \log\left(K\frac{W_k(\nu\exp(t\mathbf{A}))_k}{\sum_{j=1}^{K} W_j(\nu\exp(t\mathbf{A}))_j}\right), \quad (2)$$

where $\nu\exp(t\mathbf{A})$ denotes the product of ν and $\exp(t\mathbf{A})$, and $(\nu\exp(t\mathbf{A}))_k$ is its k-th component.

Note that the dynamic Theil's entropy $\mathbb{T}$ can be expressed as a function T of the transition matrix $\mathbf{P}_\eta$ for time values t of the form $t = m\eta$, with $m \in \mathbb{N}$. Indeed, $\exp(m\eta\mathbf{A}) = \exp(\eta\mathbf{A})^m$ gives in (2),

$$\mathbb{T}(m\eta) = T(\mathbf{P}_\eta, m) := \sum_{k=1}^{K} \frac{W_k(\nu\mathbf{P}_\eta^m)_k}{\sum_{j=1}^{K} W_j(\nu\mathbf{P}_\eta^m)_j} \log\left(K\frac{W_k(\nu\mathbf{P}_\eta^m)_k}{\sum_{j=1}^{K} W_j(\nu\mathbf{P}_\eta^m)_j}\right). \quad (3)$$

Note that T is formally defined on the set of stochastic matrices on $E = \{\mathcal{C}_1,\ldots,\mathcal{C}_K\}$, differentiable on the open set of irreducible transition matrices.

It is well known that the processes $\mathbf{C}_i$, $i \in \mathbb{N}^*$, are ergodic. Let $\pi = (\pi_1, \pi_2, ..., \pi_K)$ denote its stationary distribution; see [8] for an explicit expression. Markov chains $\mathbf{D}_i$, $i \in \mathbb{N}^*$, are also ergodic with stationary distribution π. Particularly, π satisfies $\pi \mathbf{P}_\eta = \pi$.

Consequently we can compute the asymptotic value of the dynamic Theil's entropy as follows,

$$\begin{aligned}\mathbb{T}(\infty) := \lim_{t\to\infty} \mathbb{T}(t) &= \sum_{k=1}^{K} \frac{W_k \pi_k}{\sum_{j=1}^{K} W_j \pi_j} \log\left(K \frac{W_k \pi_k}{\sum_{j=1}^{K} W_j \pi_j}\right) \\ &= T_\infty(\pi),\end{aligned} \tag{4}$$

where T_∞ is differentiable on the set of probability measures supported by E.

4 Estimation of dynamic Theil's entropy from discrete observations

This section aims at studying sequences of plug-in estimators for both dynamic and asymptotic Theil's entropy built from discrete observations of income class allocation process of one (Section 4.1) or several (Section 4.2) agent(s).

4.1 Discrete observations of one agent's wealth

Let us suppose the income class allocation process $\mathbf{C}_1$ of one agent are observed at discrete times $m\eta$, up to time $M\eta$, where $M \in \mathbb{N}$, so that a truncated trajectory of the Markov Chain $\mathbf{D}_1$ is available. Let $\widehat{\mathbf{P}}_M$ be the empirical estimator of $\mathbf{P}_\eta$ built from $D_1(0), \ldots, D_1(M)$ as follows:

$$\widehat{\mathbf{P}}_M(k,l) := \begin{cases} \dfrac{\sum_{m=0}^{M-1} \mathbb{1}_{\{D_1(m)=\mathcal{C}_k, D_1(m+1)=\mathcal{C}_l\}}}{\sum_{m=0}^{M-1} \mathbb{1}_{\{D_1(m)=\mathcal{C}_k\}}} & \text{if } k \neq l \text{ and } \sum_{m=0}^{M-1} \mathbb{1}_{\{D_1(m)=\mathcal{C}_k\}} \neq 0, \\ 0 & \text{if } k \neq l \text{ and } \sum_{m=0}^{M-1} \mathbb{1}_{\{D_1(m)=\mathcal{C}_k\}} = 0 \\ 1 - \sum_{j\neq k} \widehat{\mathbf{P}}_M(k,j) & \text{if } k = l, \end{cases} \tag{5}$$

for $(k,l) \in \{1, \ldots, K\}^2$. It is well known (see for instance Billingsley [5]) that the sequence $(\widehat{\mathbf{P}}_M)_{M\in\mathbb{N}}$ is strongly consistent and that $(\sqrt{M}(\widehat{\mathbf{P}}_M - \mathbf{P}))_{M\in\mathbb{N}}$ is asymptotically normal, with zero mean and asymptotic variance matrix $\Sigma^2_{\mathbf{P}_\eta}$ equal to the inverse of the Fisher information of the chain $\mathbf{D}_1$.

For any $m \in \mathbb{N}$ let us define the estimator $\widehat{\mathbb{T}}_M(m\eta)$ of $\mathbb{T}(m\eta)$ built from $\widehat{\mathbf{P}}_M$ by plugging it into the function T; precisely

$$\widehat{\mathbb{T}}_M(m\eta) := T(\widehat{\mathbf{P}}_M, m).$$

Because T is differentiable on the set of stochastic matrices, the sequence $(\widehat{\mathbb{T}}_M(m\eta))_{M\in\mathbb{N}}$, for $m \in \mathbb{N}$, inherits of the good asymptotic properties of $(\widehat{\mathbf{P}}_M)_{M\in\mathbb{N}}$ – thanks to the continuous mapping theorem and the Delta method; see Van der Vaart[14] for details. Precisely, for all $m \in \mathbb{N}$,

$$\widehat{\mathbb{T}}_M(m\eta) \xrightarrow[M\to\infty]{\text{a.s.}} \mathbb{T}(m\eta),$$
$$\sqrt{M}\left(\widehat{\mathbb{T}}_M(m\eta) - \mathbb{T}(m\eta)\right) \xrightarrow[M\to\infty]{\mathcal{D}} \mathcal{N}(0, \Sigma^2_{\mathbb{T}(m\eta)}),$$

where

$$\Sigma^2_{\mathbb{T}(m\eta)} = \left(\frac{\partial}{\partial \mathbf{P}_\eta} T(\mathbf{P}_\eta, m)\right)^t \Sigma^2_{\mathbf{P}_\eta} \left(\frac{\partial}{\partial \mathbf{P}_\eta} T(\mathbf{P}_\eta, m)\right). \tag{6}$$

Similarly, for estimating the asymptotic Theil's entropy $\mathbb{T}(\infty)$, we define $\widehat{\mathbb{T}}_M(\infty)$ by plugging whether the stationary distribution $\widehat{\pi}_M$ of $\widehat{\mathbf{P}}_M$ if $\widehat{\mathbf{P}}_M$ is irreducible and aperiodic or any distribution supported by E if not, into T_∞ given by (4); precisely

$$\widehat{\mathbb{T}}_M(\infty) := \sum_{k=1}^{K} \frac{W_k \widehat{\pi}_M(\mathcal{C}_k)}{\sum_{j=1}^{K} W_j \widehat{\pi}_M(\mathcal{C}_j)} \log\left(K \frac{W_k \widehat{\pi}_M(\mathcal{C}_k)}{\sum_{j=1}^{K} W_j \widehat{\pi}_M(\mathcal{C}_j)}\right),$$

with the convention $0 \log 0 = 0$.

Again, the continuous mapping theorem and delta method ensure the strong consistency of $(\widehat{\mathbb{T}}_M(\infty))_{M\in\mathbb{N}}$ and asymptotic normality of $(\sqrt{M}(\widehat{\mathbb{T}}_M(\infty) - \mathbb{T}(\infty)))_{M\in\mathbb{N}}$, with asymptotic variance $\Sigma^2_{\mathbb{T}(\infty)} = \left(\frac{d}{d\pi} T_\infty\right)^t \Sigma^2_\pi \left(\frac{d}{d\pi} T_\infty\right)$, where Σ^2_π is explicit.

To estimate $\mathbb{T}(t)$ for $t \notin \eta\mathbb{N}$ is much more complicated. Indeed, (3) does not hold true anymore. Hence, it is required to estimate the generator of $\mathbf{C}_1$ from observations of the discretized chain $\mathbf{D}_1$ and then to plug this estimator into (2). The main mathematical obstacle is due to the non-injectivity of the exponential function for matrices : given the transition matrix $\mathbf{P}_\eta$ of $\mathbf{D}_1$, it may exists several different generators $\mathbf{A}$ such that $\exp(\eta\mathbf{A}) = \mathbf{P}_\eta$, so that the statistical model may not be identifiable. Regnault [11] shows that the model of birth-death processes with constant rates, i.e. $\lambda_k = \lambda > 0$ and $\mu_{k+1} = \mu > 0$ for all $k \in \{1, \ldots, K-1\}$, is identifiable. Indeed, the generators of such processes are proven to have K distinct eigenvalues, a sufficient condition for the uniqueness of the logarithm matrix of $\mathbf{P}_\eta$. Further investigations would be necessary to extend this result to the whole family of generators given by (1).

4.2 Discrete observations of several agents' wealth

Let us now suppose that the discretized income class allocation processes $\mathbf{D}_1, \ldots, \mathbf{D}_n$ of n independent agents up to time $M \in \mathbb{N}$ is observed. Let $\widehat{\mathbf{P}}^{(i)}_M$ be the maximum likelihood estimator of $\mathbf{P}_\eta$ given by (5), obtained from the trajectory $\mathbf{D}_i$, for each $i \in \{1, \ldots, n\}$ and let

$$\widehat{\mathbf{P}}_{M,n} := \frac{1}{n} \sum_{i=1}^{n} \widehat{\mathbf{P}}^{(i)}_M,$$

define the estimator of $\mathbf{P}_\eta$ obtained as mean of the estimators $\widehat{\mathbf{P}}_M^{(i)}$, $i \in \{1, \ldots, n\}$.

The sequence $(\widehat{\mathbf{P}}_{M,n})_{M\in\mathbb{N}}$ is obviously strongly consistent. Moreover, observe that $\sqrt{M}(\widehat{\mathbf{P}}_{M,n} - \mathbf{P}_\eta) = \frac{1}{n}\sum_{i=1}^n \sqrt{M}(\widehat{\mathbf{P}}_M^{(i)} - \mathbf{P}_\eta)$ is the mean of n independent, asymptotically normal variables, with the same asymptotic variance $\Sigma^2_{\mathbf{P}_\eta}$, so that

$$\sqrt{Mn}\left(\widehat{\mathbf{P}}_{M,n} - \mathbf{P}_\eta\right) \xrightarrow[M\to\infty]{\mathcal{D}} \mathcal{N}(0, \Sigma^2_{\mathbf{P}_\eta}).$$

Observing n agents in the place of a unique one thus leads to improve the rate of convergence by a factor $\sqrt{n}$.

Hence, the estimators $\widehat{\mathbb{T}}_{M,n}(m\eta)$ for $m \in \mathbb{N}$, obtained by plugging $\widehat{\mathbf{P}}_{M,n}$ into T inherit this asymptotic behavior. Precisely,

$$\widehat{\mathbb{T}}_{M,n}(m\eta) \xrightarrow[M\to\infty]{\text{a.s.}} \mathbb{T}(m\eta),$$
$$\sqrt{Mn}\left(\widehat{\mathbb{T}}_{M,n}(m\eta) - \mathbb{T}(m\eta)\right) \xrightarrow[M\to\infty]{\mathcal{D}} \mathcal{N}(0, \Sigma^2_{\mathbb{T}(m\eta)}).$$

Similarly, the sequence of estimators of $\mathbb{T}(\infty)$ given by $\widehat{\mathbb{T}}_{M,n}(\infty) = T_\infty(\widehat{\pi}_{M,n})$, where $\widehat{\pi}_{M,n}$ is the stationary distribution of $\widehat{\mathbf{P}}_{M,n}$, is strongly consistent and asymptotically normally distributed, with rate of convergence $\sqrt{Mn}$.

4.3 Confidence intervals for dynamic Theil's entropy

Confidence intervals for dynamic Theil's entropy at times $m\eta$, where $m \in \mathbb{N}$, can be derived from the asymptotic distribution of estimators $\widehat{\mathbf{P}}_M$ and $\widehat{\mathbf{P}}_{M,n}$, for $n \in \mathbb{N}^*$ and M large enough. Precisely, given a confidence level α – typically, $\alpha = 0.95$, the confidence interval for $\mathbb{T}(m\eta)$ is

$$]\widehat{\mathbf{P}}_{M,n} - \frac{q_{(1+\alpha)/2}\widehat{\Sigma}_M}{\sqrt{Mn}}; \widehat{\mathbf{P}}_{M,n} + \frac{q_{(1+\alpha)/2}\widehat{\Sigma}_M}{\sqrt{Mn}}[,$$

where $q_{(1+\alpha)/2}$ is the quantile of order $(1+\alpha)/2$ of the standard normal distribution and $\widehat{\Sigma}^2_M$ is any strongly consistent estimator of $\Sigma^2_{\mathbb{T}(m\eta)}$ built from independent truncated trajectories of $\mathbf{D}_1$ up to time M. Such an estimator can be obtained by computing the inverse Fisher's information matrix $\widehat{\Sigma^2_{\mathbf{P}_\eta}}^{(1)}_M$ of the empirical transition matrix $\widehat{\mathbf{P}}_M^{(1)}$ and hence by plugging both $\widehat{\Sigma^2_{\mathbf{P}_\eta}}^{(1)}_M$ and $\widehat{\mathbf{P}}_M^{(1)}$ into (6).

Although a thinner study could refine the following limitation, we recommend to use the normal asymptotic distribution for M larger than 100 units of time when $K \leq 5$, so that truncated trajectories $\mathbf{D}_1, \ldots, \mathbf{D}_n$ catch enough information about system's dynamics.

Other applications of the central limit theorem would be of interest, such as computing the minimal number n of customers to observe to ensure a given accuracy of the estimation of Theil's entropy, for a given $M \geq 100$.

References

1.G. D'Amico and G. Di Biase. Generalized Concentration/Inequality Indices of Economic Systems Evolving in Time. *WSEAS TRANSACTIONS on MATHEMATICS*, 9, 2, 140-149, 2010.
2.G. D'Amico, G. Di Biase and R. Manca. Immigration effects on economic systems through dynamic inequality indices. *Global journal of business research*, 5, 5, 11-25, 2011.
3.G. D'Amico, G. Di Biase and R. Manca. Income Inequality Dynamic Measurement of Markov Models: Application to some European Countries. *Economic Modeling*, 29, 1598-1602, 2012.
4.G. Athanasopoulos and F. Vahid. Statistical inference on changes in income inequality in Australia. *The Economic Record*, 79, 247, 412-424, 2003.
5.P. Billingsley. *Statistical Inference for Markov Process,* The University of Chicago Press, Chicago, 1968.
6.M. Bladt and M. Sørensen. Statistical inference for discretely observed Markov jump processes. *Journal of the Royal Statistical Society*, B, 67, 395-410, 2005.
7.F. Cowell. *Measuring Inequality.* LSE handbook in Economics Series, 2nd edition, prentice Hall, London, 1995.
8.V. Girardin and A. Sesboüé. Maximizing the entropy rate of state-dependent M/M/1 queues. *Bayesian Inference and Maximum Entropy Methods in Science and Engineering, AIP Conference Proceedings*, 1073, 181-188, 2008.
9.D. Quah. Empiric for economic growth and convergence. *European Economic Review* 40, 951-958, 1996.
10.P. Regnault. Différents problèmes liés l'estimation de l'entropie de Shannon d'une loi, d'un processus de Markov. *PhD Thesis, Université de Caen – Basse Normandie*, 2011.
11.P. Regnault. Entropy estimation of M/M/1 queueing systems, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering, AIP Conference Proceedings*, 1443, 330-337, 2012.
12.C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal* 27, 379-423, 1948.
13.H. Theil. *Economics and Information Theory,* North Holland, Amsterdam, 1967.
14.A.W. Van der Vaart. *Asymptotic Statistics,* Cambridge Series in Statistical an Probabilistic Mathematics, Cambridge, 1998.

Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013)
International Conference, Mataró (Barcelona), Spain 25 - 28 June 2013

Longitudinal Analysis of renal function with informative observation times

Lisandra Rocha[1], Inês Sousa[1,2], and Raquel Menezes[1,2]

[1] Center of Mathematics of University of Minho, Portugal
(E-mail: ID2704@alunos.uminho.pt)
[2] Department of Mathematics and Applications of University of Minho, Portugal
(E-mail: isousa@math.uminho.pt; rmenezes@math.uminho.pt)

Abstract. In this work, we intend to analyze the progression of renal function in patients of Hope Hospital in Salford, England. The aim of this work is to model the evolution of creatinine in the blood taking into account that time measurements depend on the responses obtained. For that, we propose a model that considers dependence between the longitudinal outcome process and the follow-up time process. This model use subject-specific random effects to describe the association between the two processes of interest and a joint-likelihood-based estimation was developed for estimation. We will also analyze the same data with a standard longitudinal model which ignore the dependence assumption, aiming to compare the results of the two analyzes. The principal difference between these two models and different approaches is that, the standard longitudinal model considers the follow up time process as deterministic while the model that we propose considers the follow up time process stochastic.
Keywords: Follow up Time Process, Longitudinal Analysis, Monitoring Kidney Function.

1 Introduction

Renal failure is a disease caused by the slow and irreversible loss of renal function. It is defined by progressive loss of the ability of the kidneys to filter, responsible for eliminating toxic products to the body and producing substances essential to life. Often, the loss of renal function is silent in the absence of the manifestation of symptoms for extended periods in time. There are already some risk factors known for kidney failure such as diabetes, hypertension and patient age.

For the control of renal failure, the creatinine level in blood should be monitored. The kidneys control levels of creatinine in the blood, therefore it is an indicator for evaluating kidneys performance. High levels of creatinine in the blood warn for possible failure in the functioning of the kidneys.

The times where measurements of creatinine are observed are decided according to the patient's clinical status. For example, if the patient on the last measurement had a high value of creatinine in the blood, the patient will be measured sooner than other patient with a creatinine value within normal values. Thus, in this context, the measurement times are considered dependent of the longitudinal responses and the follow up time process can not be considered deterministic in the study design.

In standard longitudinal models (Diggle et al.[4]), the follow up time process is assumed deterministic, meaning, the follow-up time process is noninformative about the outcome longitudinal process of interest. Therefore, this type of analysis does not considered the dependence that can exist between the follow-up time process and the longitudinal outcome process. In recent years, a number of authors have considered the situation where longitudinal outcome process is correlated with the follow up time process (Sun et al.[2]; Ryu et al.[1]; Liang et al.[6]).
In this work, we intend to analyze the progression of renal function in patients of Hope Hospital in Salford, England. The aim of this work is to model the evolution of creatinine in the blood taking into account that time measurements depend on the responses obtained. For that, we propose a model that considers dependence between the longitudinal outcome process and the follow-up time process (Rocha et al.[3]). This model use subject-specific random effects to describe the association between the two processes of interest. For estimation, a joint-likelihood-based estimation was developed. We also use a standard longitudinal model which assumes follow up time process as deterministic, aiming to compare the results of the three approaches.

2 The Data Set

For study the progression of renal function in patients of Hope Hospital in Salford, England, we used two data sets: one with primary and secondary care patients and other data set with healthy patients (Hoefield et al.[5]). The response variable of interest is the indicator eGFR (estimated gromelular filtration rate). Doctors suggest that a person with an annual rate of decrease in eGFR exceeding 5% should be referred to a specialist treatment center.
Below, we show some information about the two data sets.

Healthy patients

- 162.394 patients;
- Of these 162.394 individuals, 74.222 were males and 88.172 were female;
- 9% of the patients died during the study;
- The age range 18 to 106 years, and the average is about 57 years;
- The patients follow up was maximum of 2502 days (approximately 7 years). The average length of patients follow up is 616 days (approximately 1 year and 8 months).

Primary and secondary care patients

- 18.350 patients;
- Of these 18.350 individuals, 9.457 were males and 8.893 were female;
- 25.8% of the patients died during the study;
- The age range 18 to 106 years, and the average is about 70 years;
- The patients follow up was maximum of 3066 days (approximately 8 years and half). The average length of patients follow up is 938 days (approximately 2 years and half);

- In this data set we have information about some risk factors: 45.3% of the patients have diabetes, 17.3% had a myocardial infarction, 21.9% have hypertension, 20.3% of the patients take ACE drugs and 6.4% take AT2 Blockers drugs.

Figure 1 show the logarithm of eGFR versus patient age of two databases. The solid red line in the graphs represent the mean estimate of evolution using the function smooth.spline in software R. It is known that, as will the age of patients increases, renal function worsens. It is found that, from the age of 65, measurements of the logarithm of eGFR are worse. As expected, patients in primary and secondary, on average, have values of eGFR lower than healthy patients. This means that patients in primary and secondary care are with renal function in worse condition than healthy patients.

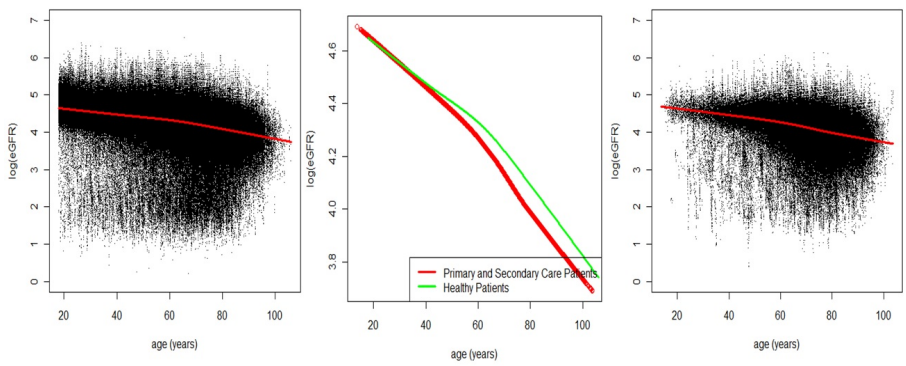


Fig. 1. Evolution of renal function in healthy patients and primary and secondary care patients

3 Notation and Model

Consider Y_{ij} the variable response observed in the individual $i = 1, , n$ at time $t_{ij} = 1, , m_i$ and X_{ij} a matrix of covariates with dimension Nxp. Consider the model defined as

$$Y_{ij} = \beta X_{ij} + W_i(t_{ij}) + U_i + \epsilon_{ij} \tag{1}$$

where β is the vector of regression parameters and ϵ_{ij} are independent realizations of a Gaussian variable with $E[\epsilon_{ij}] = 0$ and $Var[\epsilon_{ij}] = \tau^2$. Consider U_i as the random effect of individual i, where $U_i \sim Normal(0, \gamma^2)$ and $W_i(t_{ij})$ as the serial correlated variation for individual i, where $E[W_i(t_{ij})] = 0$, $Var[W_i(t_{ij})] = \sigma^2$ and $Corr[W_i(t_{ij}), W_i(t_{ij} - u)] = \rho(u)$. Also consider the follow up time process T_{ij} defined as

$$T_{ij} \sim Exp[\lambda_i(t_j)] \qquad \text{and} \qquad \lambda_i(t_j) = Exp[\alpha + \rho \int_0^{t_{i,(j-1)}} Y_i(t)dt] \tag{2}$$

where α and ρ are unknown parameters. The difference between the standard longitudinal model and our model is that, the standard longitudinal model considers the follow up time process T as deterministic while our model considers the follow up time process stochastic. The parameters are estimated by maximizing the likelihood function, using the function optim (for our model) and the function lme (for the longitudinal standard model) in software R.

4 Discussion

In this work, we propose a model for repeated measures in the presence of informative observation times. For the model adjusted to the data, the effect of age is negative. This simply means that will as, renal function decreases with age.
It was also observed that there is a changing point in age effect at 65 years of age where the decrease of renal function happens more rapidly. Is also the starting of 65 than the difference between the renal function of patients in primary and secondary and healthy patients is increased.
We believe that the model proposed is the most appropriate model to the data concerned.

5 Acknowledgements

The authors acknowledges the grant by FCT PTDC/MAT /112338/2009.
The author L. Rocha has a PhD scholarship by FCT SFRH/BD/61368/2009.
This research was financed by FEDER Funds through "Programa Operacional Factores de Competitividade COMPETE" and by Portuguese Funds through FCT - "Fundação para a Ciência e a Tecnologia", within the Project Est-C/MAT/UI0013/2011.

References

1.D. Ryu, D. Sinha, B. Mallick, S. Lipsitz and S. Lipshultz. Longitudinal studies with outcome-dependent follow-up: models and bayesian regression, *Journal of the American Statistical Association*, 102, 952-961, 2007.
2.J. Sun, D. Park, L. Sun and X. Zhao. Semiparametric regression analysis of longitudinal data with informative observation times, *Journal of the American Statistical Association*, 100, 882-890, 2005.
3.L. Rocha, I. Sousa and R. Menezes. A longitudinal model with informative observation times. (in preparation)
4.P. Diggle, K. Liang and S. Zeger. *Analysis of Longitudinal Data*, Oxford: Clarendon Press, 2002.
5.R. Hoefield, P. Kalra, P. Baker, I. Sousa, P. Diggle, M. Gibson, D. O'Donoghue, R. Middleton and J. New. The use of eGFR and ACR to predict decline in renal function in people with diabetes. *Nephrol Dial Transplant*, 1–6, 2010.
6.Y. Liang, W. Lu and Z. Ying. Joint Modeling and Analysis of Longitudinal Data with Informative Observation Times, *Biometrics*, 65, 377-384, 2009.