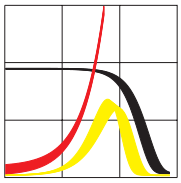

Max-Planck-Institut für demografische Forschung
Max Planck Institute for Demographic Research
Konrad-Zuse-Strasse 1 · D-18057 Rostock · GERMANY
Tel +49 (0) 3 81 20 81 - 0; Fax +49 (0) 3 81 20 81 - 202;
http://www.demogr.mpg.de

MPIDR WORKING PAPER WP 2006-041
NOVEMBER 2006

Inverse problems in demography and biodemography

Anatoli Michalski (mpoctok@narod.ru)

This working paper has been approved for release by: James W. Vaupel (jwv@demogr.mpg.de)
Head of the Laboratory of Survival and Longevity.

Inverse Problems in Demography and Biodemography

Anatoli Michalski

Institute of Control Sciences RAS, 117997, Profsoyuznaya 65, Moscow,

Russian Federation

(email: mpoctok@narod.ru)

and

The Max Plank Institute for Demographic Research,

Rostock, Germany

Abstract

Inverse problems play important role in science and engineering. Estimation of boundary conditions on the temperature distribution inside a metallurgical furnace, reconstruction of tissue density inside body on plane projections obtained with x-rays are examples. The similar problems exist in demography in the form of projection and estimation of population age distributions and age-specific mortality rates. The problem of residual demography is estimation of demographic process in wild nature on its manifestation in marked subjects with unobserved age, which again is inverse problem. The article presents examples and the ways of solution the inverse problems in demography and biodemography, discusses the ways of improving results by combination of demographic and genetic data.

Keywords: Inverse problem, ill-posed integral equation, regularization, epidemiology, HIV/AIDS, demography, cohort trends, residual demography, survival in wild, simulation, genetic data.

1 Introduction

In many fields of science and engineering a necessity exists to estimate a process using observations from another related with the estimated process. Depending on the problem setting it can be estimation of a signal in the presence of noise, numerical calculation of derivative, calculation of boundary conditions on the values of temperature distribution inside metallurgical furnace, reconstruction of tissue density inside body on plane projections obtained with x-rays. In nontechnical science similar problems arise in epidemiology, when a disease prevalence (proportion of sick people in different age groups) can be obtained but incidence rate (probability of healthy person to become sick during say one year) is to be estimated. In demography calculation of mortality rates form similar problem. Proportion of survived people is observed but chances to die during one year are of primer interest. In population projection and analysis we meet the same situation when one observes the population structure or some health related index at fixed years but is interesting in age structure or health related index levels at other not observed years.

All these problems form a class of mathematical problems called *inverse problems* contrary to *forward problems* when the estimation process follows cause-effect line. In this terms calculation of time graphic for distance covered by car using its velocity is forward problem but estimation the velocity using time graphic for distance in

inverse problem. Many important for practice inverse problems have solution which is very sensitive to disturbance in data. The example with velocity estimation is well known in technical applications where enormous efforts are applied to stabilize this kind of estimates using special frequency filters, time averaging or spline smoothing.

The same problem with solution sensitivity exists in inverse problems in epidemiology, demography, biodemography. Variations in data arise because of probabilistic nature of the process and limited number of observed people. The first means that the process can be attained only by its realizations in population in form of proportion of people in the defined states. The second means that these proportions differ from the probabilities. In the result one has extremely large changes in the estimates even when amount of data increases.

Formal consideration of inverse problems and procedures for solution stabilization is given in the next section. Examples of inverse problems from epidemiology (AIDS/HIV epidemic), demography (population dynamic), biodemography (estimation in wild on observations in laboratory) are given in section 3. Possible solutions for demography and biodemography problems are presented in sections 4 and 5. Section 6 contains discussion of possibility to improve the precision of estimates combining demographic and genetic data, section 7 contains conclusion.

2 Formal definition of inverse problem

In formal terms inverse problem is a problem of solution of an operator equation

$$Ax = y \qquad (1)$$

where A is a bounded linear operator between infinite dimensional functional Hilbert spaces *X* and *Y*, *x* and *y* are elements from these spaces. Function *y* plays role of "observations" or "effect", function *x* plays role of "cause" produced observed effect. It is supposed that operator *A* makes one to one mapping between spaces *X* and *Y*. The solution of equation (1) is a function, defined as

$$x = A^{-1}y$$

where A^{-1} is inverse to A operator which is linear as well. It is proved that if the range $\mathcal{R}(A)$ for A is non-closed, then operator A^{-1} is unbounded (Tikhonov and Arsenin, 1977; Engl *et al*., 1996). The latter means that if one substitutes a "disturbed" function $y^{\delta} \in Y$ such that $\left\|y - y^{\delta}\right\| \le \delta$ in (1) then the disturbance in corresponding solution $\left\|x - A^{-1}y^{\delta}\right\|$ may infinite. Here $\|x\|$ denotes the norm of *x*. More precisely: let δ_n be a set of nonnegative values tending to zero with *n* tending to infinity. For any small value δ and any large value Δ exists a function $y_{\Delta}^{\delta} \in Y$ such that $\left\|y - y_{\Delta}^{\delta}\right\| \le \delta$ and $\left\|x - A^{-1}y_{\Delta}^{\delta}\right\| > \Delta$.

To illustrate this fundamental property consider an integral equation for $t \in [0,1]$

$$\int_0^t x(\tau)d\tau = y(t) \qquad (2)$$

with linear bounded operator. The inverse operator is the operator of differentiation and

$$x(t) = \frac{d}{dt}y(t).$$

Consider a "disturbed" function $y^{\omega}(t) = y(t) + \sqrt{\omega}\sin(t/\omega)$ for which

$$x^{\omega} = A^{-1}y^{\omega} = \frac{d}{dt}\left(y(t) + \sqrt{\omega}\sin(t/\omega)\right)$$
$$= \frac{d}{dt}y(t) + \frac{1}{\sqrt{\omega}}\cos(t/\omega)$$

It is obvious that $\left\|y - y^{\omega}\right\| \xrightarrow[\omega\to 0]{} 0$ while $\left\|x - x^{\omega}\right\| \xrightarrow[\omega\to 0]{} \infty$. This means that for any small value δ and any large value Δ can be found small enough value ω such that $\left\|y - y^{\omega}\right\| \le \delta$ and $\left\|x - x^{\omega}\right\| > \Delta$.

The formulated property of inverse operator make impossible to guarantee that solution found on disturbed data will be close to solution, corresponding to undisturebed data. To make the inverse operator to be bounded one needs to make range $\mathcal{R}(A)$ for A to be closed. This can be done by reducing the dimension of functional spaces *X* and *Y* to finite values which corresponds to parameterization of functions *x* and *y* (in space of finite dimension linear nonsingular operator has bounded inverse operator) or by applying additional restrictions on solution of (1). In the case of example (2) this can be restriction on the maximum value of the first derivative of the solution. More methods and formal proves can be found in (Morozov, 1993).

The general approach to solution of equations with unbounded inverse operator, called ill-posed equations, is formulated in Tikhonov and Arsenin (1977) as minimization of regularized functional $J_{\alpha}(x) = \left\|Ax - y^{\delta}\right\|^2 + \alpha\left\|Bx\right\|^2$, where $\alpha > 0$ is a regularization parameter, B is unbounded operator defined at functional set $\mathcal{D}(B) \subseteq X$ such that $Bx \subseteq X$. Minimization is to be done in $\mathcal{D}(B)$ - the region of definition of operator B. The problem of proper selection of regularization parameter value is widely discussed in literature. For special case $B = D^s$, where D is a differential operator and s is some

nonnegative real number, Natterer (1984) has shown that under the assumptions $\left\|D^p A^{-1} y\right\| \le E$ and $m\left\|D^{-\alpha} x\right\| \le \|Ax\| \le M\left\|D^{-\alpha} x\right\|$ with some constants *E*, *m* and *M*, regularized solution x_α provides approximation of the real solution with bound $\left\|x_\alpha - A^{-1} y\right\| = O\left(\delta^{p/(\alpha+p)}\right)$ for $s \ge (p-\alpha)/2$ if α is chosen priory as $\alpha = c\delta^{2(\alpha+s)/(\alpha+p)}$ with some constant *c*. Posterior selection of regularization parameter can be done using Morozov's discrepancy principle (Morozov, 1993) which prescribes to select parameter α as solution of the equation $\left\|Ax_\alpha - y^\delta\right\| = C\delta$, where $C \ge 1$ is a constant. The efficiency of this approach has been proved in many applications (Nair et al., 2003; 2005). Procedures of regularization parameter selection in case of stochastic disturbances in y^δ are considered in Michalski (1987), Lukas (1998), Engl *et al.* (2005).

3 Examples of inverse problems in epidemiology, demography and biodemography

Specific operator equations arise in consideration of estimation problems in epidemiology, demography and biodemography. We will describe the problems of estimation the number of HIV infected on dynamics of AIDS cases (epidemiology), forward and back projection of population structure and population health related indexes (demography), reconstruction of survival in wild on survival of captured animals in laboratory (residual demography).

Epidemiology

In epidemiology event of infection causes with some probability event of the disease development which in turn with some probability causes the diagnoses establishing. If the time lags between these three events are not large and the probabilities are not small then the dynamics of the disease diagnosed cases reflects the infection process. This is not a case of HIV/AIDS epidemic. The lag between HIV infection and AIDS manifestation (incubation period) is reported to last up to 15 years. There are some evidence that the shorter the incubation period is the more amount of virus was transmitted in the blood. Cases of HIV infection at young ages are characterized by enlargement of incubation period (Gigli and Verdecchia, 2000). These conditions make epidemic HIV/AIDS not only specific in terms of demographic consequences but to be difficult for monitoring and control. Better understanding of HIV/AIDS epidemics can be gained using inverse problems approach. Denote $\psi(t,x)$ HIV infection rate at age *x* in time *t*, $\mu_c(t,x)$ total mortality at age *x* in time *t*, $L(x,s)$ probability density function to develop AIDS at age *x* being infected wit HIV at age *s*, $u(t,x)$ prevalence of diagnosed AIDS cases at age *x* in time *t*. The relationship between AIDS prevalence and HIV infection rate is given by

$$u(t,x)=\int_0^x L(t,s)\exp\left(-\int_s^x \mu_c(t-x+\tau,\tau)d\tau\right)\psi(t-x+s,s)ds$$

which is integral equation in respect to $\psi(t,x)$. Obtaining estimates for $\psi(t,x)$ one can calculate the total number of HIV infected in population, prevalence of HIV positive people by age groups, age and time trends in HIV infection process. More details and examples can be found in Michalski (2005).

Demography

Population projections are based on current population structure, projections of mortality and birth rates, scenarios of in and out migration. One can skips birth rates by projecting cohort dynamics, but mortality rates are to be estimated on numbers of deaths and numbers of alive people. The estimate for age-specific mortality rate used in demography is $m_{xt} = \frac{d_{xt}}{n_{xt}}$, where d_{xt} - number of deaths in age group *x* in year *y*, n_{xt} - number of people alive in age group *x* in the beginning of year *y*. Because of small number of people alive in advanced age groups the estimate m_{xt} has high variance and is to be improved by application additional data and approaches. One possibility is to consider mortality as solution of integral equation

$$S(x) = 1 - \int_0^x \mu(y) \exp\left(- \int_0^y \mu(\tau) d\tau \right) dy , \qquad (3)$$

which follows from consideration of two-states Semi-Markov model with one state "alive", the other state "deceased", rate of transition from the first state to the second one $\mu(y)$ and probability to stay in the "alive" state – survival function $S(x)$. Solution of equation (3) is equivalent to calculation of logarithmic derivative $\mu(x) = -\frac{d}{dt} \ln S(x)$ which means that (3) has unstable solution. One possibility to reduce instability is proper parameterization of mortality as is done in Lee and Carter (1992).

Similar to (3) equation with unstable solution emerges in estimation of cohort gradients in population characteristics, say in risk factors, health related indicators. Denote $h(x,t)$ a value for an indicator at age *x* and time *t*, $g(x,t)$ a value of cohort

gradient in health indicator at age *x* and time *t*. Relationship between $h(x,t)$ and $g(x,t)$ is given by cohort dynamic equation

$$h(x,t) = h(x_0,t_0) + \int_0^{x-x_0} g(x_0+\tau, t_0+\tau) d\tau .$$

Effective way of this equation solution using data from cross-sectional surveys is presented in Moltchanov *et al.* (2005).

Survival in the wild

Muller *et al.* (2004) formulated a problem of investigation the survival in flies living in wild on the observation of survival in flies captured at unknown age and kept in laboratory. This direction in demography is entitled *residual demography*. The experiment setting is as follows. A flies in wild are captured at random and put in laboratory where the captured cohort is observed and proportion of survivors till day x after capture $P_c(x)$ is calculated. At the same time a cohort of flies is reared from fruits, collected in the same region where the flies were captured. This cohort has known age and is called a reference cohort. Survival $S_r(x)$ observed in this cohort is reference survival, reflecting survival in laboratory conditions, while in captured cohort function $P_c(x)$ is not survival because of unknown age at capture. The problem is how to estimate survival in wild nature $S_w(x)$ using functions $S_r(x)$ and $P_c(x)$.

This is typical inverse problem where survival in wild $S_w(x)$ causes proportion of survivors among captured flies $P_c(x)$. Probability for a fly captured at age *a* to survive in

laboratory x days is $P\{X > x \mid a\} = S_r(a+x)/S_r(a)$, where X is life span in laboratory. Probability for a fly captured random to survive in laboratory x days is

$$P\{X > x\} = E_a(P\{X > x \mid a\}) = \int_0^\omega \frac{S_r(a+x)}{S_r(a)} p_w(a)da$$

where $p_w(a)$ denotes probability for a wild fly to be captured at age interval $[a, a+da]$. In the case of stationary wild population $p_w(a) = \frac{1}{e_0} S_w(a)$ and

$$P_c(x) = \frac{1}{e_0} \int_0^\omega \frac{S_r(x+a)}{S_r(a)} S_w(a)da\,, \qquad (4)$$

where e_0 is life expectancy at birth in wild. Equation (4) is typical convolution equation with kernel function $K(x,a) = S_r(x+a)/S_r(a)$.

4 Lee-Carter Method and Dynamic Regression Method

Demographers want to predict age structure of population as in future so in the past. The first case is prognosis for planning while the second case is for interest of historical demography. Lee and Carter (1992) proposed to use for mortality forecasting presentation $m_{xt} = \exp(a_x + b_x k_t)$, which is a decomposition of mortality at age effect and at time effect. Estimates for vectors a_x, b_x and k_t are obtained by minimization least square error between model and given data

$$\sum_{x,t} \left(\ln\left(\frac{d_{xt}}{n_{xt}}\right) - a_x - b_x k_t \right)^2 \xrightarrow[a_x, b_x, k_t]{} \min$$

or weighted least square error

$$\sum_{x,t} d_{xt}\left(\ln\left(\frac{d_{xt}}{n_{xt}}\right)-a_x-b_x k_t\right)^2 \xrightarrow[a_x,b_x,k_t]{} \min .$$

The estimates can be obtained by iterating the three expressions Wilmoth (1993)

$$\hat{a}_x=\frac{1}{\sum_t d_{xt}}\sum_t d_{xt}\left(\ln\left(\frac{d_{xt}}{n_{xt}}\right)-\hat{b}_x\hat{k}_t\right),$$

$$\hat{b}_x=\frac{1}{\sum_t d_{xt}\hat{k}_t^2}\sum_t d_{xt}\hat{k}_t\left(\ln\left(\frac{d_{xt}}{n_{xt}}\right)-\hat{a}_x\right),$$

$$\hat{k}_x=\frac{1}{\sum_t d_{xt}\hat{b}_t^2}\sum_t d_{xt}\hat{b}_t\left(\ln\left(\frac{d_{xt}}{n_{xt}}\right)-\hat{a}_x\right)$$

starting from the proper guess values.

Some modifications of the Lee-Carter method for inverse projection of population structure, age distribution of death, reproduction patterns are presented in Barbi *et al.* (2004). Among them are *differential inverse projection*, based on separation of mortality on child and adult mortality. In this case probability of dying by age is presented as

$$q_{xt}=\begin{cases} q_{xt_0}k_t^1 & x\le 4 \\ q_{xt_0}k_t^2 & x>4 \end{cases},$$

where q_{xt} - probability of dying in year *t*, q_{xt_0} - probability of dying in reference year t_0, k_t^1 and k_t^2 are age group specific tome effects. A stochastic inverse projection is described in Barbi *et al.* (2004) in form

$$N(x,t)=N(x+1,t+1)+M(x,t),$$

where $N(x,t)$ is distribution of population by age *x* in year *t*, $M(x,t)$ is distribution of deaths by age *x* in year interval (*t*,*t*+1), which is calculated by simulation of death events with probability

$$P(x \mid t+u) = \frac{[SV(x,t) - N(x,t+u)]\dfrac{\pi(x+u)}{1-\pi(x+u)}\mu(x,t)}{\sum_{y=0}^{\omega-1}[SV(y,t) - N(y,t+u)]\dfrac{\pi(y+u)}{1-\pi(y+u)}\mu(y,t)}$$

where $SV(x,t)$ is the random total number of survivors of age *x* at time *t*, simulated with the program from the time series of the people born from year $t-\omega+1,(t-\omega+2)$ to year $(t-1,t)$ by surviving functions induced by the mortality rates $\mu(x,t)$. Probability for a person born at time *t-y* to survive a period *y*+*u* is denoted as $\pi(y+u)$. In the formula quantity *u* denotes a set of times $u = r\Delta t$ with $r = M^*(t)-1,...,1$ and $M^*(t)$ equals to number of individuals, who died in the interval (*t*,*t*+1).

The Lee-Carter method and its modifications explore population dynamics and uses parametric by time presentation for mortality. Many related to population health indices (*risk factors*) can not be presented in parametric form but should be reconstructed nonparametrically as a function of age and time. This can be done by Method of Dynamic Regression (Moltchanov *et al.*, 2005) in which cohort dynamics of the indicator is given by equation

$$h(x,t) = h(x_0,t_0) + \int_0^{x-x_0} g(x_0+\tau, t_0+\tau)d\tau, \qquad (5)$$

where $h(x,t)$ is a level of an indicator at age *x* and time *t*, $g(x,t)$ is a value of cohort gradient for the indicator at age *x* and time *t*, (x_0,t_0) - initial point of the cohort observation. To formulate the problem in discrete time and age a supposition is made that

the cohort gradient $g(x,t)$ takes constant value writhing the parallelogram $\mathbf{P_{ij}} = \{x,t : i \le t < i+1, t+j-i-1 < x \le t+j-i\}$ and level $h(x,t)$ takes constant value by age *x* writhing the same parallelogram. With proper indexing equation (5) takes form $h(x,t) = h(i,j) + g(i,j) \times (t-i)$. Denote h_k level of the index observed at the point (x_k, t_k) and present it in form using (5) for the cohort, which at time t_k was x_k years old

$$h_k = h_o^k + \sum_{m=1}^{\delta_k} g(i-m, j-m) + (t_k - i) \times g_{ij} + \varepsilon_k \qquad (6)$$

where h_0^k - initial level, corresponding to the beginning of the cohort observation, δ^k - number of years between t_k and t_0 , corresponding to the beginning of the cohort observation, ε_k a random term with zero mean value. In the case of rectangular region of investigations where $i = 0,...,I$, $j = 0,...,J$ (6) leads to matrix equation

$$\mathbf{h} = \mathbf{Bz} + \boldsymbol{\varepsilon} \qquad (7)$$

with $\mathbf{B}$ - matrix composed by 1th and 0th, $\mathbf{h} = (h_1,...,h_k)$, $\mathbf{z} = (\mathbf{h_0} \mid \mathbf{g})$, $\mathbf{h_0} = (h(I+1,0),...,h(0,0),...,h(0,J+1))^{\mathbf{T}}$, $\mathbf{g} = (g(0,0),...,g(0,J),...,g(I,0),...,g(I,J))^{\mathbf{T}}$, $\boldsymbol{\varepsilon} = (\varepsilon_1,...,\varepsilon_k)$, $E(\boldsymbol{\varepsilon}) = 0$, $Cov(\varepsilon_l, \varepsilon_m) = 0$. Because of ill-posed nature of equation (5) solution of equation (7) is to be done with stabilization procedures. In the Method of Dynamic Regression two types of stabilization are implemented: smoothing and aggregation (Moltchanov *et al.*, 2005).

Smoothing is done by constrained minimization

$$L(\mathbf{z}) = (\mathbf{h} - \mathbf{Bz})(\mathbf{h} - \mathbf{Bz})^{\mathbf{T}} \xrightarrow[\mathbf{z} \in L_\alpha]{} \min$$

$$L_\alpha = \{\mathbf{z} : \mathbf{z}^{\mathbf{T}} \mathbf{B}_1^{\mathbf{T}} \mathbf{B}_1 \mathbf{z} \le \alpha\},$$

which leads to unconstrained minimization problem

$$L_\lambda(\mathbf{z}) = (\mathbf{h} - \mathbf{B}\mathbf{z})(\mathbf{h} - \mathbf{B}\mathbf{z})^{\mathrm{T}} + \lambda \mathbf{z}^{\mathrm{T}} \mathbf{B}_1^{\mathrm{T}} \mathbf{B}_1 \mathbf{z} \xrightarrow[\mathbf{z}]{} \min .$$

The matrix of the first differences can be used as stabilization matrix

$$\mathbf{B}_1 = \begin{pmatrix} -1 & 1 & 0 & \dots \\ 0 & -1 & 1 & \dots \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots \end{pmatrix}.$$

Aggregation reduces the dimension of the problem by assigning the same values of cohort trends in several adjusted parallelograms. In this case

$$\mathbf{g} = \mathbf{G}\mathbf{f}$$

with $\mathbf{G}$ - matrix composed by 1th and 0th with number of rows equal to length of vector $\mathbf{g}$ and number of columns less than length of vector $\mathbf{g}$, $\mathbf{f}$ is vector of aggregated values for cohort trends. Recall that $\mathbf{B} = (\mathbf{B}_{\mathbf{h}_0} \quad \mathbf{B}_{\mathbf{g}})$ and write

$$\begin{aligned} \mathbf{B}\mathbf{z} &= \mathbf{B}_{\mathbf{h}_0}\mathbf{h}_0 + \mathbf{B}_{\mathbf{g}}\mathbf{g} \\ &= \mathbf{B}_{\mathbf{h}_0}\mathbf{h}_0 + \mathbf{B}_{\mathbf{g}}\mathbf{G}\mathbf{f} \end{aligned}.$$

Estimation of initial levels and cohort trends is doe by minimization of the functional

$$L_G(\mathbf{z}) = (\mathbf{h} - \mathbf{B}_{\mathbf{h}_0}\mathbf{h}_0 - \mathbf{B}_{\mathbf{g}}\mathbf{G}\mathbf{f})(\mathbf{h} - \mathbf{B}_{\mathbf{h}_0}\mathbf{h}_0 - \mathbf{B}_{\mathbf{g}}\mathbf{G}\mathbf{f})^{\mathrm{T}} \xrightarrow[\mathbf{h}_0, \mathbf{f}]{} \min .$$

Smoothing can used in combination of aggregation as well.

5 Survival in the wild nature

In animals links between age and mortality, fecundity, disability can be made only in laboratory conditions, where the date of the animal birth is recorded. Survival in wild nature can be assessed only by observation in laboratory of life spans after moment of the

animal capture. Under the hypotheses of stationarity of wild population the survival in wild and probability to survive *x* days in laboratory are linked by equation (4). This equation can be simplified under a hypothesis that mortality in laboratory does not differ from mortality in wild, which does not looks realistic but can be used as a starting point if the reference cohort is not available. Equation (4) reduces to

$$P_c(x) = \frac{1}{e_0}\int_0^{\omega} S_w(x+a)da = \frac{1}{e_0}\int_x^{\omega} S_w(a)da\,. \qquad (8)$$

It is easy to obtain by differentiation an analytical solution for this equation in form

$$S_w(x) = \frac{d}{dx}P_c(x) / \frac{d}{dx}P_c(0)\ .$$

To estimate survival in wild one is to estimate the derivative from the probability to survive in laboratory in captured animals $P_c(x)$. This leads to unstable solution. Muller *et al.* (2004) used non-parametric kernel density estimation for the derivative estimation

$$\frac{d}{dx}P_c(x) = \frac{-1}{nh(n)}\sum_{i=1}^{n} K\left(\frac{x - x_i^*}{h(n)}\right),$$

and estimation of its value

$$\frac{d}{dx}P_c(0) = \frac{-1}{nh(n)}\sum_{i=1}^{n} K_0\left(\frac{-x_i^*}{h(n)}\right)$$

where $x_1^*,...,x_n^*$ is a sample of observed life spans after capture, $h(n)$ is a sequence of proper selected values for bandwidth. The kernel functions were defined for the sample normalized at [-1,1] in the derivative estimation and [-1,0] in estimation the derivative value at *x*=0 $K(x) = 0.75(1-x^2)$, $K_0(x) = 12(x+1)(x+0.5)$. The asymptotic confidence intervals for the estimate are presented in Muller *et al.* (2004).

The alternative way of estimation survival in the wild is numerical solution of equation (8) which leads to a matrix equation

$$P_c = \mathbf{A}S_w \tag{9}$$

with triangular matrix

$$\mathbf{A} = \begin{pmatrix} 1/e_0 & 1/e_0 & \dots & 1/e_0 \\ 0 & 1/e_0 & \dots & 1/e_0 \\ 0 & 0 & \dots & 1/e_0 \\ 0 & 0 & \dots & 1/e_0 \end{pmatrix}.$$

To investigate the solution of equation (9) a numerical investigation was done. The "survival in wild" was modeled by survival in reference cohort of flies reared in laboratory (J.Carey's data). The graph for this survival is presented in figure 1 by empty circules. By multiplication this survival curve by matrix *A* the "survival among captured" has been produced. Small disturbances in this curve were added to simulate an effect of survival estimation by finite number of animals in captured group. Resulting survival curve is presented in figure 1 by crosses.

Figure 1 presents the results of solution of equation (9) by regularization functional minimization

$$J_\alpha(S) = (P_c - \mathbf{A}S)^{\mathbf{T}}(P_c - \mathbf{A}S) + \alpha S^{\mathbf{T}}\mathbf{B}^{\mathbf{T}}\mathbf{B}S$$

where $\mathbf{B}$ is matrix of the first derivatives like matrix $\mathbf{B}_1$ above. Dashed line in figure 1 presents solution obtained for *a*=0.001. One can see instability in solution. Solid line in figure 1 presents solution obtained for *a* selected by statistical elimination criterion (Michalski, 1987). Procedure of the regularization parameter selection is as follows. Denote by S_w^α solution of the problem

$$J_a(S) \xrightarrow[S]{} \min$$

under fixed value for the regularization parameter α. The value for regularization parameter α is selected by minimization on α of an expression

$$I(\alpha) = J_\alpha\left(S_w^\alpha\right) \Big/ \left(1 - \frac{2}{m} \operatorname{Tr} \mathbf{A}\left(\mathbf{A}^{\mathbf{T}}\mathbf{A} + \alpha \mathbf{B}^{\mathbf{T}}\mathbf{B}\right)^{-1} \mathbf{A}^{\mathbf{T}}\right).$$

Minimization was done in the range of α where criterion $I(a)$ is positive and S_w^α is monotone decreasing by age, m – number of rows in matrix $\mathbf{A}$.

6 Combination of demographic data with genetic data

The notion *inverse problem* is conditional in a sense that there exists the other problem in respect to which the considering inverse problem is a forward problem. For example senescence can be considered as a cause for mortality increase with age, which has reflection in longevity. Senescence process in turn in significant proportion is determined by genetics. In this example the same process can be the cause of some effect and the effect of some other cause. If we want to assess senescence based on longevity data we have an inverse problem with all troubles, related to instability of solution. Alternatively we can take into account the links between genetics and senescence and solve the problem in two steps. First on longevity data we estimate the proportions of different alleles in investigated group and then assess the conditions of health taking into account the estimated proportions. The large errors in estimates, made at the first step in solution of the inverse problem, may be not important in estimation of related health conditions and mortality. This leads to improvement in the result in comparison with solution, which does not take into account genetic information.

The other possible effect of using genetic data is employment of genetic links between relatives which will stabilize the solution in comparison with the case when all members of a family are considered as independent persons. Combination of genetic and demographic data is considered in Yashin et al. (1998), Tan et al. (2004[a], 2004[b]). Simulation studies were conducted in Begun and Yashin (2005). The results show big potential of combining data of different types.

7 Conclusion

Presented in the paper consideration of inverse problems in different branches of science demonstrate an example of a unified methodology, which can be effective in demography and biodemography. Implementation of inverse problem approach allows to improve the precision of traditional demographic methods and construct new models. This approach will be especially effective in solution of complex problems of regulation in human health and longevity, in investigation of links between longevity, genetics and environment conditions.

References

Barbi E., Bertino S., Sonio E. (Eds.) (2004) *Inverse Projection Techniques. Old and New Approaches*. Demographic Research Monographs, Springer, Berlin [et al.].

Begun A.Z., Yashin A.I. (2005) Genetic markers data in survival studies of twins: the results of a simulation study. Twin Res Hum Genet. 8: 34-8.

Engl H.W., Hanke M. and Neubauer A. (1996) *Regularization of Inverse Problems*. Dordrecht: Kluwer.

Engl H.W., Hofinger A. and Kindermann S. (2005) Convergence rates in the Prokhorov metric for assessing uncertainty in ill-posed problems. *Inverse Problems* 21: 399-412.

Gigli A., Verdecchia A. (200) Uncertainty of AIDS incubation time and its effects on back-calculation estimates. *Statistics in Medicine* 19: 175-189.

Lee R. D. and Carter L.R. (1992) Modeling and Forecasting U.S. Mortality. *JASA* 87: 659-671.

Lukas M.A. (1998) Comparison of parameter choice methods for regularization with discrete noisy data. *Inverse Problems* 14**:** 161–184.

Michalski A. I. (1987). Choosing an algorithm of estimation based on samples of limited size. Automatization and Remote Control. 48: 909-918.

Michalski A.I. (2005) Estimation of HIV infected number in population on the dynamics of observed AIDS cases. In Denisov B.P. (ed.) *Demography of HIV, Population and Crisises*, 11, MSU: 75-99 (*in Russian*).

Moltchanov V., Sarti C., Antikainen R. and Tuomilehto J. (2005) Application of the Dynamic Regression Method (DRM) to the assessment of the Body Mass Index dynamics in population using sequential cross-sectional survey data *EUROPEAN CONFERENCE ON CHRONIC DISEASE PREVENTION*, Helsinki**.**

Morozov V.A. (1993) Regularization Methods for Ill-Posed Problems. Florida: CRC Press.

Muller H.-G., Wang J.-L., Carey J.R., Caswell-Chen E.P., Chen C., Papadopoulos N., and Yao F. (2004) Demographic window to aging in the wild: constructing life

tables and estimating survival functions from marked individuals of unknown age. *Aging Cell* 3**:** 125-131.

Nair M.T., Schock E. and Tautenhahn U. (2003) Morozov's Discrepancy Principle under General Source Conditions. *Journal for Analysis and its Applications* 22: 199–214.

Nair M.T., Pereverzev S.V. and Tautenhahn U. (2005) Regularization in Hilbert scales under general smoothing conditions. *Inverse Problems* 21: 1851–1869.

Natterer F. (1984) Error bounds for Tikhonov regularization in Hilbert scales. *Appl. Anal.* 18**:** 29–37.

Tan Q., De Benedictis G., Yashin A.I., Bathum L., Christiansen L., Dahlgaard J., Frizner N., Vach W., Vaupel J.W., Christensen K., Kruse T.A. (2004[a]) Assessing Genetic Association with Human Survival at Multi-Allelic Loci. *Bioger.* 5: 89-97.

Tan Q, Yashin AI, Christensen K, Jeune B, De Benedictis G, Kruse TA, Vaupel JW (2004[b]) Multidisciplinary approaches in genetic studies of human aging and longevity. *Current Genomics* 5: 409-416.

Tikhonov A.N. and Arsenin V.Y. (1977) Solution of Ill-Posed Problems. New York: Wiley.

Wilmoth J.R. (1993) Computational methods for fitting and extrapolating the Lee-Carter model of mortality change. *Technical report, Department of Demography, University of California, Berkeley*.

Yashin A.I., Vaupel J.W., Andreev K.F., Tan Q.; Iachine I.A., Carotenuto L., De Benedictis G., Bonafe M., Valensin S., Franceschi C. (1998) Combining genetic and demographic information in population studies of aging and longevity. *Journal of Epidemiology and Biostatistics* 3: 289-294.

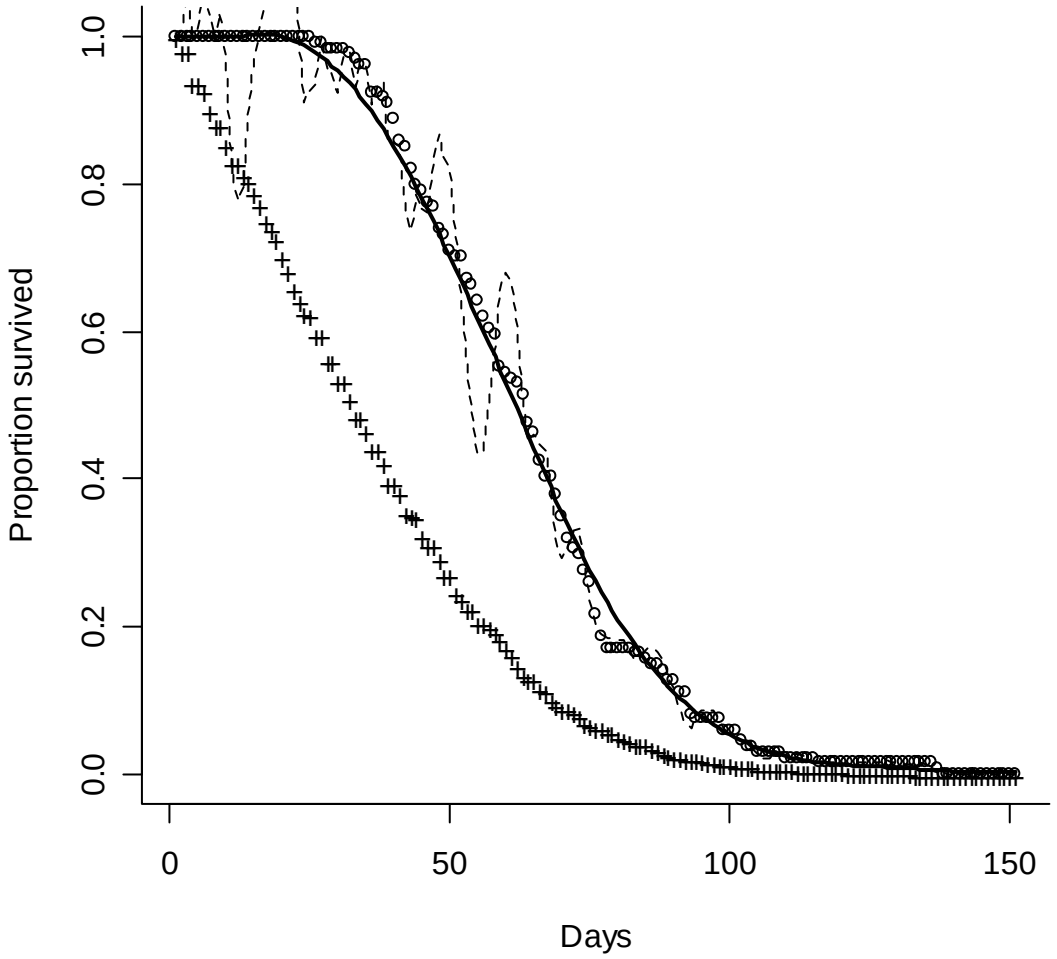

Figure 1. Survival in wild (open circuits), calculated and randomly disturbed survival among captured flies (crosses), estimate for survival in wild corresponding to small value for regularization parameter (dashed line), estimate for survival in wild corresponding to selected value for regularization parameter (solid line).