CHARLES UNIVERSITY IN PRAGUE

Faculty of Science

Department of Demography and Geodemography

Study program: Demography

Mgr. Klára Hulíková Tesárková

SELECTED METHODS OF MORTALITY ANALYSIS FOCUSED ON ADULTS AND THE OLDEST AGE-GROUPS

Vybrané způsoby zkoumání procesu úmrtnosti se zaměřením na dospělou populaci a nejvyšší věkové skupiny

Doctoral Thesis

SELECTED PART – SAS MACRO

Thesis Supervisor: RNDr. B. Burcin, Ph.D.

Prague, 2012

Chapter 5

Smoothing of mortality rates using the SAS software

5.1 Introduction

SAS is modern statistical software and its possibilities of usages in demography are not yet fully discovered. There are many important advantages within the SAS software. First of all many inbuilt procedures could be used and so many difficult tasks could be solved quite easily. Secondly, the usage of SAS software enables to solve many problems and tasks in one moment or to repeat a procedure many times even for large data sets. As a result the solution is efficient and quick. Thanks to macros, the procedures or whole programs could be easily repeated or adjusted to another data. On the other hand, it must be pointed out that it is quite difficult to use the SAS software for someone who does not know even the basics of the SAS programming code (or language).

For the purpose of this work a simple macro (with several other sub-macros inbuilt) was prepared and it is part of the electronic Appendix of this Thesis. Also some model data are distributed with it. The main purpose of this macro is to produce the estimate of unknown parameters of several functions of mortality smoothing (mortality laws) and so to produce the first columns of the life table (that means the smoothed probabilities of dying). These values can be easily taken as the base for the life table construction according to selected method. Because this macro could be easily repeated for many calendar years or more populations, it is also possible to study the development of the parameter values itself. It can help to describe the mortality development as a whole. In the proposed macro a user can easily define own conditions (minimal and maximal ages used for the estimation, to select calendar years where the estimation process should be done, or select the function of mortality smoothing, etc.).

This chapter contains a short description of all the methods used in the proposed macro, a short description of the whole macro, and possibilities of its usage.

5.2 Basic description of the macro

Only parametric functions were implemented to the macro and so it could be possibly used in any of the following analyses. The first basic comparison of the most important functions, or rather mortality laws, was published in the Czech demographic journal *Demografie* (Burcin *et al.*, 2010). The method of estimation of the unknown parameters could be taken as an independent topic within the basic analysis of the mortality laws. In the proposed macro the weighted non-linear least squares method was used (will be described later). In some commonly used software, like Microsoft Excel or other types of spreadsheets, there the estimation of the unknown parameters using the mentioned method would be

very complicated and almost impossible. Moreover, usage of such software for the estimation for more calendar years or for more populations is at least inefficient. That is why the statistical software (SAS) was chosen for the solution of this task. So that the usage of the macro was possible also for an inexperienced user of SAS, it will be described in detail within this chapter.

The main features of the macro program:

1) when the input data file respects the demands on its design, the macro could be submitted for data of many calendar years in one moment;
2) user can choose from several methods of mortality smoothing such as Gompertz, Gompertz-Makeham, Coale-Kisker and other functions (see below);
3) results are exported directly to Excel file where individual sheets contain particular years (populations). Name of the sheet contains not only the particular calendar year but also the abbreviation of the used method;
4) the estimation method minimizing the sum of weighted non-linear squares is implemented to the macro (more detailed description follows later in the text);
5) not only the estimated values of the parameters are exported but also the smoothed values of mortality rates with its confidence intervals, smoothed values of probabilities of dying and many other characteristics and results;
6) user of the macro has the possibility to choose the age-range for the estimation of the parameters and also the maximal theoretically attainable age – up to that age the smoothed values are calculated. Theoretically there are no limits for this age selection but it should be selected reasonably – the age-range should be width enough for the parameter estimation (at least 15 or better 20 years).

5.3 Methodological background

5.3.1 Life table construction at higher ages – usage of the laws of mortality

Because of the variability of the mortality rates at higher ages some methods of smoothing and extrapolation of the mortality rates are traditionally used. Available and reliable empirical data are used for the estimation of the unknown parameters in the selected function. Through this function the values of mortality rates (hazard function μ_x) are extrapolated also for those ages where the reliable empirical data are not accessible or do not exist at all.

There could be many of those models (or laws of mortality) presented and used in the calculation. Some of the most important ones are presented in the article of Burcin *et al.* (2010). The practice of the Czech Statistical Office is to use the Gompertz-Makeham function (Czech Statistical Office, 2009b) as was mentioned in the previous chapter. The same method is used also by some other national Statistical Offices, like in Slovakia or Estonia (European Commission, 2003), because the Gompertz-Makeham method is one of the best known one.

When the Gompertz-Makeham (or any other parametric method of smoothing and extrapolation) is used, then there arose the necessity of the estimation of the unknown parameters. Only reliable empirical data should be used in the calculation and in the same time the extrapolated values have to fit well the empirical values for the highest ages.

Several methods of the estimation could be used. One of the simplest ones usable for the estimation of the Gompertz-Makeham parameters is described in Pavlík, *et al.* (1986). It is the estimation of the three unknown parameters from the three empirical values of the age-specific mortality rates for three selected ages. The procedure is very easy but still there is the risk that one or more of these three selected values of mortality rates will reach extreme values in some way and the result could be biased due to that (as was already mentioned above, see Chapter 4).

Application of some more sophisticated methods could lead to more accurate results of the estimation procedure. One the basic methods frequently used in similar kind of needs is the method which minimizes the sum of differences between the empirical and estimated values (sum of the least squares; used for example in Burcin, *et al.* 2010). It could be more complicated to use such a method without some specialized statistical software.

Slightly compromise procedure is the King-Hardy method of estimation of the parameters. Briefly it could be described as the estimation of the three unknown parameters on the basis of the empirical mortality rates from three age-intervals (more details are in the Chapter 4 and in Fiala, 2005; Pecka, 1989). This method is used also for the construction of the Czech official life tables (Hartmannová, Fesenko, 1973; Czech Statistical Office, 2009b). Also this method has important disadvantages (see Chapter 4 or Pecka, 1989; Burcin, Hulíková Tesárková, 2011). When any compromises need not to be done and a demographer could use some suitable statistical software, not only the minimization of the least squares could be used but even more sophisticated and complicated methods.

In the proposed macro the non-linear regression is implemented where the weighted sum of the generalized (non-linear) least squares is used. The successful usage of the weighted sum of squares could be found also in the article of Wilmoth (1995).

5.3.2 Weighted generalized least squares method for the parameter estimation

It is reasonable to expect a changing variability of mortality rates with age (heteroskedasticity in data). In accordance to that, specifically constructed weights were used in the computational procedure in the introduced macro. The weights are taken as being equal to the reciprocal values of variance of the mortality rates with age.

The weights were constructed in accordance to the assumption that when the number of deaths is considered to be binomially distributed, then also the mortality rates are binomially distributed (or "relatively binomially distributed" as it was called by Gerylová and Holčík, 1988, p. 69). The variance of such distribution could be written as (*ibid.*):

$$\frac{\pi*(1-\pi)}{n},$$

where π symbolize the relative variable (in this case the mortality rate or the force of mortality) and n is the theoretical number of events (the theoretical number of deaths during the time interval is equivalent to the total number of people living during the same time interval). So we can rewrite the variance as:

$$\frac{m_x*(1-\mathrm{m_x})}{P_x},$$

where m_x is the mortality rate at age *x* and P_x is the population living at completed age *x* in the middle of the studied time interval. The reciprocal value of the variance at age *x* will be taken as the weight at that age in the estimation procedure. Because the method of estimation (generalized/nonlinear weighted least squares) is an iterative method, in each step (iteration) the weights are recalculated with the actual estimate of m_x (based on the actual estimate of the parameters). The same formula for the weights is introduced also by Koschin (1981) or Fiala (2005).

The Gauss-Newton method of estimation was used in the introduced macro. The Gauss-Newton iterative method "regress the residuals onto the partial derivatives of the model with respect to the parameters until the estimates converge" (SAS Institute Inc., 2009, p. 2926). The criterion of convergence has to be specified in the estimation procedure. In the macro attached to this work, the maximum change of parameter estimates as the convergence criterion was used. The iterations are said to have converged for `CONVERGEPARM = c` if

$$\max_j \left(\frac{\left| \beta_j^{(i-1)} - \beta_j^{(i)} \right|}{\left| \beta_j^{(i)} \right|} \right) < c,$$

where $\beta_j^{(i)}$ is the value of the *j*-th parameter after the *i*-th iteration.

When the parameters are estimated, values of the hazard function could be calculated easily only by substituting the parameter values into the equation of the selected model of extrapolation (below). For this purpose we have to describe the considered relation between the mortality rate (m_x) and the force of mortality (intensity of mortality μ_x). From the definition of those both terms it could be easily derived, that the intensity of mortality in the centre of the one-year age interval could be taken as equal to the mortality rate, and vice versa. Therefore, it is supposed that for all *x*, except the age zero, it holds (Thatcher, 1999):

$$\mu_{x+1/2} \approx m_x \, .$$

The probability of dying derived from the smoothed values of mortality rates were calculated as they usually are in the life table:

$$q_x = 1 - e^{-m_x} \, ,$$

Other life table functions could be then calculated easily from the column of probabilities of dying.

5.3.3 Description of the mortality laws used within this work

One of the most important advantages of the introduced macro is the possibility to use one or more different laws of mortality which are used as the parametric functions of mortality smoothing and extrapolation. Traditionally the exponentially or logistically increasing functions are used for the life

table construction. Both these methods are represented in the group of models which are incorporated into the introduced macro.

The exponentially increasing models are represented by the Gompertz and Gompertz-Makeham formulas. The logistically increasing functions are represented by two formulas labeled as "Kannisto" and "Thatcher". The last two functions are slightly different – it is the Coale-Kisker model modified by Wilmoth (1995) to the quadratic function, and the modified Gompertz-Makeham function (proposed by Koschin *et al.*, 1998). The more detailed description of all the models could be found below:

Gompertz and Gompertz-Makeham functions

Too much description of those two (in demography probably the best known) mortality laws is not needed. The Gompertz function (Gompertz, 1825) is used in the form:

$$m(x) \cong \mu\left(x + \frac{1}{2}\right) = a * b^{(x+1/2)},$$

where *a* and *b* are the two unknown parameters which have to be estimated, *x* stands as usually for age. The Gompertz-Makeham function (Makeham, 1860) differs only by the usage of the constant *a* representing that part of the total mortality which does not change with age:

$$m(x) \cong \mu\left(x + \frac{1}{2}\right) = a + b * c^{(x+1/2)}.$$

As it was said already, the intensity of mortality (mortality rates) increases exponentially with age in both the mentioned functions. Because the estimated values of mortality rates approach the infinity with increasing age and because of the relationship

$$p_x = e^{-m_x},$$

the probability of survival approaches limitedly zero and the probability of dying approaches one for the highest ages in the Gompertz and also in the Gompertz-Makeham function if the intensity of mortality increase limitedly to infinity.

Modified Gompertz-Makeham function

The traditional Gompertz-Makeham function is based on the assumption that the rate of increase of mortality is constant with age and it models the age-related mortality by 3 parameters. Nevertheless, the empirical data show that this is not valid and at the highest ages the rate of increase of the mortality rate is very likely to slow down, so it is not constant, but rather a decreasing function. The model formulated on the basis of this mentioned assumption would contain one parameter more in comparison to the Gompertz-Makeham function, i.e. 4 parameters. The new parameter γ represents the mentioned fall in the rate of increase of mortality with age. This modified Gompertz-Makeham function could be in a form expressed as (Koschin *et al.*, 1998; Koschin, 1999):

$$\mu(x) = a + b * c^{x_0 + \frac{1}{\gamma} * ln[\gamma * (x - x_0) + 1]},$$

where $x > x_0$; a, b and c are the parameters of the traditional Gompertz-Makeham function, and $\mu(x)$ is the intensity of mortality at age x.

The age where the function starts to be applied, is denoted as x_0 and it is recommended to be chosen around the age of 85. Therefore, it is a model specifically focused only on the highest age groups. Ages entering into the calculation of the parameter estimates thus start at the age of around 85 years and end at ages where the empirical values of mortality rates begin to fluctuate significantly, i.e. slightly over 90 years (around ages 93 to 95 years). The smoothed values of the modified function resemble the course of the classic Gompertz-Makeham function (on which it is linked smoothly thanks to its construction), however, at the highest ages it does not grow so rapidly, what reflects the latest knowledge about the development of the oldest-old mortality (Koschin *et al.*, 1998; Koschin, 1999).

Kannisto and Thatcher functions

Just as the two previously mentioned basic functions (Gompertz and Gompertz-Makeham), differ the Kannisto and Thatcher functions only by the constant used in one of them.

The Thatcher function was proposed by Thatcher in 1999 (Thatcher, 1999), in the macro it is used in the form

$$m(x) \cong \mu\left(x + \frac{1}{2}\right) = \frac{a * e^{b*(x+\frac{1}{2})}}{1 + a * e^{b*(x+\frac{1}{2})}} + c\ ,$$

where a, b, c are the unknown parameters. The Kannisto formula does not contain the constant c again representing the component of mortality which is independent on age:

$$m(x) \cong \mu\left(x + \frac{1}{2}\right) = \frac{a * e^{b*(x+\frac{1}{2})}}{1 + a * e^{b*(x+\frac{1}{2})}}\ .$$

Both these functions are approaching the value of one with increasing age. As a consequence of that the probability of dying tends limitedly to $1 - e^{-1} = 0.632$.

Coale-Kisker

This model was originally proposed as a relational one – the mortality rate at each age was measured in relation to the mortality rate at the age which was taken as the initial one in the model – usually 80 or 85. In the macro, there the model is used in its quadratic form derived by Wilmoth (1995):

$$m(x) = e^{a*x^2 + b*x + c}\ .$$

5.4 Technical remarks – construction of the macro

The introduced macro is constructed from seven particular macros, each macro solves one of the 6 implemented functions plus one macro serves for running the whole process – this main "outer" macro reads the "setup row" where the demanded features of the calculation are defined by the user and automatically selects the proper "inner" macro solving the particular parameter estimation of the selected function.

Each "inner" macro starts by the import of relevant data – for the whole calculation only the exposure time (number of people living in the middle of the studied time period / year) and empirical values of death rates are needed. These values have to be available for at least one year for individual ages. The unified design of the input data is needed so that the correct data could be extracted from the data set within the macro. The input data set will be described later below.

The core of the macro is the estimation of the parameters. The procedure NLIN is used for this purpose. The general structure of the procedure used in the macro could be illustrated as show below:

```
PROC NLIN DATA=input_data_set MAXITER=&maxiter OUTEST=outest SAVE
CONVERGEPARM=1e-12;
BOUNDS a>&bounda;
PARMS a=&initial_value_of_a b=&initial_value_of_b;

IF
      (&minimal_age_of_estimation-1<x<&maximal_age_of_estimation+1)
THEN
      MODEL equation_of_the_model;

      IF (x<&maximal_age_of_estimation + 1) then
            _WEIGHT_ = description_of_the_weights_calculation;
      ELSE
            _WEIGHT_=0;

OUTPUT OUT=output_data_set PREDICTED=mxp PARMS=a b L95=lower L95M=lowerM
U95=upper U95M=upperM WEIGHT=w;

RUN;
```

The first row starts the NLIN procedure. Then the input data set is defined by the option DATA. The next option, MAXITER, defines the maximal number of iterations. In our macro the value of this parameter was selected as to be equal to 10,000. In most of the cases this value was not reached at all, usually only a few iterations are needed. In the macro the maximal number of iterations was stated as a macro variable (starts by the sign "&") so it is easy to change its value. The command OUTEST defines the data set for the estimates of parameters, in our macro the name of this data set is "Outest". The command SAVE is needed so that the final estimates of the parameters are saved to the OUTEST data set. The last option of the first row is CONVERGEPARM, its value is compared with the maximal change of the value of estimated parameters in each iteration. When the maximal change of the parameters is higher than this value, the next iteration starts, when the value of the maximal parameter change is lower than the defined value the estimation process is finished.

By the option BOUNDS some defined bounds of parameters could be set when needed. Again the value of the bound is defined as the macro variable so that it could be changed easily. Macro variables are also used for the definition of initial values of the parameters introduced by the command PARMS. The procedure showed out not to be very sensitive to these initial values. When these values are

chosen within a reasonable range (where the real parameter value could be expected) then the results do not depend on the initial values.

The IF statement starts the main part of the procedure – it defines that the computation should be done only for the ages within the defined interval (user of the macro defines the minimal age of the calculation and the maximal age to which the estimated values should be produced). Then, after the statement MODEL, the description of the model is defined. For this particular purpose the model has to be stated like an equation. Then another IF process is started where the weights are defined. For values of ages higher than the defined maximal age (defined by the user) all weights are supposed to be equal to zero. Otherwise, the weights are calculated as described in the code, also the weights have to be defined in the equation form.

On the last row the parameters of the output are defined and labeled in the output data set. After the statement OUT= the output data set is defined. Then the predicted values of the mortality function are produced (labeled as "mxp") and the final estimation of the parameters (PARMS = a b). "L95" and "U95" specify variables that contain the lower/upper bound of an approximate 95 % confidence interval for an individual prediction. This includes the variance of the error as well as the variance of the parameter estimates. "L95M" and "U95M" specify variables that contain the lower/upper bound of an approximate 95 % confidence interval for the expected value (mean). In the described macro also the values of the weights from the last iteration are exported. The statement RUN; finishes the procedure and starts its processing when submitted.

Then the procedure SQL is used for the creation of the output table which is exported from the whole process to a defined Excel file, procedure SGPLOT produces the graphical output as defined.

Each macro for a particular function is finished by the export procedure:

```
PROC EXPORT DATA=&output_data_set_exported_to_excel
            OUTFILE= &adress_of_the_output_file
            DBMS=EXCEL replace;
     SHEET=&b&&function;
RUN;
```

On the first row of the procedure the data set with final results is specified, this data set was created in the SQL procedure and will be exported to Excel. This table is then exported to the defined address by the statement OUTFILE=. On the same row, there the Excel format is stated and the option REPLACE means that when a sheet with an already existing name is to be exported then the original (older) one would be replaced. For better orientation in the results the name of the sheet is defined in the form "Yxxxx", where "xxxx" signifies a calendar year (for example Y1980) – this is created by the calling of macro variable *b* ("&b"). After the 5 characters ("Y" and 4 numbers for the calendar year), there the abbreviation of the function would be added – each function is represented by one letter (the abbreviations are the same as at the beginning of the macro, each user has to select one abbreviation for the desired function.

When the currently (March 2012) latest version of the SAS software was released (SAS 9.3) it was necessary to prepare a slightly modified version of the macro. The latest SAS software in its 64-bit

version uses a slightly different code of the procedure IMPORT and EXPORT. According to that a second macro with the same functions was prepared specifically for SAS 9.3, 64-bit users. This could be found also in the attachment. Except the way of importing and exporting data this second macro is the same as the first one.

5.5 How to use the macro – short manual for the user

The whole description of the usage of the macro below is prepared for the user who is not used to work with SAS software; a more experienced user would not need all the instructions, of course.

5.5.1 Input data file

The macro is designed for a standardized dataset, one such a model dataset is distributed (within this work) with some model data (in an Excel-format). The structure needed for the input data file is not complicated (see Figure 11). It should be an Excel file (xls-format) with two sheets:

1) sheet labeled as "Drates" containing age-specific death (mortality) rates, and
2) sheet labeled as "Exposures" with numbers of the exposed population (survivors to the middle of the year according to age or time exposure calculated in any possible way).

Both the sheets should have the same structure. The first column is expected to be labeled as "x" and contains the values of age ("0", "1", etc.). The last value has to be a number (not interval like "100+"). Other columns have to be labeled by the number of the particular calendar year and contain values of the death rates or numbers of survivors. It is no problem when the input data file contains more sheets – for example it is possible to use a sheet with numbers of deaths and from this sheet to create the sheet "Drates" by calculating the rates. When there is an unexpected structure of the input data sheet the macro will warn the used by an error-statement. It is possible to use the model data sheet. In this example, there is also the sheet called "Deaths" so that the values of death rates could be calculated. The sheet "Deaths" could remain as a part of the input data file, it will not be imported into SAS.

Figure 1: Example of the structure of the input data file

	A	B	C	D	E	F	G	H
1	x	1816	1817	1818	1819	1820	1821	1822
2	0	0,186986	0,181727	0,185714	0,196792	0,180915	0,18184	0,207284
3	1	0,046702	0,054247	0,061039	0,066216	0,05613	0,056687	0,060173
4	2	0,033928	0,038904	0,041661	0,045566	0,0393	0,041396	0,042095
5	3	0,022912	0,02705	0,028556	0,030022	0,026246	0,027983	0,02909
6	4	0,015995	0,018839	0,020295	0,021489	0,018371	0,019456	0,020213
7	5	0,013834	0,015245	0,016523	0,018145	0,015834	0,016255	0,016414
8	6	0,012102	0,012863	0,013993	0,015814	0,014054	0,014179	0,013975
9	7	0,01043	0,010745	0,011766	0,013601	0,012284	0,012121	0,011853
10	8	0,008907	0,008928	0,009821	0,011566	0,010563	0,010189	0,009885
11	9	0,007595	0,007492	0,008198	0,009657	0,008925	0,008444	0,008258
12	10	0,006286	0,006471	0,006964	0,007948	0,007366	0,006897	0,006979

Deaths | Exposures | Drates

5.5.2 Preparation of the macro

First of all, the folder "macro_store_smooth_32" or "macro_store_smooth_64" (for SAS 9.3 in 64-bit version users) distributed within this work should be copied to a home directory of the user. Also the program labeled as "SMOOTH32.sas" or "SMOOTH64.sas" should be saved to the home directory of

the user and an input data file should be prepared. Then the SAS session have to be started. Here it will be described how to use the macro in the SAS software version 9.2 and we use the SAS BASE working space (not Enterprise Guide).

Figure 2: Icon of the SAS software

When the SAS session is started (icon is shown in the Figure 12) we need to open the code of the macro. When the Editor-window is activated, a user has to open the file "SMOOTH32.sas" or "SMOOTH64.sas" from his or her home directory (File – Open Program). The code will be opened in the Editor window and it is commented by many notes and instructions.

The first address which should be defined within the "Macro settings" is the directory where the folder "macro_store_smooth_32" (or "macro_store_smooth_64") was stored before starting the SAS session.

In "Graphical settings" the folder have to be defined where the graphical outputs should be exported. Finally the input and output file should be defined.

The macro is started by some general instructions which should be read at least before the first usage of the macro (Figure 13). Then the user has to define addresses of several files as mentioned above (Figure 14).

Figure 3: General instructions of the macro

```
************************************************************
* READ ME BEFORE THE FIRST PROCESSING                      *
*                                                          *
* Please read all the instructions below first.            *
*                                                          *
* Then fill the information which should be filled and     *
* submit the whole code or separatedly the particular      *
* parts of the code step by step.                          *
*                                                          *
* Please be sure that the input data file respects the     *
* needs for its design. If it does not, the programm       *
* wouldn`t be successful!                                  *
*                                                          *
* Close the output file before submiting the code -        *
* without it the programm wouldn`t be successful!          *
*                                                          *
* It is highly recomended to read the log file after       *
* finishing the process.                                   *
************************************************************
```

Figure 4: First part of the macro – definition of the input and output data file

```
*Macro settings:                                                    ;
*Please fill the address where the file "macro_store_smooth_32" was saved and
submit the two rows (the code could be also submited as a whole);
libname smooth32 'g:\SAS\macro_store_smooth_32\';
options mstored mautosource spool SASMSTORE=smooth32;

*Graphical settings:                                                ;
*Please fill the address where graphical outputs should be stored and
submit the two rows (the code could be also submited as a whole);
ods html path="g:\SAS\pictures_store\";
ods graphics on;

*Please define the adress of the input data file;
%let adress_input="g:\SAS\data_example\FRA_F.xls";
*Please define the adress for the output file;
%let adress_output="g:\SAS\output_store\FRA_F_out.xls";
```

Figure 5: Second part of the macro – the setup row

```
*"SETUP ROW"
In the setup row (below) user has to input the values defining the output:
start = the initial (calendar) year for which the parameters should be estimated
stop = the last (calendar) year for which the parameters should be estimated
minimal = the initial age used for the estimation procedure
omega = the last age for which the smoothed values should be calculated
maximal = the last age used for the estimation procedure
function = the selected function of mortality smoothing
            - user can choose from this options (more detailed description
              of the functions could be found in the original Doctoral Thesis):
        K - Kannisto function
        M - Gompertz-Makeham function
        T - Thatcher function
        C - Coale-Kisker function
        F - modified Gompertz-Makeham function
        G - Gompertz function;
%smoothing (start=2000, stop=2005, minimal=30, omega=100, maximal=90, function=m);
```

The main part of the macro is the "setup row" (Figure 15). Several rows of instructions are written above the setup row, those should be read. The user has to fill the values in the setup row (in the brackets). The "start" value signifies the first calendar year for which the parameters of the model should be estimated. The "stop" value signifies the last calendar year for which the parameters should be estimated.

Three age-values should be defined then. The first one, it is the "minimal" value; it is the minimal age which should be used for the estimation. The "maximal" value signifies the maximal age which should be used for the estimation. By those two values the age interval entering the estimation process is defined. The empirical values of death rates at these ages are used in the estimation procedure. It is important that there are enough ages selected for this interval – at least ca 20 years should be used for the estimation of the parameters. Usually the more ages are in the selected interval, the better. On the other hand, only the ages where we suppose reliable data should be selected to this interval. Moreover, the ages where the selected function could be supposed to be valid should be used for the estimation of its parameters – that means that ages at least 25–30 years should be used because in lower ages the mortality development could be supposed to be different from the implemented mortality functions. For the modified Gompertz-Makeham function even higher ages should be selected, this model was

originally developed for the highest ages (ca above 80 years, as mentioned above in the description of the model). Therefore, when the minimal age for this function would be selected as being equal to ca 80 years the length of the age interval used in the estimation process hardly could be 20 years.

The third value which has to be filled in is labeled as "omega", what is the highest age for which the smoothed values are estimated. When the parameter values are estimated, then the smoothed values of the death rates are calculated using the estimated parameters and the formula of the selected model of smoothing. These smoothed values are calculated up to the age of "omega" and exported in the export data file.

The last value that should be defined is the selection of the function of smoothing and extrapolation. One can choose from six models (mortality laws) defined above – the Kannisto function, Gompertz-Makeham function, Thatcher function, Coale-Kisker function, the modified Gompertz-Makeham function and the traditional Gompertz function. There is an abbreviation defined for each of these functions (one letter – "K", "M", "T", "C", "F", "G"). This letter should be written at the end of the "setup row" – see the Figure15 where the selected function is the Gompertz-Makeham function (`function = m`).

5.5.3 Setup the macro

When all the needed information is filled in the setup row the whole macro can be submitted. It is possible to submit the macro as a whole or to submit particular parts. The first way could be done easily when the Editor window is activated by clicking on the "submit" icon.

Figure 6: The "submit" icon in SAS

When one wants to submit particular parts of the macro one by one, it is necessary to select the particular part of the code (it is important to select the whole rows) and then again click on the "submit" icon (see Figure 16). This way of submitting is more practical for a user who wants to read the log output after each step of the code. The log output informs the user about the submitted process (part of the code). The log output is good to read after each submitting of the macro and it could be found in the "Log" window. After submitting the code, it could take several seconds (or even minutes) to process the code. The output file should not be opened when the process is running and it should be closed before the macro is submitted again.

5.5.4 Results

As a result the Excel file and graphical outputs in the png-format are created. In the output Excel file particular sheets contain results for particular selected calendar years and selected function (see Figure 17). The first 5 characters in the sheet name signify the calendar year ("Y2000" to "Y2005" in our example). The last character signifies the selected function (in our example it is "M" for the above selected Gompertz-Makeham function).

Figure 7: Sheet-names in the output data file from the SAS macro

37	65	0.007043	0.006927	0.004815	0.009039	0.006519	0.007335	0.006903
38	66	0.007639	0.007731	0.005505	0.009957	0.007287	0.008175	0.007701
39	67	0.008353	0.008644	0.006298	0.01099	0.008163	0.009126	0.008607
40	68	0.009383	0.009682	0.007217	0.012146	0.00916	0.010203	0.009635
41	69	0.010684	0.01086	0.008249	0.01347	0.010297	0.011423	0.010801
42	70	0.011604	0.012198	0.009402	0.014993	0.011591	0.012804	0.012123
43	71	0.012905	0.013717	0.010712	0.016722	0.013066	0.014369	0.013623
44	72	0.014511	0.015443	0.012229	0.018657	0.014745	0.016141	0.015324
45	73	0.015773	0.017403	0.013967	0.020839	0.016656	0.01815	0.017253

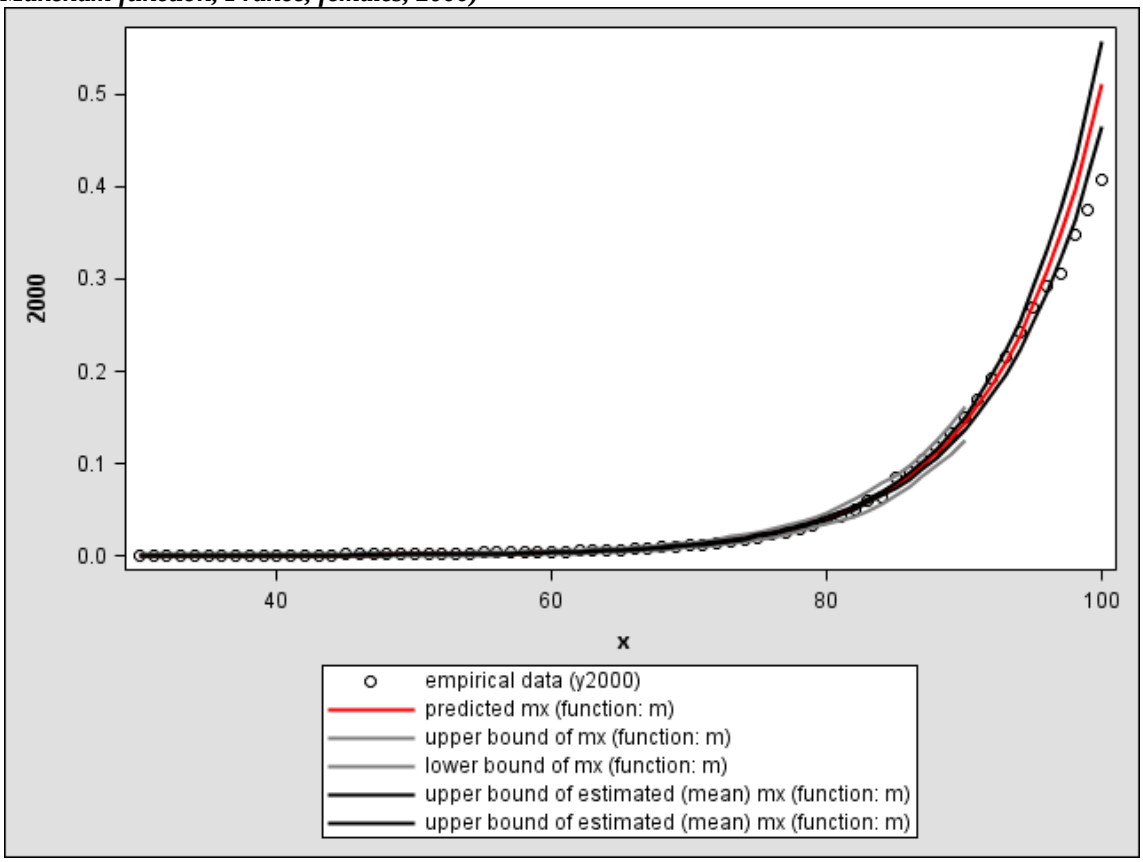
Y2000M | Y2001M | Y2002M | Y2003M | Y2004M | Y2005M

There is a unified structure of all the sheets in the output data file. In the first column, there is the age (labeled as "*x*" in accordance to the input data file). Then there the input data are repeated because there are empirical (imputed) death rates in the column named as "empirical_mx". In the "predicted_mx" column, there are predicted values calculated through the usage of the formula of the selected function and the estimated values of parameters. Columns "lower"/"upper" and "lowerM"/"upperM" contain the lower/upper bounds of an approximate 95 % confidence intervals for the individual prediction ("lower"/"upper") or for the expected value, the mean ("lowerM"/"upperM").

The "predicted_qx" column contains the values of the probability of dying which could be used for the construction of the smoothed life table. These values are constructed on the basis of estimated values of the mortality rates. Using the lower and upper bounds of an approximate 95 % confidence intervals of the expected value (the mean) of mortality rates ("lowerM"/"upperM") the bounds of the estimated values of probability of dying are calculated ("predicted_qx_lowerM"/"predicted_qx_upperM"). For comparison also the calculated empirical probabilities of dying are in the output file ("empirical_qx"). Its calculation is based on the empirical values of mortality rates imputed into the macro. Residuals (the difference between the empirical mortality rates and the estimated ones) are outputted in the column "residual_mx". In the next column, "Px", the original imputed numbers of survivors (or exposure time in general) are repeated. The final values of the weights are in the column named as "weight". The next three columns contain the minimal age of the estimation, the highest age "omega", and the maximal age used in the estimation as were defined by the user in the "setup row" (see above). The last columns contain the final estimations of the parameters. The name of the column corresponds with the name of the parameter in the selected model as described above.

The graphical results are represented by one graph for each calendar year and the selected model. In these graphs there are the empirical values of mortality rates, the estimated values of mortality rates and upper and lower intervals of the predicted values and of the estimated values (the mean) of the mortality rates (see Figure 18). All the graphical outputs are stored in the defined folder and they are in the png-format.

Figure 8: Example of the graphical output from the macro without any further modifications (Gompertz-Makeham function, France, females, 2000)



Note: Output from SAS 9.2 software

Source of data: Author's calculation based on Human Mortality Database (2010)

5.6 Summary

In the previous chapters it was shown that not only the selected function of the method of smoothing and extrapolation can influence significantly the resulted life tables but also the method used for its parameter estimation. In the Chapter 4 it was illustrated for the Gompertz-Makeham function and King-Hardy method of parameter estimation. Within this chapter a more sophisticated method, the non-linear weighted least squares method, was not only introduced but also a tool for its calculation was prepared. In the attachment to this Thesis there a programming code for the SAS software could be found. It could be used for calculation of the parameter estimation for six selected mortality laws. The calculation could be easily repeated for more years or more populations in general. This macro is prepared in two modifications so as also users of the SAS 9.3 64-bit version could use it. This macro will be used also in several following parts of this Thesis.