DeRaS documentation

B. Burcin, K. Hulíková Tesárková, and D. Kománek

with the support of O. Sivková and T. Kučera

Contents

1. Introduction

DeRaS (Death Rates Simulations) is, briefly said, a software application designed for an efficient life table construction under several detailed conditions defined by the user. DeRaS offers several smoothing and extrapolating methods of empirical mortality rates or probabilities of dying at higher ages. In the most recent version of DeRaS, there are six generally recognized mortality models included in the application – the Gompertz and Gompertz-Makeham as two representatives of exponential functions, two logistic functions (Kannisto and Thatcher with constant) and two less frequently used models (Coale-Kisker, Heligman-Pollard). This documentation provides insights into the application installation, its features and theoretical background.

2. Software requirements

The DeRaS application consists of two basic components: console application processing in a command line window and graphical interface (i.e. frontend). The structure was intentionally chosen so as to be able to work in a batch mode via application commands from pre-defined scripts. In addition this choice allows the use of different programming languages and their advantages for various implemented components. Though the most recent version runs only on MS Windows, the application can be easily portable on various operation systems: user interface is programmed in Java, which can be executed on all devices with an installed Java runtime, and console components can be subsequently compiled for the target device. This concept makes it easy to create another frontend, e.g. web-based interface.

Basic software requirements:

- MS Windows XP/Vista/7 (in edition “Home” or higher) in both 32-bit and 64-bit versions, incl. server editions. It is recommended to install the latest Service Pack and updates.
- Java Runtime Environment 6 or 7 (JRE Standard Edition) – required only by graphical interface; JRE is preinstalled on computers with MS Windows, because it is commonly used by other applications there. In case that JRE has not been installed yet, it can be downloaded free of charge from http://java.oracle.com. It is recommended to use JRE's latest version.

3. Installation

Single-user computer

- Run setup (setup_deras.exe), and then setup wizard will guide you through the following steps:

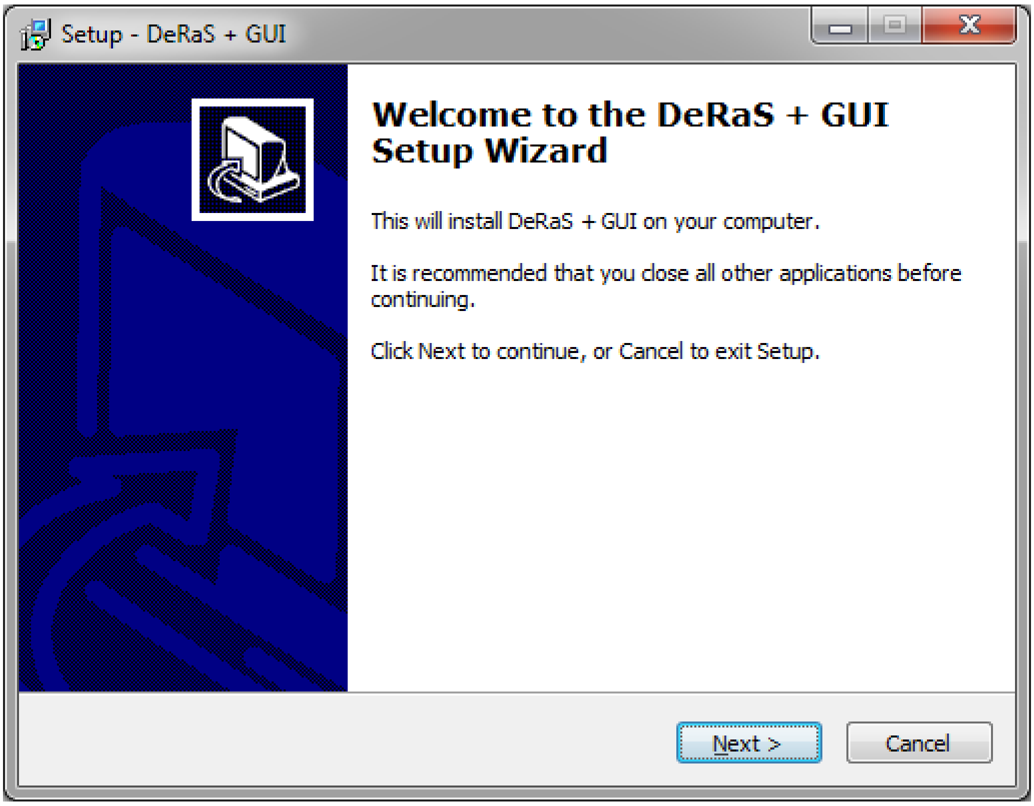

Fig. 1 – Introductory screen of setup wizard

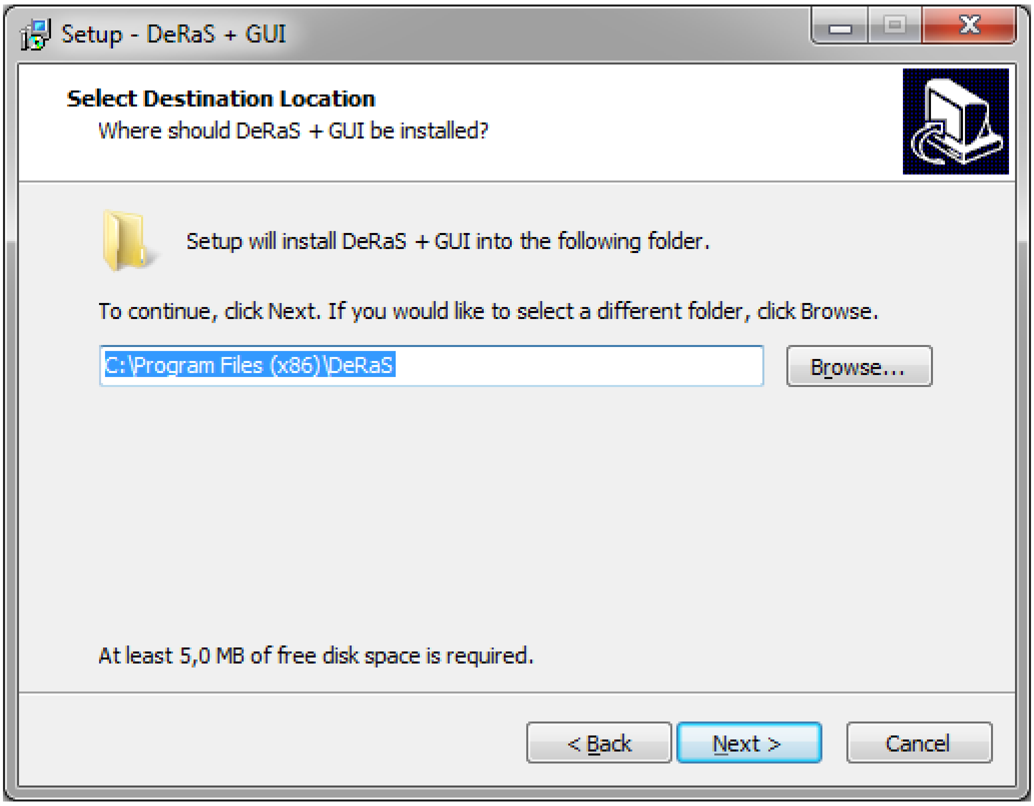

Fig. 2 – Program's installation folder (may differ according to MS Windows version)

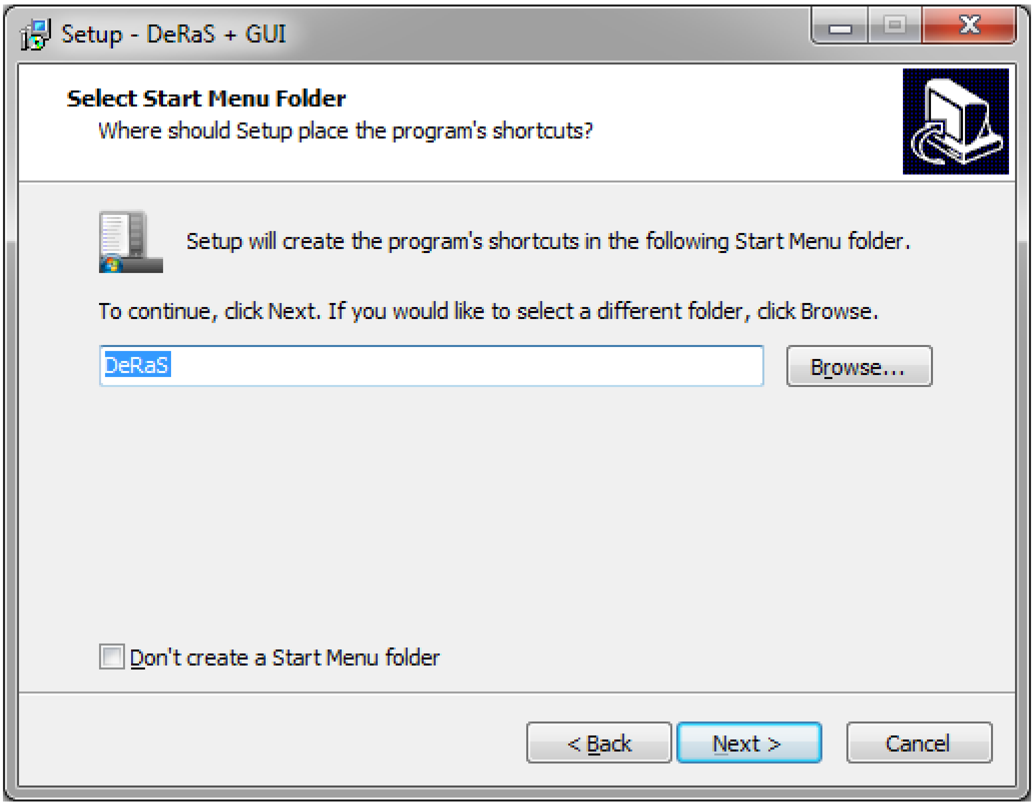

Fig. 3 – Start Menu Folder

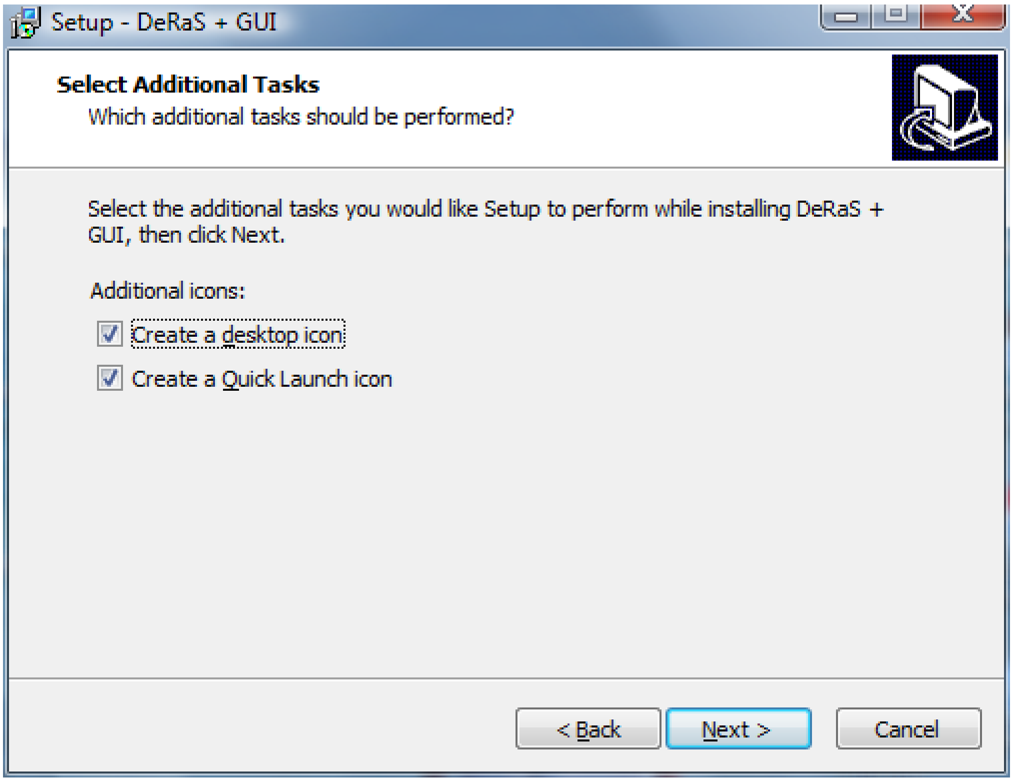

Fig. 4 – Adding Quick Launch and desktop icons

- The installer always uses *Program Files* folder as the default installation directory (see Figure 2). If the setup is not run with administrative rights, it is up to user to choose a different folder with sufficient access rights (e.g. user's profile folder).

- Request a license file: please send us (deras@natur.cuni.cz) your “request.dat” file from your installation folder. See “License” chapter below for more details.

Multi-user system

- Same installation as for single-user computer.
- Single-user installation creates an application’s configuration file in the installation folder. If the application is run under different user’s account, the configuration file is automatically duplicated for this user.

Network installation with Active Directory integration

- The application must be installed on each computer in the domain. Installation process is the same as for single-user computer. The network administrator is advised to create icons and a Start Menu folder into users’ profiles: this can be achieved by logon scripts or moving icons/Start Menu folder into “All Users” profile.

4. License

DeRaS, as a commercial product, is subject to license conditions, as follows:

- non-transferable to other machines, once installed
- multiple users can only work with the licensed application locally (not via network shared folder). Remote Desktop (RDP) usage or desktop virtualization (VDI) are not restricted.
- upon installation (or in extraordinary cases) “request.dat” file is generated in the installation folder. This file must be verified on the activation page at DeRaS site (see activation email sent to you after payment for your link to the activation page). Follow instructions on the site to cryptographically sign the request file and obtain "verify.dat" file, obligatorily saved in the installation folder. Then you will be able to run the application with no limitations. If you are doing a reinstall procedure to a different directory, please move the previously obtained file named “verify.dat” to the new DeRaS target directory.
- If the application does not find a valid license file, it informs the user and terminates itself. If the license was lost, or the file is damaged, you can contact your Reseller to obtain a new license file. A new license file is also required for every significant hardware change, or when the application is copied/moved onto a different computer.

Trial or evaluation license provides time expiration feature. Time-unlimited commercial license acts as follows:

- License includes program updates and new program releases. Release dates are completely up to the manufacturer. The user has right to get the list of program changes.
- Significant future changes and new features in the software may require a new license file too. The use of the older (already unsupported versions) with corresponding license files is allowed for unlimited time (effectively the time period the device hardware and software is capable to running this application).
- License is not liable for any other damages whatsoever arising out of the use of or inability to use the software.

- License does not allow modifying or reverse engineering of algorithm responsible for license code handling and verification.

5. Running the application

It can be executed by desktop icon, or from Start Menu folder (both options were created by setup). The main executable one (called “deras.jar”) is stored in the installation folder. Application scripts can be executed from Console window, see Appendix B of this Documentation for input commands.

Graphical interface

The main window can be used to set up input parameters and modify calculation parameters. It can also be used to check for newer program versions.

6. Application description

Input data

An integral part of the application is a standardized input file, which provides a hint how to format input data for the calculations of life tables (see file 1950-2010_Czech_Republic_males.csv). The file is available on the application web deras.natur.cuni.cz in two versions taking into account a delimiter which generally depends on national settings and language version of the underlying operating system at the time and place of creation of the input file. The input file is in CSV format which means it can be easily imported into MS Excel, Quattro Pro, Open Office Calc, Libre Office Calc and other spreadsheet applications. Users may duplicate the original file and insert own data. When renaming the file, it is recommended to insist on the same file naming convention. Considering the example of the standardized input file, the first part of the name “1950_2010” denotes period, for which data are available, the part “Czech Republic” corresponds with the territory for which data were collected and finally, the last part “male” specifies gender of the given population. If input data are in different format (e.g. in formats listed above), it is simple to transform a given format into the CSV format, e.g. data in MS Excel are transformed via command “Save as-CSV”.

As mentioned above, input data have standardized structure in DeRaS. Therefore violation of the data structure may cause problems in processing, resulting in accurate outputs. Following rules have to be observed:

- **Metadata section:** the first rows contain required metadata. In the first column, the name of the item is presented (Version, Country, Region, Sex), in the second column corresponding value is inscribed. Taking into account the item Sex, user has to insert one of the compulsory notations (F stands for females, M stands for males, and T stands for both sexes). It is important to distinguish the gender of a given population (i.e. F, M or T), because some of applied methods utilize different approaches according to sex (see Figure 5 or file 1950-2010_Czech_Republic_males.csv). The section is closed by one empty row.

- **D_x section (deaths by age)** – the variable required name is located in the upper left corner (D_x). Further columns are named according to the year of data. The name consists of four digit number and columns are in ascending order from left to right. It does not matter which time period is involved as long as the time covered is a continuous series. For independent time periods, new separate data files must be created. Names of rows correspond to the completed ages, which are in increasing order. Ages form the unabridged series from the age of 0 to 110. Thus the matrix of two dimensions is created: it contains the number of deaths for a given completed age and given year. Please bear in mind that data have to be inserted by age unit and not by age intervals. When input data are available, e.g. only up to the age of 100, the rest of rows remain empty (except for the first column where the names are located) (see Figure 5 or file 1950-2010_Czech_Republic_males.csv). In addition, on has to consider that DeRaS demands data at least up to the age of 80. This section is closed by one empty row.

- **P_x section (mid-year population by age)** – the structure of this section corresponds with the previous Dx section. The only difference is in the name of the variable Px located in upper left corner and in the values in the matrix. Otherwise, the name of rows and columns are the same (see Figure 5 or file 1950-2010_Czech_Republic_males.csv). The section is closed by one empty row again.

- **B_v section (number of live births)** – the variable required is located (Bv) in the upper left corner. The name of columns corresponds with the previous two sections, i.e. the year. The data consist only of one row, while the first field has to remain empty (see Figure 5 or file 1950-2010_Czech_Republic_males.csv).

Version	1.0
Country	Czech Republic
Region	
Sex	M

Dx	1950	1951	1952	1953	1954	1955	1956	:	2003	2004	2005	2006	2007	2008	2009	2010
0	6890	6029	4703	3361	2943	2724	2381	:	207	210	211	214	218	200	185	172
1	451	480	360	332	268	201	199	:	19	15	15	13	15	14	15	16
2	241	216	175	177	154	124	110	:	10	6	10	8	16	6	16	15
3	167	155	116	133	114	109	87	:	11	9	6	12	11	13	5	10
4	131	120	91	98	73	82	62	:	9	11	3	9	7	6	5	10
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
96	19	12	19	18	16	23	29	:	104	104	88	104	101	107	155	124
97	3	12	14	10	8	12	8	:	64	50	69	74	79	89	90	85
98	11	4	9	3	5	5	3	:	35	30	46	37	44	53	46	53
99	5	4	8	5	6	4	9	:	12	18	24	17	24	26	33	42
100	0	0	0	1	0	0	0	:	21	17	32	33	22	45	52	48
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
109								:								
110								:								

Px	1950	1951	1952	1953	1954	1955	1956	:	2003	2004	2005	2006	2007	2008	2009	2010
0	89136	90866	90399	88054	85367	83853	82432	:	47909	49109	51384	53555	56392	59947	60990	60443
1	91020	88586	89709	89477	87221	84844	83453	:	47285	47961	49255	51429	53591	56684	60123	61142
2	95503	90548	88937	90077	89261	87135	84805	:	46555	47353	48110	49331	51530	53946	56887	60277
3	94818	95436	90610	88962	89959	89251	87128	:	45656	46546	47437	48111	49354	51813	54070	56946
4	84397	94735	95554	90684	88838	89914	89222	:	45550	45711	46721	47521	48177	49660	51989	54176
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
96	33	37	38	27	10	5	82	:	230	254	278	304	328	360	375	380
97	16	21	24	28	20	5	3	:	98	112	137	161	178	195	209	214
98	8	6	10	17	12	4	3	:	78	95	103	129	148	159	171	176
99	4	4	4	7	9	3	2	:	21	28	32	47	57	65	81	94
100	0	2	4	2	4	4	2	:	47	64	81	89	109	137	75	197
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
109								:								
110								:								

Bv	1950	1951	1952	1953	1954	1955	1956	:	2003	2004	2005	2006	2007	2008	2009	2010
	97490	95778	93047	88949	86615	85580	83370	:	48131	50262	52453	54612	58475	61326	60368	60220

Fig. 5 – Example of input data structure

Graphical interface

Graphical interface (see Figure 6) enables to load input data, set parameters for life table calculations, specify text and graphical outputs, execute calculations, and to check availability of DeRaS updates.

Computation parameters

IO Parameters

Input File – selects input data file which is structured in accordance with the rules described in the section "Input data" of this Documentation. Moreover one can select a text file, which contains list of paths, 1 per line, with input data files (it is recommended to use ".flist" file extension), i.e. application can load several input data files at the same time (e.g. for males and females: 1950_2010_Czech_Republic_males.csv and 1950_2010_Czech_Republic_females.csv (see Figure 7), or for several territories).

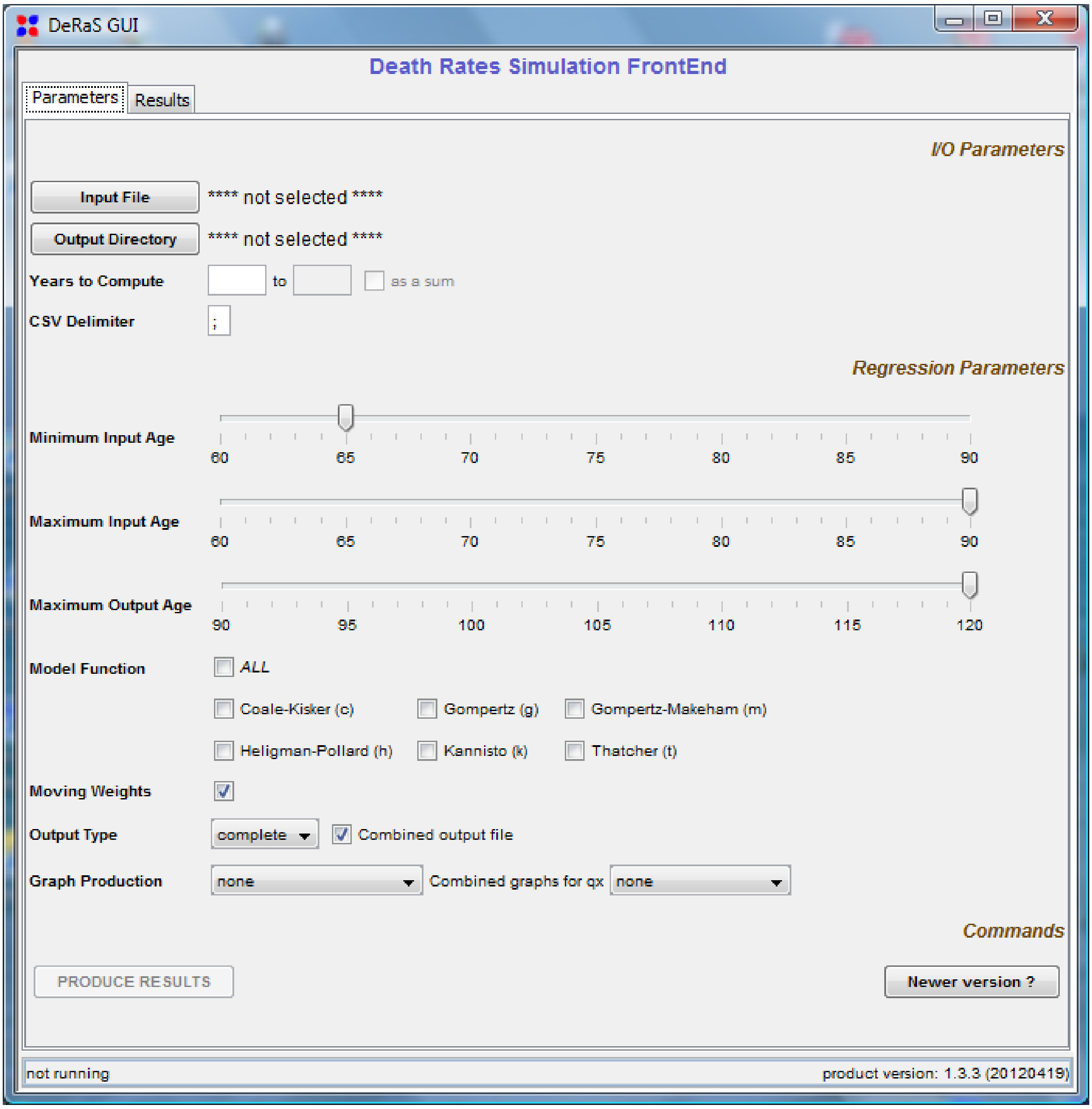

Fig. 6 – Input computation parameters in graphical interface

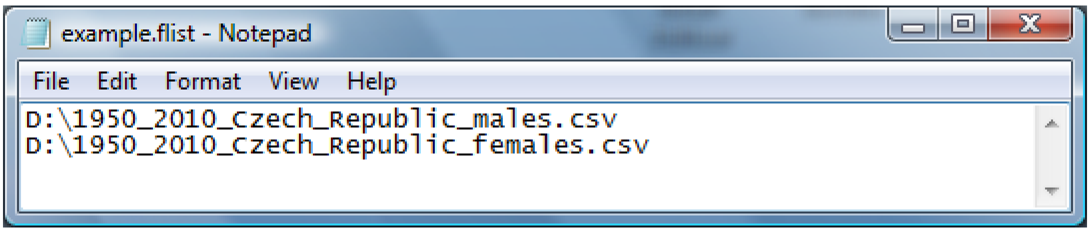

Fig. 7 – Example of text file for loading several input files in one step

Output Directory – sets output path for calculations results. An existing folder to which outputs, both text and graphical will be saved, must be defined. Use of an empty folder is recommended otherwise existing files can be overwritten.

Years to Compute – specifies calendar year or time period for which life tables will be calculated. The selected year or time period has to correspond with input data. In both cases, life tables for a given year or years will be calculated. In addition, if time period is specified, than a parameter "as a sum" can be applied, calculating a table for the whole period. Therefore, the output will not involve separate tables for given years but a single table for entire given period.

CSV Delimiter – field separator in CSV format is local-dependent. User may download from the application web (deras.natur.cuni.cz) an example of the input data file with corresponding delimiter either comma or semicolon.

Regression Parameters

Minimum Input Age – the lowest age for input data at regression model

Maximum Input Age – the highest age for input data at regression model

To obtain unbiased parameter estimates, the minimum interval in regression was set to 20 years. Default setting correspond to the interval between the ages 65 and 90. Empirical values of this interval are used to estimate unknown parameters in implemented models/functions (see Appendix A of this Documentation), applied to smooth and extrapolate mortality rates by age within life tables computation processes.

Maximum Output Age – this parameter enables user to define the maximum age limit for life table computation. The maximum is the age of 120 and over.

Model Function – allows user to choose the function for regression and extrapolation.

Moving Weights – if this option is enabled, each step of iteration recalculates the weights to get more accurate parameters of model function (for details see Appendix A of this Documentation).

Output Type – specifies the type of standardized output, either a complete or an abridged life table.

Combined output file – enables to export all text outputs to one file.

Graph Production – allows user to choose a type of figure for the selected function of the life table. If this option is enabled, figures are also generated together with text outputs. The list of graphical files saved in the output folder is presented on the tab "Graphs". Please note that figures are produced one by one only for the completed life table. The future possibility of exporting all figures at one time is under development.

Combined graphs for qx – allows user to generate comparative charts at one time.

Commands

Produce Results – runs the computation: button is active if all required items are correctly filled out.

Newer version– allows users to check if a new version of the application is available. An internet connection is required.

Text outputs

Basic information about processing is available on the tab “Results” (see Figure 8). All important information is saved into output files. Text outputs are structured in accordance with pattern of input files. The format of text output is CSV thus enabling users to import outputs to any application for table processing.

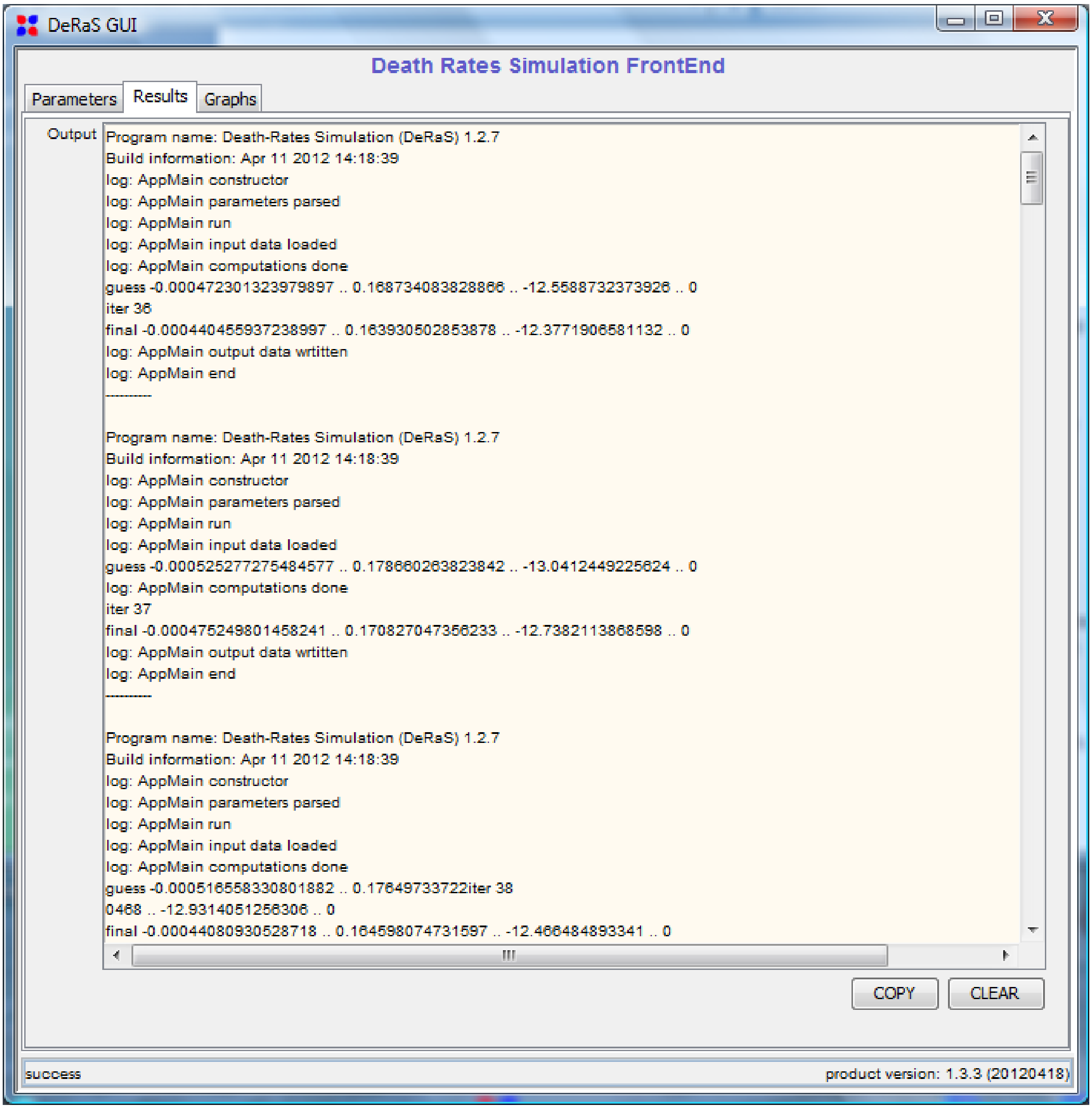

Fig. 8 – Calculation results information

The application provides two types of outputs:

1. Output file, which involves output data only for one calendar year, or period and one model. In other words, if the life table is calculated for several years and several models, than the number of outputs, which is saved into output folder, equals to product of both parameters, i.e. number of years * number of utilized models. It is a default setting.

2. Output file, which involves output data for all selected calendar years and all selected models. This output is friendlier for subsequent comparison of obtained results via, e.g. table calculator. This option can be disabled by “Combined output file” checkbox.

File name

The name of text outputs, so called the first kind of outputs, has a standardized structure as follows:

A_B_C_D_E.extension (.CSV in case of text output and .PNG in case of graphical output), where:

A = name of the input file (in the structure period_territory_sex)

B = type of life table: CLT equals to complete life table, ATL equals to abridged life table

C = calendar year (period) for which life table is constructed

D = abbreviation for selected model: c = Coale-Kisker, g = Gompertz, h = Heligman-Pollard,
k = Kannisto, m = Gompertz-Makeham, t = Thatcher

E = the age interval used in the procedure of the parameter estimation

Examples of output file name for the first kind of text outputs:

1950_2010_Czech_Republic_males_CLT_1951_c_65_90.csv = completed life table, the Czech Republic, 1951, males, model Coale-Kisker, age interval in regression 65-90

1920_2008_France_females_ALT_1951–1955_t_65_85.csv = abridged life table, France, 1951, females, model Thatcher, age interval in regression 65-90

The name for the so called the second kind of text outputs is as follows:

A_COMBINED.csv

where

A = name of the input file (in the structure period_territory_sex)

Examples of output file name for the second kind of text outputs:

1950_2010_Czech_Republic_males_COMBINED.csv

The standardized structure of the name for this kind of outputs is currently being developed.

File content

Text output consists of four sections (see Figure 9).

The first section involves basic information about processed data (type of output table, country, region, sex, year/period, program version and copyright). In batch processing additional information are included into the file name.

The second section contains data organized in columns. The structure depends on the output type (input parameter "Output type"). The required life table with its functions is located in this section (columns *x*, q_x, l_x, d_x, L_x, T_x, e_x – functions description and life table construction is described in the Appendix A of this Documentation). Life table type (complete or abridged) and the scope (the value of the highest age) of the life table correspond with the specification of input parameters "Output type" and "Maximum Output Age" (see paragraphs above).

Complete Life Table	
Country	Czech Republic
Region	
Sex	M
Year / Period	2010
Program Version	1.2.7
Copyright	Copyright (c)2010-2012 by B. Burcin, K. Hulikova-Tesarkova, D. Komanek

x	qx	lx	dx	Lx	Tx	ex
0	0,002856	100000	286	99729	7440703	74,41
1	0,000262	99714	26	99701	7340973	73,62
2	0,000249	99688	25	99676	7241272	72,64
3	0,000176	99663	17	99655	7141596	71,66
4	0,000150	99646	15	99639	7041942	70,67
:	:	:	:	:	:	:
96	0,349832	1391	486	1147	2956	2,13
97	0,371790	904	336	736	1809	2,00
98	0,393299	568	223	456	1073	1,89
99	0,414157	345	143	273	617	1,79
100	0,434189	202	88	158	343	1,70
101	0,453247	114	52	88	185	1,62
102	0,471218	62	29	48	97	1,55
103	0,488023	33	16	25	49	1,49
104	0,503615	17	9	13	24	1,44
105	0,517980	8	4	6	12	1,39
106	0,531126	4	2	3	5	1,35
107	0,543085	2	1	1	2	1,31
108	0,553907	1	0	1	1	1,28
109	0,563651	0	0	0	0	1,26
110+	1,000000	0	0	0	0	1,23

Input data			
x	Dx	Px	Bv
0	172	60443	60220
1	16	61142	
2	15	60277	
3	10	56946	
4	10	54176	
:	:	:	
96	124	380	
97	85	214	
98	53	176	
99	42	94	
100	48	197	
101			
102			
103			
104			
105			
106			
107			
108			
109			
110+			

	Selected model				
Unadjusted	Thatcher, 65-90	Adjusted			
qx	parameters	qx	sdmin	R-squared adj.	Final status
0,002856	9,65734E-07	0	79	0,99703718	SUCCESS
0,000262	0,140001404	0			
0,000249	0,014773936	0			
0,000176		0			
0,000185		0			
:		:			
0,278423		0,349832			
0,327798		0,371790			
0,260023		0,393299			
0,360334		0,414157			
0,216242		0,434189			
		0,453247			
		0,471218			
		0,488023			
		0,503615			
		0,517980			
		0,531126			
		0,543085			
		0,553907			
		0,563651			
		0,572388			

Fig. 9 – Example of output data file

Input data are located in the next section (columns *x*, D_x, P_x, B_v). They can be used for additional analysis.

The last fourth section is devoted to the recapitulation of the parameters estimation of the selected smoothing and extrapolating function – parameters estimated values and basic evaluation of the selected model (coefficient of determination) could be found there.

The "unadjusted qx" column contains the values of the probability of dying computed from original input data.

In the column labeled "parameters", estimated values of parameters included in selected models (two or three values/parameters) are located. In all cases, the upper value is for the parameter a, the second value for the parameter b, and the eventual third one is the value of the parameter c (as defined in this Documentation methodological part, describing incorporated functions).

Column "Adjusted qx" lists values of the probability of dying as estimated according to the smoothed and extrapolated values of age-specific mortality rates based on the used model.

Value of "sdmin" represents the value where the first smoothing (moving average described in this Documentation methodological part) is switched to the second method of estimation (selected parametrical function).

Then the adjusted coefficient of determination is used for the basic evaluation of the quality of fitting of the selected model to empirical data ("R-squared adj.").

Final status – shows the state of processing at the final stage. If the calculation is finished then the report "SUCCESS" is shown. The other states indicate divergence in calculation, i.e. the limit solution was not approached, or other error states.

Graphical outputs

Choosing the graphical output, an independent tab in application's interface is created ("Graphs" – see Figure 10). The list of paths to files with figures in a standard graphical format (png) is shown there. This format (png) is suitable because it is easy to use in other applications, e.g. it can be directly published on web.

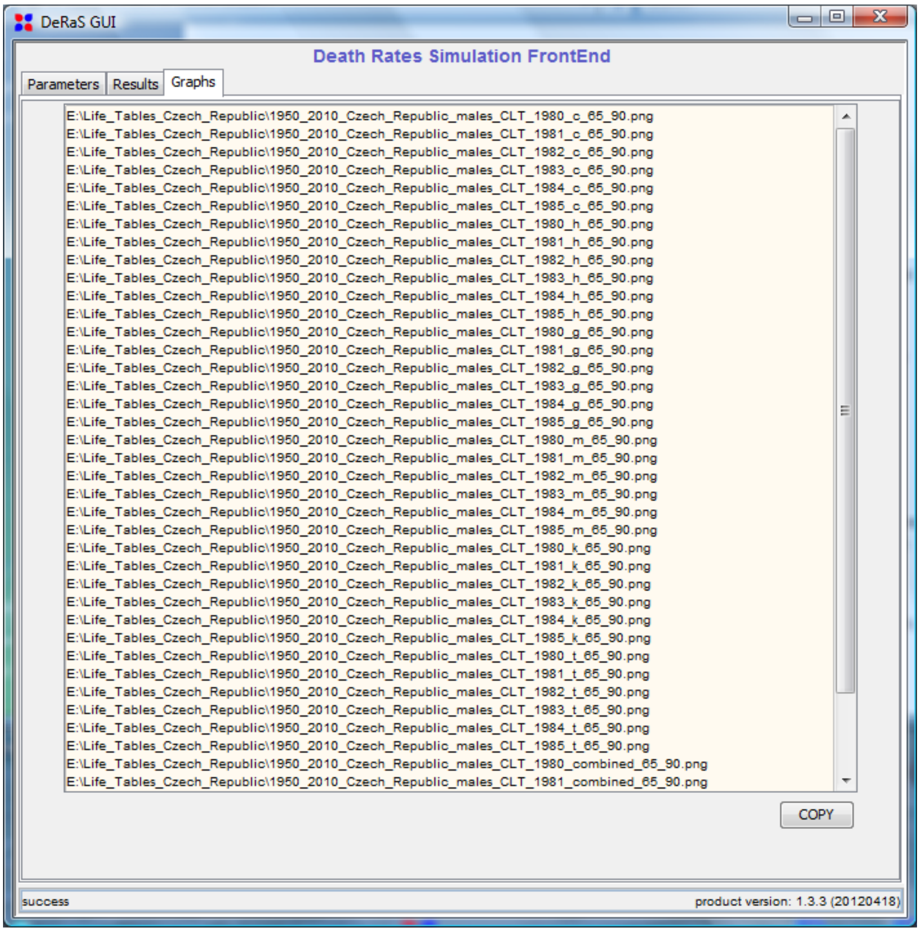

Fig. 10 – List of the generated graphical outputs

Basic graphical outputs serve for four selected table functions description (probability of deaths in comparison with empirical probability of dying, life table survivors, life table deaths, and life expectancy – see Figures 11a and 11b). These graphical outputs cannot be produced based on the abridged life table. Names of figures are structured in the identical way as the text outputs names.

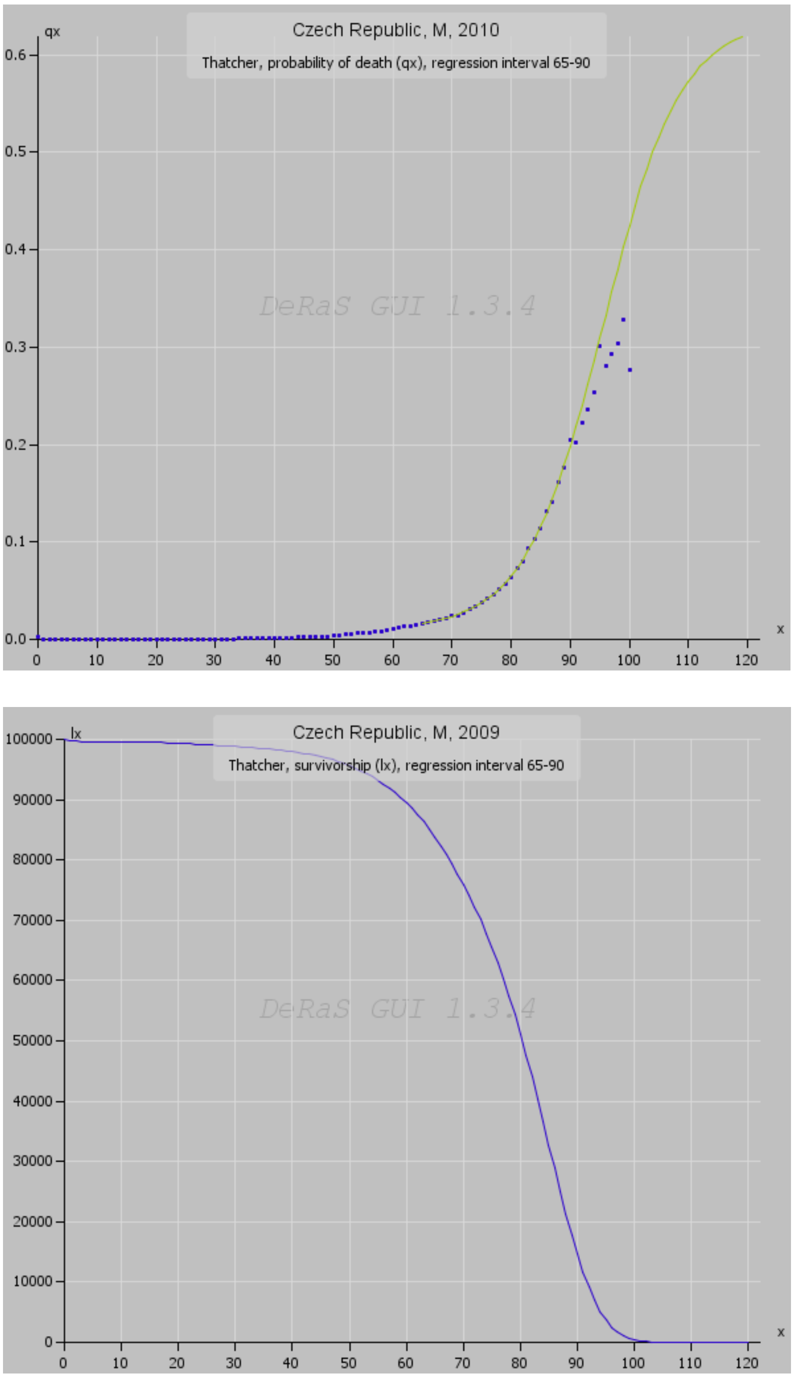

Fig. 11a – Examples of graphical outputs for selected table functions

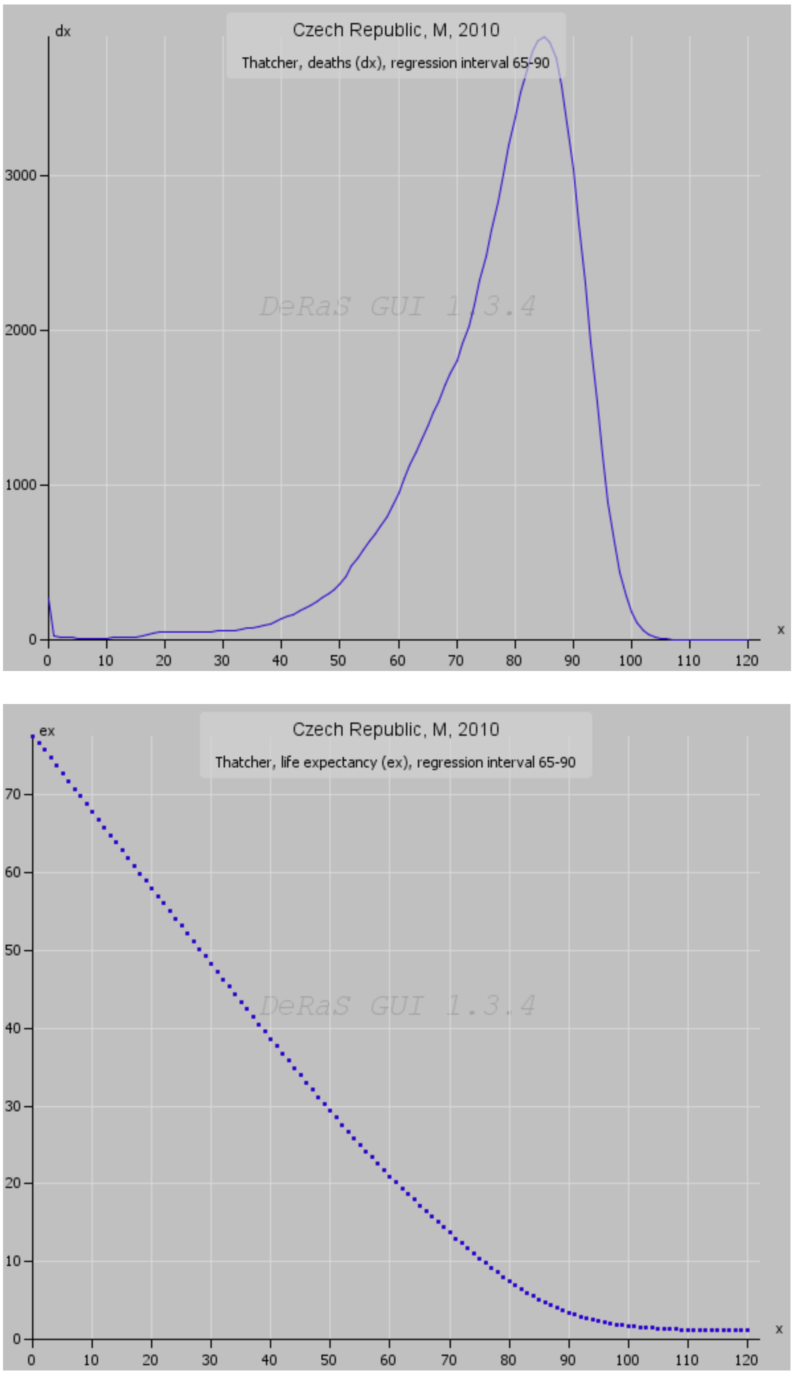

Fig. 11b – Examples of graphical outputs for selected table functions

Basic graphical outputs names are identically structured as text outputs ones.

Advance graphical outputs enable to produce two types of combined figures (see Figure 12 and Figure 13). The first type presents development trends in empirical and smoothed (extrapolated) values of the probability of dying for a given calendar year (period), selected models and sex. The name structure of given graphical output is as follows:

A_B_C_combined_E.png

where

A = name of the input file (in the structure period_territory_sex)

B = type of life table: CLT equals to complete life table, ATL equals to abridged life table

C = calendar year (period) for which life table is constructed

E = regression interval used in parameter estimation processing

Examples of output file name for the first type of the combined graphs

1950_2010_Czech_Republic_males_CLT_2010_combined_65_90.png

The second type of combined figures shows behavior of the probability of dying for empirical and smoothed (extrapolated) values in a maximum period of 11 years. The name structure of given graphical output is as follows:

A_B_combined_D_E.png

Where

A = name of the input file (in the structure period_territory_sex)

B = type of life table: CLT equals to complete life table, ATL equals to abridged life table

D = abbreviation for selected model: c = Coale-Kisker, g = Gompertz, h = Heligman-Pollard,
k = Kannisto, m = Gompertz-Makeham, t = Thatcher

E = regression interval used in the procedure of the parameter estimation

Examples of output file name for the second type of combined graphs

1950_2010_Czech_Republic_males_CLT_combined_c_65_90.png

These graphical comparative outputs are obtained via tab “Combined graphs for q_x“.

These outputs can be generated only from graphic interface. They cannot be generated from console windows.

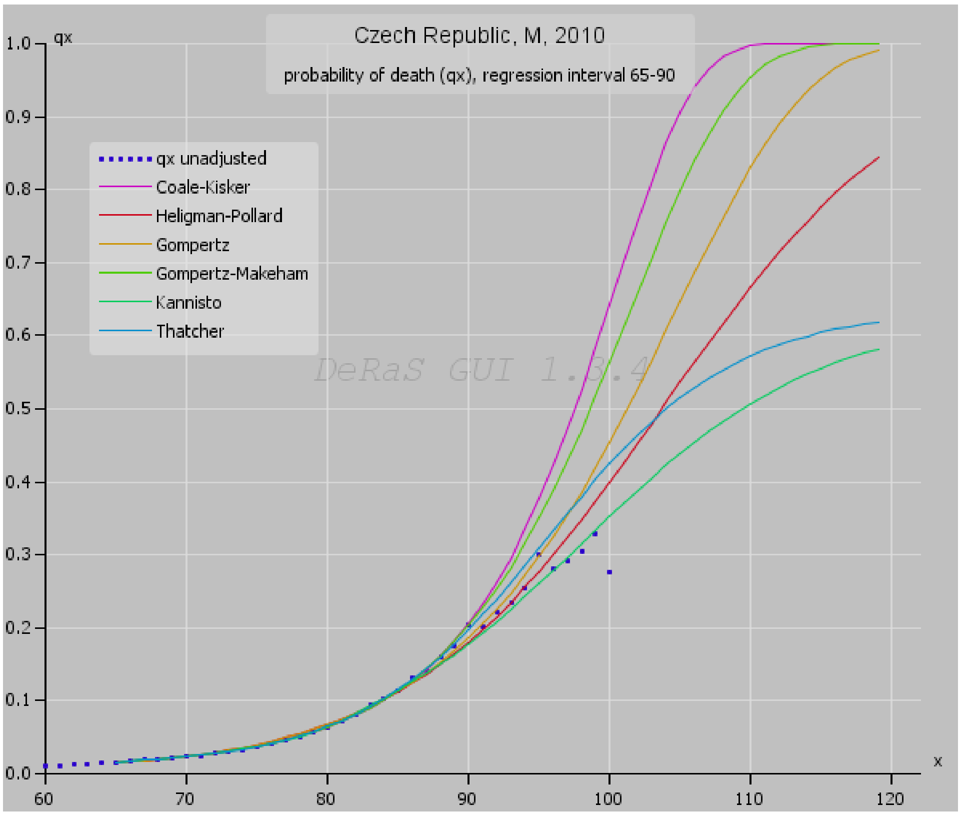

Fig. 12 – Example of the combined chart

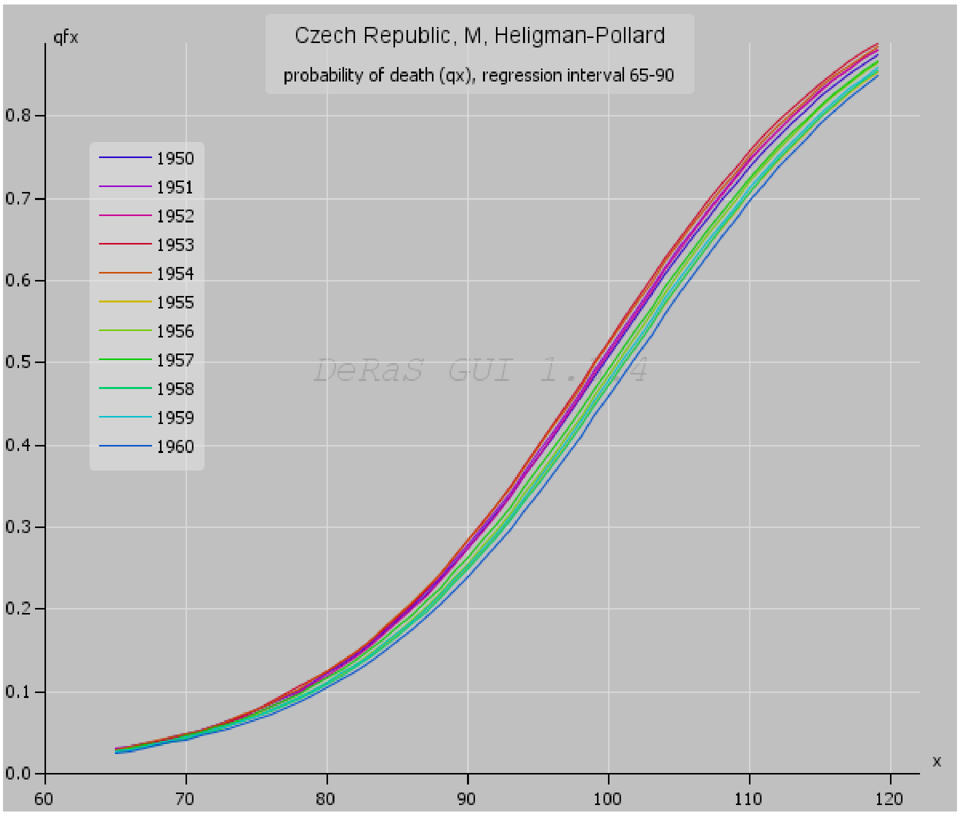

Fig. 13 – Example of the combined chart

7. Appendix A: Methodological background of DeRaS

This part of the Documentation deals with the application theoretical background. Constructions of complete and abridged life tables, as well as mortality laws are described.

Theoretical description of the complete life table construction used in DeRaS

The basics for the methodology used in DeRaS arose from the current (in 2009) way of life table construction used by the Czech Statistical Office. The main difference consists in using different method of parameter estimation. While the Czech Statistical Office uses the King-Hardy method of estimation, parameters of the Gompertz-Makeham formula used in their life tables (*Czech Statistical Office 2009*), in DeRaS the weighted nonlinear least squares are used.

Probability of dying (q_x)

The probability of dying (q_x), the life table basic function, is calculated using so called indirect method stemming from the transformation of the age-specific mortality rates (m_x). It could be written according to the Czech statistical office (2009) as:

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

$$m_x = \frac{Dx}{Px},$$

where D_x denotes the absolute number of deaths by completed age x, P_x presents mid-year population by completed age.,

For the age 0 holds:

$$q_0 = \frac{D_0}{B_v},$$

where D_0 equals to the absolute number of deaths by completed age 0, B_v is the number of lives births in a given year.

From the addition to the value of one, the probability of surviving between the exact age x and x+1 (p_x) is calculated. It can be written in the form:

$$p_x = 1 - q_x.$$

Values of the numbers of deaths and survivors could be taken from the input dataset. Calculated probabilities of dying and/or surviving are considered as the basis for the whole life table. Because of its variability, these functions are traditionally smoothed in some way before the whole table is constructed (*Koschin 1993*). Within the Czech official methodology, the moving average for the ages above 4 years is used for the smoothing (*Hartmannová et al. 1973*; *Czech Statistical Office 2009*):

$$q_x^{smo\_I} = \frac{1}{315} * [105 * q_x + 90 * (q_{x-1} + q_{x+1}) + 45 * (q_{x-2} + q_{x+2}) - 30 * (q_{x-3} + q_{x+3})],$$

where $q_x^{smo\_I}$ represents the smoothed values of the probability of dying.

In the DeRaS calculation, the method of moving average is utilized in the same way as in the Czech official methodology.

In case the probability of dying (q_x) equals to zero in at least two neighboring age groups, than given formula produces negative probabilities of dying. Therefore, in DeRaS an additional approach for probability smoothing is applied:

$$q_x^{smo\_II} = \frac{1}{375} * [105 * q_x + 90 * (q_{x-1} + q_{x+1}) + 45 * (q_{x-2} + q_{x+2})]$$

where $q_x^{smo\_II}$ represents the smoothed values of the probability of dying.

Usage of mortality laws

Because of the variability of the mortality rates at higher ages, some methods of smoothing and extrapolating mortality rates are traditionally used. In this phase, available and reliable empirical data are used to estimate unknown parameters in some selected function. Through this function, values of mortality rates (hazard function μ_x) are extrapolated also for specific ages where reliable empirical data are not available, or do not exist at all.

Numerous such models (or laws of mortality) could be 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 2009*). The same method is also used by a few other national Statistical Offices, because the Gompertz-Makeham method is one of the best known.

When the Gompertz-Makeham (or any other method of smoothing and extrapolating) is used, the necessity of the estimation of the unknown parameters arises. Only reliable empirical data should be used in this estimation though extrapolated values have to substitute well empirical values for the highest ages concurrently.

Several methods of extrapolation could be used. One of the simplest methods 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 a risk that one or more of these three selected values of age-specific mortality rates will reach extreme values in some way thus the result could be biased.

Using additional sophisticated methods could lead to more accurate results of the estimation procedure. One the basic methods frequently used in similar types of needs, is the method which minimizes the sum of differences between 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 any specialized statistical software.

A slightly compromising procedure is the King-Hardy method of parameters estimation. It could be briefly described as the estimation of the three unknown parameters on the basis of the empirical mortality rates from three age-intervals (more details in *Fiala 2005; Pecka 1989*). This method is used also in the construction of Czech official life tables (*Hartmannová et al. 1973*; *Czech Statistical Office 2009*). However this method has important disadvantages (*Pecka 1989; Burcin, Hulíková 2011*). When compromises are deemed needless, and a demographer could use some suitable statistical software, not only minimization of the least squares could be used but even more sophisticated and complicated methods.

Our proposal is to use nonlinear regression where weighted least squares method is used in an iterative manner (output values in one iteration are subject to next round of regression with newly computed weights until the requested precision of estimated parameters is reached). The successful usage of the weighted sum of squares could be found in article of *Wilmoth* (*1995*).

Iteratively reweighted (non-linear) least squares method for parameter estimation

Because the variability of mortality rates changes with age, heteroskedasticity could be expected in data. Consequently specifically constructed weights are used in DeRas computational procedure. Weights are taken as being equal to the reciprocal value of variance of the mortality rates with age.

Variance of age-specific mortality rates could be written according to *Gerylovová and Holčík (1988*) 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 the completed age *x*. In this case the method of estimation is an iterative one. Weights are recalculated in each iteration of the procedure. In addition, the Gauss-Newton method of estimation is applied.

After parameters are estimated, values of the hazard can be calculated easily only by substituting parameter values into the equation of the selected model of extrapolation (see below). It is supposed that for all *x* except the age zero it holds (*Thatcher 1999*):

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

Switch from weighted average to parametric smoothing of mortality rates

When both methods of smoothing are applied, a method of fluent transition from one onto another one has to be used. In DeRaS, the method of minimal difference is applied for detection the age where the switch takes place – this age is characterized by the minimal difference of values smoothed by moving averages from values smoothed by a chosen parametric function.

Construction of other life table functions

In DeRaS, the standard methodology of life table construction is used. After the estimation, smoothing and extrapolating of mortality rates, probabilities of dying and surviving are calculated in the way described above and repeated here below:

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

$$q_0 = \frac{D_0}{B_v},$$

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

From these functions (or from one of them) other life table functions are derived:

lx = number of persons alive at exact age *x,* out of the original number of births

dx = number of deaths between exact ages *x* and *x* + 1

Lx = number of person-years lived in the interval between exact ages *x* and *x* + 1

Tx = total number of person-years lived from exact age *x*

ex = average number of years lived by a person from exact age *x*

are computed easily as (the radix of the table (survival function at exact age zero, l_0) is taken as equal to 100 000):

$$l_{x+1} = l_x * p_x,$$

$$d_x = l_x - l_{x+1} = l_x * q_x,$$

$$L_0 = l_0 - a_0 * \ d_0,$$

where a_0 denotes the ratio of infant deaths up to the age of six months on the total number of infant deaths in given year. This parameter depends on the intensity of infant mortality in given population. Therefore the following approach in accordance with the methodology applied in the Human Mortality Database (2007, p. 38) is utilized.

If $m_0 < 0.107$, than

for males: $a_0 = 0.045 + 2.684 * \ m_0$

and

for females: $a_0 = 0.053 + 2.800 * \ m_0,$

otherwise for m_0 ≥ 0.107,

for males: $a_0 = 0.33$

for females: $a_0 = 0.35$

In case of the total population, the approach had to be modified. Taking into account the fact that input data in DeRaS contain raw data for males and females separately, following relations were derived.

If $m_0 < 0.107$, than

$a_0 = 0.04935 + 2.273308 * m_0$

otherwise for m_0 ≥ 0.107,

$a_0 = 0.34$

The Human Mortality Database (2007, p. 38) applies similar approach to a_0 calculation in case of both sexes.

With the exception of the last age interval, the number of person-years lived in the interval between exact ages *x* and *x* + 1is calculated as follows:

$$L_x = \frac{l_x + l_{x+1}}{2},$$

For the last age group x, which is in DeRaS defined by user, calculation of L_x corresponds to following approach. The extrapolated value of m_x for the open interval is unknown. Thus the life table is first calculated up to the age of 130 years. It is assumed that no one can live longer. Consequently equality of L_{130} and l_{130} holds, which means that the number of person-years lived in the open interval x in the final mortality table, equals to the sum of L_x from the age of 130 to the age x, which is defined by user. One must note that maximum x can be in the application age 120 years. Therefore

$$\infty L_x = \sum_x^{130} L_x.$$

Afterwards it holds:

$$T_x = \sum_x^{\omega} L_x$$

and

$$e_x = \frac{T_x}{l_x},$$

where e_x is life expectancy at age *x* and ω is maximal attainable age assumed in the table.

Theoretical description of abridged life table construction used in DeRaS

DeRaS enables to construct both complete and abridged life tables. The abridged life table is calculated based on the complete life table as follows:

1. The values of the life table function l_x for the age x =0, 1, 5, 10,… correspond to the sum of values for a given age group in the complete life table.

2. In a similar way, the life table function L_x is calculated. Therefore, ${}_nL_x$ in the abridged life table equals to the sum of values L_x for the age interval n in the complete life table.

3. According to the following formula:

$$nq_x = 1 - \frac{l_{x+n}}{l_x},$$

probability of dying at the age interval *n* is calculated. For the open age interval ∞q_x equals one.

4. Remaining life table functions are calculated with respect to above mentioned formulas.

Note: If the maximum output age within parameters' settings does not correspond to the age ending in 0 or 5, e.g. the age of 113 years, then DeRaS selects the nearest lower value ending in 0 or 5 in calculation of the abridged life table, i.e. in given example the age of 110 and over is selected.

Description of mortality laws used in DeRaS

One of the most important advantages of DeRaS is the possibility to use one or more different laws of mortality which are used as the parametric functions of mortality smoothing and extrapolating. Traditionally 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 DeRaS.

Exponentially increasing models are represented by the Gompertz and Gompertz-Makeham formulas. 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 last term of the formula proposed by Heligman and Pollard which is smoothing probabilities of dying instead of mortality rates. A more detailed description of all 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:

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

and the Gompertz-Makeham (*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 stated earlier, the intensity of mortality (mortality rates) increases exponentially with age. Because estimated values of mortality rates approach the infinity with increasing age and because of the relationship

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

the probability of surviving approaches limitedly zero and, the probability of dying approaches one for the highest ages.

Kannisto and Thatcher functions

As is the previous functions, Kannisto and Thatcher functions merely differ by the constant.

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

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

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.

Heligman-Pollard model

This model was originally suggested as a nearly universal model which would be able to describe the mortality development in all ages. In DeRaS, only the last term of this function is used: it is the term which is supposed to describe the development in adult ages (*Caselli 2006, Boleslawski, Tabeau 2001*). In DeRaS it is used in the form:

$$q_x = \frac{b*e^{a*x}}{1+b*e^{a*x}},$$

Thus probabilities of dying have to be calculated from the input data, not mortality rates.

Coale-Kisker model

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

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

Calculation of the adjusted coefficient of determination

The adjusted coefficient of determination is used for basic evaluation of fitting quality of the selected model to empirical data ("R-squared adj.") – see output text files. This measure is computed according to the formula:

$$R^2_{adj.} = 1 - (1 - R^2) * \frac{n-1}{n-k} = 1 + (R^2 - 1) * \frac{n-1}{n-k},$$

where *k* is the number of parameters estimated in the model, *n* is the number of empirical values used for the estimation and R^2 is the standard coefficient of determination. The closer is the value of the $R^2_{adj.}$ to one, the better the selected model fits the empirical data.

8. Appendix B: Use of the console window (without graphical interface)

This application is command-line executable, behaving according to numerous command line parameters. It is suitable for running in scripts/batches. For an intelligent GUI, covering most of the functionalities of this application and offering even more features see other chapters of this documentation.

A. General usage info:
deras.exemodelScenarioweightsScenariomoving_weightsregression_bound_minregression_bound_maxprint_bound_maxyear_startyear_endoutput_typeinput_fileoutput_file [CSV delimiter]

B. Parameters are positional and have the following meaning and variants:
1. modelScenario - shortcut name for the regression modeling function. Possible values are:
  k - Kannisto
  t - Thatcher
  g - Gompertz
  m - Gompertz-Makeham
  c - Coale-Kisker
  h - Heligman-Pollard

2. weightsScenario - shortcut for the type of weights used in regression computations. Possible values are:
  e - equal
  l - pure logistic
  d - demographic

3. moving_weights - weights recalculation mode for regression computations. Possible values are:
true - recalculated after every iteration
false - calculated only at the beginning

4. regression_bound_min - minimum age of the regression input interval
integer value in <60;90>

5. regression_bound_max - maximum age of the regression input interval
integer value in <regression_bound_min+15;90>

6. print_bound_max - maximum age predicted
integer value in <regression_bound_max;120>

7. year_start - beginning year from the life tables
integer value in <1000;2100>

8. year_end - ending year from the life tables
integer value in <year_start+1;2100> - computations are performed for sums over the years interval
year_start - computations are performed for the given single year

9. output_type - form of the resulting table
complete - complete table
abridged - abridged table

10. input_file - full or relative path to the input data tables, on Windows both back- and forward slashes in the paths are supported

11. output_file - full or relative path to the resulting data table, on Windows both back- and forward slashes in the paths are supported

12. CSV_delimiter - character which is the data delimiter in the CSV files (i.e. ";", "," or "|", default: ";")

9. References

BOLESLAWSKI, L., TABEAU, E. 2001. Comparing Theoretical Age Patterns of Mortality Beyond the Age of 80. In Tabeau, E., van den Berg Jeths, A., Heathcote, Ch. (eds.): *Forecasting Mortality in Developed Countries : Insights from a Statistical, Demographic and Epidemiological Perspective*, Springer Netherlands, pp. 127–155. <http://www.springerlink.com/content/vv65332142w06u2h/>.

BURCIN B., TESÁRKOVÁ K., ŠÍDLO L. (2010): Nejpoužívanější metody vyrovnávání a extrapolace křivky úmrtnosti a jejich aplikace na českou populaci. *Demografie*, Vol. 52, No. 2, pp. 77–89.

BURCIN B., HULÍKOVÁ TESÁRKOVÁ K. (2011): Úmrtnost obyvatelstva České a Slovenské republiky, návrh konstrukce nových úmrtnostních tabulek. *Česko a Slovensko ve společném státě a samostatně, podobnosti a odlišnosti – XLI. konference České demografické společnosti.* Příspěvek na konferenci, 26.–27. 5. 2010, Přírodovědecká fakulta Univerzity Palackého v Olomouci, Olomouc.

CASELLI, G. 2006. Mortality Forecasts: Hypotheses and Methods. In G.Caselli, J. Vallin and G. Wunsch (eds.): *Demography: Analysis and Synthesis*, Vol. 3, pp. 177–188.

CZECH STATISTICAL OFFICE. 2009. Úmrtnostní tabulky za ČR, oblasti a kraje 2008 : Metodické poznámky. [Online] 2009. [Cited: 12- 01- 2009] <http://czso.cz/csu/2009edicniplan.nsf/t/BD00408CCA/$File/400209m.pdf>.

FIALA, T. 2005. *Výpočty aktuárské demografie v tabulkovém procesoru*. Praha : Oeconomica.

GERYLOVOVÁ, A., HOLČÍK, J. 1988. Základy statistické metodologie. Brno: Univerzita J. E. Purkyně, 128 p.

GOMPERTZ, B. 1825. On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies. *Philosophical Transactions of the Royal Society of London*, 115, pp. 513–585.

KOSCHIN, F. 1993. *Aktuárská demografie : (úmrtnost a životní pojištění)*. Praha : Vysoká škola ekonomická.

HARTMANNOVÁ, H., FESENKO, P. 1973. Úmrtnostní tabulky 1970. *Demografie*, No. 3, pp. 235–241.

HUMAN MORTALITY DATABASE. 2007. Methods Protocol for the Human Mortality Database (Version 5). [Online]. [Cited: 20-04-2012] <http://www.mortality.org/Public/Docs/MethodsProtocol.pdf >.

Makeham, W.M. 1860. On the law of mortality and the construction of annuity tables. *Journal of the Institute of Actuaries*, Vol. 8, pp. 301–310.

PAVLÍK, Z., RYCHTAŘÍKOVÁ, J., ŠUBRTOVÁ, A. 1986. *Základy demografie*. Praha : Academia, pp. 736.

PECKA, J. 1989. Příspěvek k problematice výpočtu Československých úmrtnostních tabulek. *Demografie*. Vol. 3, pp. 229–238.

THATCHER, A.R. 1999. The long-term pattern of adult mortality and the highest attained age. *Journal of the Royal Statistical Society (Series A)*, 162 (Part 1), pp. 5–43 .

WILMOTH, J.R. 1995. Are Mortality Rates Falling et Extremely High Ages : An Investigation Based on a Model Proposed by Coale and Kisker. *Population Studies*, Vol. 49, No. 2, pp. 281–295.