GENERAL CHARACTERISTICS OF MODGEN APPLICATIONS: EXPLORING THE MODEL RISKPATHS

Martin Spielauer

Statistics Canada – Modeling Division
R.H. Coats Building, 24-O
Ottawa, K1A 0T6

martin.spielauer@statcan.gc.ca

RiskPaths is a simple, competing risk, case-based continuous time microsimulation model. Its main use is as a teaching tool, introducing microsimulation to social scientists and demonstrating how dynamic microsimulation models can be efficiently programmed using the language Modgen.

Modgen is a generic microsimulation programming language developed and maintained at Statistics Canada.

RiskPaths as well as the Modgen programming language and other related documents are available at www.statcan.gc.ca/microsimulation/modgen/modgen-eng.htm

1 Introduction

Modgen is a microsimulation model development package developed by and distributed through Statistics Canada. It was designed to ease the creation, maintenance, and documentation of microsimulation models without the need for advanced programming skills as a prerequisite. It accommodates many different model approaches (continuous or discrete time, case-based or time-based, general or specialized, etc.) Modgen also provides a common visual interface for each model that implements useful functionality such as scenario management, parameter input, the display of output tables from a model run, graphical output of individual biographies, and the display of detailed Modgen-generated model documentation.

In this discussion we introduce a simple microsimulation model called RiskPaths that has been implemented using Modgen. We start with a description of its underlying statistical models and

then explore follow-up questions, such as what microsimulation can add to the initial statistical analysis and what other benefits microsimulation can bring to the overall analysis. We then demonstrate parts of Modgen's visual interface to examine elements of the RiskPaths model.

RiskPaths can be used as a model to study childlessness and was developed for training purposes. Technically, RiskPaths is a demographic single sex (female only), data-driven, specialized, continuous time, case-based, competing risk cohort model. It is based on a set of piecewise constant hazard regression models.

In essence, RiskPaths allows the comparison of basic demographic behaviour before and after the political and economic transitions experienced by Russia and Bulgaria around 1989. Its parameters were estimated from Russian and Bulgarian data of the Generations and Gender Survey conducted around 2003/04. Russia and Bulgaria comprise interesting study cases since both countries, after the collapse of socialism, underwent the biggest fertility declines ever observed in history during periods of peace. Furthermore, demographic patterns were very similar and stable in socialist times for both countries, which helps to justify the use of single cohorts as a means of comparison (one representing life in socialist times, the other the life of a post-transition cohort). In this way, the model allows us to compare demographic behaviour before and after the transition, as well as between the two countries themselves.

2 RiskPaths: The underlying statistical models

2.1 General description

Being a model for the study of childlessness, the main event of RiskPaths is the first pregnancy (which is always assumed to lead to birth). Pregnancy can occur at any point in time after the 15th birthday, with risks changing by both age and union status. The underlying statistical models are piecewise constant hazard regressions. With respect to fertility this implies the assumption of a constant pregnancy risk for a given age group (e.g. age 15-17.5) and union status (e.g. single with no prior unions).

For unions, we distinguish four possible state levels:

- single
- the first three years in a first union
- the following years in a first union
- all the years in a second union

(After the dissolution of a second union, women are assumed to stay single). Accordingly, we model five different union events:

- first union formation

- first union dissolution
- second union formation
- second union dissolution
- the change of union phase which occurs after three years in the first union.

The last event (change of union phase) is a clock event – it differs from other events in that its timing is not stochastic but predefined. (Another clock event in the model is the change of the age index every 2.5 years) Besides unions and fertility, we model mortality--a woman may die at any point in time. We stop the simulation of the pregnancy and union events either when a women dies, or at pregnancy (as we are only interested in studying childlessness), or at her 40^{th} birthday (since later first pregnancies are very rare in Russia and Bulgaria and are thus ignored for this model).

At age fifteen a woman becomes subject to both pregnancy and union formation risks. These are competing risks. We draw random durations to first pregnancy and to first union formation. There are two additional competing events at this stage—mortality and change of age group. (As we assume that both pregnancy and union formation risks change with age, the risks underlying the random durations only apply for a given time period--2.5 years in our model--and have to be recalculated at that point in time.)

In other words, the 15^{th} birthday will be followed by one of these four possible events:

- the woman dies
- she gets pregnant
- she enters a union
- she enters a new age group at age 17.5 because none of the first three events occurred before age 17.5

Death or pregnancy terminates the simulation. A change of age index requires that the waiting times for the competing events union formation and pregnancy be updated. The union formation event alters the risk of first pregnancy (making it much higher) and changes the set of competing risks. A woman is then no longer at risk of first union formation but becomes subject to union dissolution risk.

2.2 Events and parameter estimates

2.2.1 First pregnancy

As outlined above, first pregnancy is modeled by an age baseline hazard and relative risks dependent on union status and duration. The following Table 1 displays the parameter estimates for Bulgaria and Russia before and after the political and economical transition.

Table 1: First pregnancy risks

	Bulgaria	Russia
15-17.5	0.2869	0.2120
17.5-20	0.7591	0.7606
20-22.5	0.8458	0.8295
22.5-25	0.8167	0.6505
25-27.5	0.6727	0.5423
27.5-30	0.5105	0.5787
30-32.5	0.4882	0.4884
32.5-35	0.2562	0.3237
35-37.5	0.2597	0.3089
37.5-40	0.1542	0.0909

	before 1989 transition		10 years after transition: 1999+	
	Bulgaria	Russia	Bulgaria	Russia
Not in union	0.0648	0.0893	0.0316	0.0664
First 3 years of first union	1.0000	1.0000	0.4890	0.5067
First union after three years	0.2523	0.2767	0.2652	0.2746
Second union	0.8048	0.5250	0.2285	0.2698

The data from Table 1 is interpreted as follows in the model. As long as a woman has not entered a partnership, we have to multiply her age-dependent baseline risk of first pregnancy by the relative risk "not in a union". For example, the pregnancy risk of a 20 year old single woman of the pre-transition Bulgarian cohort can be calculated as 0.8458*0.0648 = 0.05481. At this rate of λ =0.05481:

- The expected mean waiting time to the pregnancy event is $1/\lambda = 1/0.05481 = 18.25$ years;
- The probability that a women does not experience pregnancy in the following 2.5 years (given that she stays single) is $\exp(-\lambda t) = \exp(-0.05481*2.5) = 87.2\%$.

Thus at her 20th birthday, we can draw a random duration to first pregnancy from a uniform distributed random number (a number that can obtain any value between 0 and 1 with the same probability) using the formula:

RandomDuration = -ln(RandomUniform) / λ;

As we have calculated above, in 87.2% of the cases, no conception will take place in the next 2.5 years. Accordingly, if we draw a uniform distributed random number smaller than 0.872, the corresponding waiting time will be longer than 2.5 years, since -ln(RandomUniform) / λ = -ln(0.872)/0.05481 = 2.5 years. A random draw greater than 0.872 will result in a waiting time smaller than 2.5 years–in this situation, if the woman does not enter a union before the pregnancy event, the pregnancy takes place in our simulation.

To continue this example, let us assume that the first event that happens in our simulation is a union formation at age 20.5. We now have to update the pregnancy risk. While the baseline risk still stays the same for the next two years (i.e. 0.8458), the relative risk is now 1.0000 (as per the reference category in Table 1) because the woman is in the first three years of a union. The new hazard rate for pregnancy (applicable for the next two years, until age 22.5) is considerably higher now at 0.8458*1.0000 = 0.8458. The average waiting time at this rate is thus only 1/0.8458 = 1.18 years and for any random number greater than exp(-0.8458*2)=0.1842 the simulated waiting time would be smaller than two years. That is, 81.6% (1 – 0.1842) of women

will experience a first pregnancy within the first two years of a first union or partnership that begins at age 20.5.

2.2.2 First union formation

Risks are given as piecewise constant rates changing with age. Again we model age intervals of 2.5 years. These are the rates for women prior to any conception, as such an event would stop our simulation.

Table 2: First union formation risks

	before 1989 transition		10 years after transition: 1999+	
	Bulgaria	Russia	Bulgaria	Russia
15-17.5	0.0309	0.0297	0.0173	0.0303
17.5-20	0.1341	0.1342	0.0751	0.1369
20-22.5	0.1672	0.1889	0.0936	0.1926
22.5-25	0.1656	0.1724	0.0927	0.1758
25-27.5	0.1474	0.1208	0.0825	0.1232
27.5-30	0.1085	0.1086	0.0607	0.1108
30-32.5	0.0804	0.0838	0.0450	0.0855
32.5-35	0.0339	0.0862	0.0190	0.0879
35-37.5	0.0455	0.0388	0.0255	0.0396
37.5-40	0.0400	0.0324	0.0224	0.0330

The parameterization example given in Table 2 has the following interpretation: the first union formation hazard of Bulgarian women of the first cohort is 0 until the 15^{th} birthday; afterwards it changes in time steps of 2.5 years from 0.0309 to 0.1341, then from 0.1341 to 0.1672, and so on. The risk is highest for the age group 20-22.5--at a rate of 0.1672, the expected time to union formation is 1/0.1672=6 years. A women who is single on her 20^{th} birthday has a 34% probability of experiencing a first union formation in the following 2.5 years (p=1-exp(-0.1672*2.5)).

2.2.3 Second union formation

A woman becomes exposed to the second union formation risk if and when her first union dissolves. As a difference to the first union formation which is based on age, this process does not start at a fixed point in time but is triggered by another event (first union dissolution). Accordingly, the time intervals of the estimated piecewise constant hazard rates refer to the time since first union dissolution.

Table 3: Second union formation risks

	before 1989 transition		10 years after transition: 1999+	
	Bulgaria	Russia	Bulgaria	Russia
<2 years after dissolution	0.1996	0.2554	0.1457	0.2247
2-6 years after dissolution	0.1353	0.1695	0.0988	0.1492
6-10 years after dissolution	0.1099	0.1354	0.0802	0.1191
10-15 years after dissolution	0.0261	0.1126	0.0191	0.0991
>5 years after dissolution	0.0457	0.0217	0.0334	0.0191

2.2.4 Union dissolution

Both first and second unions can dissolve, with such processes starting at the first and second union formations, respectively. As the sample size is very small for the modeling of the second union dissolution event we do not distinguish the before and after transition cohorts for this event.

Table 4: First union dissolution risks

	before 1989 transition		10 years after transition: 1999+	
	Bulgaria	Russia	Bulgaria	Russia
First year of union	0.0096	0.0380	0.0121	0.0601
Union duration 1-5	0.0200	0.0601	0.0252	0.0949
Union duration 5-9	0.0213	0.0476	0.0269	0.0752
Union duration 9-13	0.0151	0.0408	0.0190	0.0645
Union duration >13	0.0111	0.0282	0.0140	0.0445

Table 5: Second union dissolution risks

	Bulgaria	Russia
First 3 years of union	0.0371	0.0810
Union duration 3-9	0.0128	0.0744
Union duration 9+	0.0661	0.0632

2.2.5 Mortality

In this sample model, we leave it to the model user to either set death probabilities by age or to "switch off" mortality allowing the study of fertility without interference from mortality. In the latter case, all women reach the maximum age of 100 years. If the user chooses to simulate mortality, the specified probabilities are internally converted to piecewise constant hazard rates (based on the formula $-\ln(1-p)$ for $p<1$) so that death can happen at any time in a year. If a probability is set to 1 (as is the case when age=100), immediate death is assumed.

3 What do we expect from the microsimulation model RiskPaths?

3.1 What can simulation add to statistical analysis?

Before we can answer the question of what simulation can add to statistical analysis, we first need a good understanding of what the statistical results presented in the previous section reveal. The estimation results for the two countries and two cohorts allow us to study similarities and differences between the countries, as well as the changes in parameters over time *separately* for each of the individual processes. We see a remarkable similarity in parameters across the two countries especially for the pre-transition cohorts. Bulgaria differs from Russia basically only in the three times lower union dissolution risks and the slower speed of second union formation. Accordingly, comparing the pre- and post-transition cohorts, we find dramatic changes in most processes. The risk of first births was halved in the first three years of the first union with no later recovery, although the parameters stayed relatively unchanged after three years in a union. Also, in second unions, fertility dropped by more than 50%. The biggest difference between the two countries after the transition is in first union formation--rates halved in Bulgaria but stayed stable in Russia. For first union dissolution we see the opposite picture--union dissolution risks increased by around 40% in Russia while staying almost unchanged in Bulgaria.

These are typical examples of insights we can gain by *single process analysis*. We have separated a complex system into its component processes and studied the changes within those processes. In the case of fertility we have introduced relative risks--we study how certain factors (here, different union statuses) influence a *single* process. This is a very typical analytical question; scientific literature is rich of this kind of research.

The power of microsimulation unfolds when we study various processes simultaneously. Even in our very simple demographic example, results are difficult to interpret when we are interested in the effect of changes in single processes on aggregate outcomes. For example, what is the effect of Russia's 40% increase in union dissolution risks on childlessness? The effect will depend on fertility out of unions and in second unions as well as the speed of second union formation. The relative risk of fertility is higher in second unions than after three years in the first union, but second union formation takes time (during which fertility is very low) and not all women enter a second union. Do these effects cancel themselves out or does union dissolution affect fertility - and in which direction? Such questions invite us to use microsimulation for *sensitivity analysis*. How do aggregate outcomes change in response to the change of a single parameter? Note that we now have moved analysis from the level of a single process to an *analysis of system behaviour*.

A comparison of the two cohorts invites a further type of system analysis--what is the relative contribution of the change in single processes to the aggregated outcome? Comparing the two

simulated cohorts we see that childlessness has increased considerably in both countries but even more so in Bulgaria. We can use microsimulation to *decompose* the contributions of the changes in the various processes to the aggregate change. How much would childlessness have changed if only fertility parameters changed? What is the contribution of changes in union formation? Has the increase in union dissolution risk contributed to the increase in childlessness in Russia? Of course, the aggregate change is not the simple arithmetic sum of partial effects. Some process changes might have a stronger or weaker effect in the presence of changes in other processes. For example, the effect of the change in fertility in second unions will heavily depend on the likelihood of being in a second union which is subject to first union formation and dissolution risks. Microsimulation can help us to identify and better understand such interactions.

Looking at the post-transition cohort, we have already entered the domain of *predictions*. As data were collected 14 years after the transition, in reality no post-transition cohort has gone through its whole reproductive period. Thus, for cohort measures like childlessness, the assessment of consistency with other data sources is limited to a comparison with other projections. But we can also use our model for predictions under alternative assumptions on future changes in processes. We might have a theory that leads to the assumption that only parts of the observed changes are of a permanent nature (e.g. caused by cultural change) while others are transitory (e.g. resulting from economic crisis, therefore reversible with economic recovery). What would happen if fertility rates moved back to their initial values while slower (later) union formation persisted--or vice versa? Such an analysis can produce surprising results, as it is not always a reversal of the process which initially had the biggest overall impact that will generate the biggest opposite effect.

Are there policy implications? While our model is of course too simple for policy analysis, it does not require much imagination to see how microsimulation can support policy making.

- In many cases, the studied phenomena are of direct policy relevance. Fertility decline will for example impact the sustainability of social security systems. A good demographic projection model can therefore produce valuable data input for subsequent good-quality planning. Microsimulation is the tool to combine separate statistical models into projection models.
- Events simulated in a microsimulation model can also be policy targets themselves. A government might aim at influencing fertility. This is possible if policies exist which are capable of influencing the modeled processes. However, we first have to be able to understand the individual contribution of those processes to the aggregated outcome we aim to change; thus we need microsimulation. If we can attach price tags on such policies, we are also able to use microsimulation to find the most cost-efficient policy mix – and to study possible side effects. (Socialist Russia and Bulgaria actually had a set of powerful policies in

place, such as bachelor taxes and privileged access to housing for young married couples. The price tag for regulating individual life choices turned out to be rather high.)

- Microsimulation allows us to complement the models resulting from statistical analysis with detailed policy scenarios and economic accounting models. It provides a very natural tool for policy simulation, as policies are defined at the individual or micro level. This leads to applications which integrate demographic and economic modeling.

3.2 Desired features of a RiskPaths microsimulation model

3.2.1 Input: Parameter tables, scenarios, and simulation settings

Even being a very simple model, RiskPaths has around 130 parameter values which users should be able to set and store conveniently. We would expect these parameters to be well-organized in the microsimulation application, appearing as easy-to-access (or navigate) labelled tables which could be read or modified as required

When using a model we typically create different scenarios, i.e. different parameterizations of the model. We need to be able to save these scenarios so that certain simulations can be reproduced in future. Scenarios contain all parameter tables and, ideally, supplementary text descriptions or notes that outline the specific changes embedded in each scenario. Additionally, scenarios should include scenario settings, such as the number of simulated cases (given that RiskPaths is a case-based model), A large sample size will reduce Monte Carlo variation but comes at the cost of slower simulation runs. If we are only interested in broad aggregates, then smaller sample sizes might suffice. On the other hand, a detailed analysis of rare events or a detailed breakdowns of results (e.g. by age groups) would require large samples. Additionally, users might not wish to produce all available output. Narrowing down the desired output can again speed up simulations but also leads to a more concise and focused presentation of results according to user needs.

All of the above (parameter tables, descriptive notes, number of cases, choice of output to produce) is part of a scenario. For our RiskPaths applications, we would expect all this information to be stored together for a given scenario and we would expect it to be easily retrieved, viewed, and modified.

3.2.2 Output and output views

Microsimulation models can produce output on two levels: micro and macro. A microsimulation application could conceivably write all individual level characteristics and all their changes over time into a file and leave it to the user to analyse the resulting data file with statistical software. In the RiskPaths case, this would lead to a file storing the dates of all simulated events that occur over the simulated life course of each single individual. Only six events can happen in a simulated life, so each data record would contain at most six variables: four union formation /

dissolution events, conception, and death. For more complex applications, file size and complexity could be enormous.

As well as such a longitudinal file, we might also be interested in cross-sectional output, recording the states of all individuals at a certain point in time. While the use of such a file is rather limited when simulating a single cohort, it would resemble a cross-sectional survey or population census in a population model.

Usually, a model user will not be interested in micro files per se but in the analysis that is performed on them. The user will typically aggregate data and produce summary indicators and tables. If model developers already know how simulated data will or should be analyzed, such measures and tables can already be calculated and produced within a microsimulation application run. In this case, users would not need to run additional statistical routines; they could see results immediately after a simulation was performed. In our RiskPaths model, output does not exceed a small number of tables and summary indicators which we expect to be produced within the application. We are interested in age-specific fertility rates, childlessness, the mean age at first conception, first conception by union status, and some mortality measures.

Just as with parameter tables, aggregated model output also requires organization. We might want to present some summary measures of one or several related behaviours together in a table and we surely want to order table output in a meaningful way. Additionally, as with parameters, we would expect table results to be labelled for easy reading and understanding.

Because all microsimulation results are subject to Monte Carlo variation, aggregated numbers are only one view of the results. We might also be interested in getting distributional information on each table value. Such information would help us to set an appropriate population size sufficient for a desired level of result precision.

A special type of micro-data output is the graphical display of individual careers. This can be a helpful feature, as it provides users with a window to the simulated individuals, and thus a way to see the operation of the statistical models. This can also be useful for model developers as it supports model debugging. Since RiskPaths is a training tool, we are interested in displaying how individual biographies result from statistical processes. Thus, besides life course events, we might also want to see how the risks of the alternative events change over time and life course situations.

3.2.3 User interface and documentation

So far, we have formed expectations about the content, display, and organization of model input and output data. From the user perspective, do we just have to add a start button to complete the microsimulation application? Almost all contemporary software applications contain help files. As users of microsimulation models, we should expect access to detailed online help, not only on

the use of the modeling software itself but also on the model's specific elements and the interrelationships amongst those elements.

4 Exploring the Modgen application RiskPaths

In the remainder of this discussion we provide a quick explorative tour of the visual interface provided by Modgen for the RiskPaths model. To run RiskPaths, both the Modgen Prerequisites application and the RiskPaths executable have to be installed on your computer. As is true for all Modgen applications, RiskPaths contains a help system including documentation on the (model-independent) Modgen user interface and the actual RiskPaths model itself. Accordingly, in the description below, we concentrate on the central steps of running RiskPaths, leaving it to you to explore the model and software in depth with the assistance of the detailed help files.

4.1 The user interface

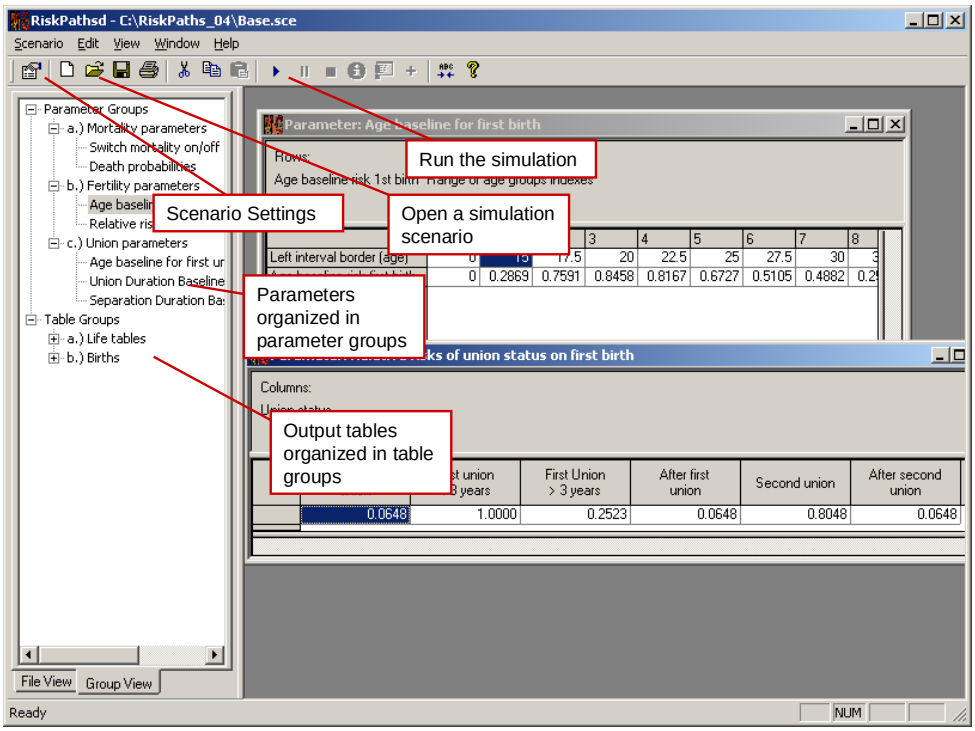
All Modgen applications have the same graphical user interface (Figure 1) which consists of the following parts:

- a menu bar and a toolbar to administer and run scenarios, as well as to get help
- a selection window containing a hierarchically grouped list of all model parameters and output tables
- a frame in which all corresponding parameters or tables can be displayed

When starting the RiskPaths.exe application, the selection window and table frame are empty, as we first have to load (or create) a simulation scenario. To do so, follow the following steps:

- Open the simulation scenario 'base.sce'. This can be done by clicking the 'Open' button or by selecting 'Open…' from the 'Scenario' menu.
- Choose the scenario settings--the settings dialog box can be accessed by clicking the 'Settings' button or by selecting 'Settings…' from the 'Scenario' menu. Specify a small number of simulated cases (e.g. 10,000) so that your first model runs quickly. Also, ensure that 'MS Access tracking' is switched on. This will allow you to view individual biographies using the BioBrowser tool that comes with Modgen
- Save your scenario under a new name by selecting 'Save as…' from the 'Scenario' menu.

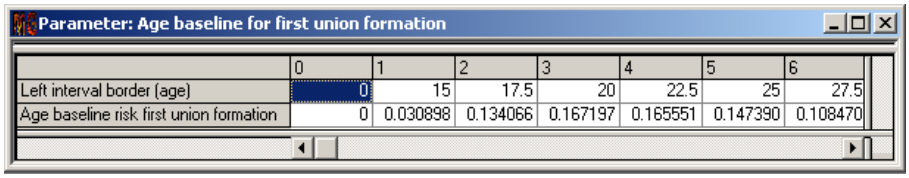
Figure 1: The RiskPaths application



4.2 Parameter tables

Users of a Modgen application have control over all parameters contained in the model's parameter tables. An individual parameter table can be selected by clicking its list entry in the selection window. The table is then displayed in the display frame in which it can also be edited. Modgen parameter tables can have any number of dimensions, ranging from a parameter with a single checkbox to parameters with numerous characteristics or dimensions (e.g. region, sex, age, time).

Figure 2: Parameterization of first union formation risks



	0	1	2	3	4	5	6
Left interval border (age)	0	15	17.5	20	22.5	25	27.5
Age baseline risk first union formation	0	0.030898	0.134066	0.167197	0.165551	0.147390	0.108470

4.3 Performing a simulation run

Click the 'Run/resume' button or select 'Run/resume' from the 'Scenario' menu. The progress of the simulation is displayed in a progress dialog box. A small sample of 10,000 actors takes around 20 seconds to run. After the model run is complete, all output tables will have been updated by Modgen.

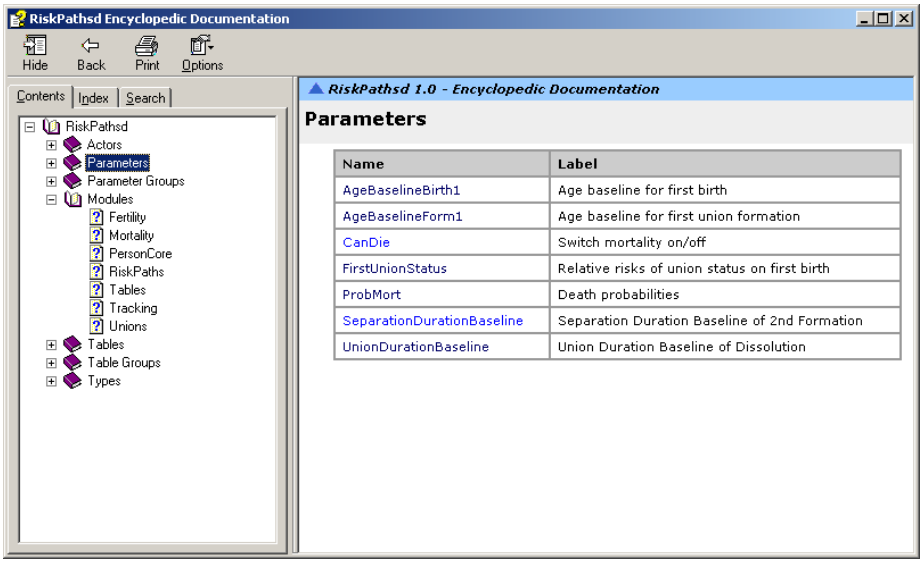
4.4 Table output: aggregates and distributions

Simulation results are written to predefined output tables. Note that the values displayed in the output table represent only one of several possible views on the results. By right-clicking a table, a properties sheet for the table can be accessed. Among other things, this allows the display of distributional information (standard errors and the coefficient of variation) of all simulated values. Table contents can also be copied and pasted. You have the choice to copy the table as displayed, or all dimensions of the table at once (if there are more than two dimensions).

4.5 Model help and documentation

As is true with all Modgen applications, RiskPaths provides help files of various types. Two are related to Modgen itself--a general user guide for the visual interface plus release notes for Modgen. The other help files are model-specific. All Modgen applications contain a detailed encyclopaedic model documentation file. This documentation is automatically created from properly commented code.

Figure 3: Model documentation



4.6 Graphical output of individual histories

The Modgen Biography Browser (BioBrowser) application is a tool for the graphical display of individual life courses. This view on the simulation results is especially useful for model debugging. In order to use the tool, the tracking feature has to be switched on in the scenario settings. The list of variables to be tracked also has to be declared by the model developer in the model code via a tracking statement. Modgen than tracks all changes of those variables included in the tracking statement for a sample of simulated actors (where the size of this sample is specified as one of the scenario settings).

To display biographies created by RiskPaths, just start the BioBrowser application and load the tracking-file of your simulation scenario, e.g. Base(trk).mdb.

Figure 4: BioBrowser