

# FPATH - Program Description

Victoria Yanulevskaya and Juha M. Alho

*FPATH* is a program intended to extend the application of results from *PEP* (for a description, visit <http://www.joensuu.fi/statistics/juha.html>) to economic OverLapping Generations (OLG) models, in which agents are allowed to revise their lifetime economic plans, as they notice that population evolution has not followed the expected path. For this purpose *FPATH* calculates a numerical approximation to the conditional expectation of future population at future years for a (typically random) *subset of paths*. (Alternatively, we can think of the conditional expectation as being *a forecast of what would be a forecast* in a future year.) The conditional expectations relating to a given path are calculated as the average (or median) of paths that belong to the neighborhood of the path.

This subset of paths, and the accompanying conditional expectations that *FPATH* produces form the input that is fed into the OLG model.

## 1. Input Information

*Number of paths.* It is assumed that a user of *FPATH* has available a stochastic population forecast produced with *PEP* with  $N + n$  sample paths, where  $n$  = number of paths for which output of *FPATH* is needed, and  $N$  is the number of paths to be used in the calculation of conditional expectations (e.g.,  $N = 9,000$  and  $n = 300$ ). Since  $n$  is the number of paths fed into the OLG model, it limits the accuracy of those final calculations. On the other hand,  $N$  determines the density of paths around the  $n$  output paths, so it limits the accuracy of the conditional expectation calculations.

*Age and sex grouping.* As a default, *PEP* produces output by single years of age for both males and females. These are called  $Px\_d1.s1$ , where  $x$  is the number of sample path. These data can serve as input data for *FPATH*. However, the default expectation for *FPATH* is that the user has somehow aggregated data (e.g., using 5-year age groups and combining males and females together). When this is the case, *PEP* should be requested to produce such aggregate files. When females and males are combined together, the files will have names  $Px\_0.c1$ . (Extensions  $m1$  and  $f1$  are also possible if only males or females are considered.) After a *PEP* run, the program *COMBINE* can also produce such files. (However, *FPATH* does not recognize extensions  $c2$ ,  $c3$  etc. That corresponds to repeated *COMBINE* run.) In all cases these files have names of age-groups (e.g., 0-4, 5-9,... if males and females are combined) on the first row. *FPATH* calculates the number  $m$  of age-groups. This is the number of columns in the files. For example, if the highest age is 100+ and 5-year age grouping is used, the number of columns is  $m = 21$ .

In addition to the paths, the calculation of conditional expectations requires annual files  $Yy\_d1.s1$  produced by *PEP* or *COMBINE* that use the same age and sex grouping.

*Timing of Conditional Expectations.* *FPATH* asks the user what is the last year  $L$  for which a conditional forecast is desired, and the frequency  $S$  of intermediate years. Thus,  $T = L/S$  is the number of years for which a conditional expectation is produced.

*Specification of Path Neighborhoods.* The  $n$  paths are chosen from the beginning of *PEP* output, so they will have numbers  $x = 1, 2, \dots, n$ . For each of the  $n$  paths, conditional expectations of future population vector (based on chosen age and sex grouping, and for the chosen future years) is estimated using either the *mean* or *median* of the size of the age group

among paths that belong the neighborhood of a given path. The median is a more robust measure of location but the mean corresponds more closely to the conditional expectation.

Distance is measured using a standard Euclidean distance. Having a distance measure available, *FPATH* determines those paths that are uniformly the closest to the target path. The number of paths belonging to the neighborhood is allowed to diminish as we go towards a more distant future. The procedure is given using two parameters. First, the user is requested to give the minimum number of points of the neighborhood must contain. As a rule of thumb, this number  $Q$  should never be below 30. The second parameter gives the relative decrease  $\alpha \geq 0$  in the number of points in the neighborhood when we move  $S$  years ahead. Thus, if the neighborhood contains  $Q$  points when the last conditional expectation is calculated at time  $(T - 1)*S$  for year  $T*S$ , the neighborhood should contain  $\exp(\alpha)Q$  points at time  $(T - 2)*S$ , when the conditional expectations are calculated for the years  $(T - 1)*S$  and  $T*S$ , etc.

## 2. Form of Output Files

Corresponding to each of the  $n$  paths numbered  $x = 1, \dots, n$ , another file is produced as output, called  $Fx\_0.c1$ . Suppose years  $S, 2*S, \dots, T*S$  are considered, where  $L = T*S$  is the last forecast year of interest. Then, the structure of files  $Fx\_0.c1$  is as follows:

- first row is the same as the first row of files  $Px\_0.c1$  (i.e., it contains the names of the age-groups).
- after that there will be  $T$  blocks.
- the first is obtained from the original point forecast  $P0\_0.c1$  by taking (after the title row) rows  $S*1, S*2, \dots, S*T$ . Thus, there are  $T$  rows in the first block. (This is *not* the same as the conditional expectation, although empirically the differences noted have been small. However, the use of the point forecast corresponds to the actual practice in many if not most decision making circumstances.) The leftmost column of the file contains labels for the  $T$  years in question in parenthesis (e.g., (5), (10), ..., (50))
- after that  $T - 1$  blocks follow that are produced separately by the program *FPATH* using files  $Yy\_0.c1$ . Out of the blocks, the first has  $T - 1$  rows and they represent a forecast with jump-off population coming from row  $x$  of file  $Yy\_0.c1$ , where  $y = S$ , and the forecast is made  $S*(T - 1)$  years ahead, or for years  $2*S, 3*S, \dots, T*S$ . Again, labels for the years are given in the first column (e.g., (10), (15), ..., (50)). The second block has  $T - 2$  rows and they represent a forecast with jump-off population from row  $x$  of file  $Yy\_0.c1$ , where  $y = 2*S$ , and the forecast is made  $S*(T - 2)$  years ahead. The last block contains one row. It represents a forecast with jump-off population from row  $x$  of file  $Yy\_0.c1$ , where  $y = S*(T - 1)$ , and the forecast is made for year  $S*T$ . The last item of the leftmost column is a label for the year  $L = S*T$  (e.g., (50)).

## Appendix. Definition of a Relative Euclidean Distance.

*F*PATH uses a standard Euclidean distance to measure the proximity of population vectors and only the ordered distances matter, not their actual values. However, for diagnostic purposes it may be desirable to develop a feeling of large the neighborhoods are. As the Euclidean distance depends on the number of age groups  $m$  and on the absolute size of the population, it is difficult to form a view of its practical magnitude. To define a more interpretable measure, suppose  $V(i,t) = (V_1(i,t), \dots, V_m(i,t))$  is the “target population vector” corresponding to path  $i = 1, \dots, n$  in a given future year  $t$ . Consider an arbitrary other path  $V(j,t) = (V_1(j,t), \dots, V_m(j,t))$ ,  $j = n + 1, \dots, n + N$ . Then, a normalized Euclidean distance between the two vectors at  $t$  can be defined as

$$D(i,j,t) = \left\{ \frac{1}{m} \sum_{k=1}^m (V_k(i,t) - V_k(j,t))^2 \right\}^{1/2} / \frac{1}{m} \sum_{k=1}^m V_k(i,t).$$

For the purpose of finding the closest paths, the use of the normalized measure is equivalent to the use of the standard Euclidean distance. But intuitively, the normalized measure is akin to a coefficient of variation. Or, it measures average distance between the age groups relative to the average size of the target vector.