Chapter 4
Adding New Dimensions

Abstract This chapter shows how implementing new dimensions into the microsimulation model. As examples, we add two dimensions that can hardly be implemented in a classical projection model: the labour force participation and the sector of activity (formal/informal). Those modules are implemented through statistical modeling using regression parameters. They use as predictors individual characteristics, such as age, sex, region, education, and for women, a binary variable indicating if she gave birth to a child within the last five years. Those two new variables are thus dynamically implemented, as assumptions on fertility have a direct impact on their outcomes.

Keywords Microsimulation · Population projection · Demography · Method · SAS

4.1 Adjusting the Workspace for the Addition of New Dimensions

In Chap. 3, we replicated in a microsimulation framework what a standard multistate model can do. In this chapter, we will increase complexity by adding two dimensions: the labour force participation and the sector of activity. While these variables can be derived from the outcome of a multistate model (e.g. using the resulting population and applying predefined participation rates), the microsimulation can implement them dynamically. In the example, we present in this chapter, the labour force participation and the sector of activity are calculated using parameters from statistical models. Predictors include age/cohort, sex, region of residence, education and a binary variable stating whether the individual is a woman who gave birth within

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-79111-7_4) contains supplementary material, which is available to authorized users.

G. Marois and S. KC, *Microsimulation Population Projections with SAS*,
SpringerBriefs in Population Studies,
https://doi.org/10.1007/978-3-030-79111-7_4

the last 5 years. As this last predictor suggests, assumptions as to fertility have an impact on the labour force outcome.

The modules for labour force participation and the sector of activity reassess the individual outome of these variables at the end of each period using personal characteristics as determinants. Because of the availability of data used in the statistical model, these modules do not take into account the past labour force participation and sector of activity of the individual. In other words, what is modelled is the probability of being in the labour force rather than the probability of entering or leaving the labour market, or the probability of changing the sector of activity. Consequently, the modeling can project reliable crossectional values, but it does not allow for longitudinal analysis, as life courses may be inconsistent.

The code file "Chapter 4 – Adding new dimensions.sas includes" the complete code of the microsimulation with two additional dimensions that are explained in this chapter. Below, we will explain the difference from the file used in Chap. 3, which replicated a multistate model. First, we change the name of the scenario. In the support documents provided with this book (Chapter ESM), all necessary files can be found in the folder "chapter4".

```
%let scenario_name=Chapter4;
```

The parameter files for modules for labour force participation and the sector of activity were already imported in Chap. 2 (with the macro function *import* and the *sort procedure*). As a reminder, the code lines for this purpose were:

```
%import
("C:\Users\Guillaume\Desktop\Microsimulation\%scenario_name\parameters\
parameters\lfp.csv",param.lfp);
%import
("C:\Users\Guillaume\Desktop\Microsimulation\%scenario_name\parameters\
parameters\formal.csv",param.formal);
%import
("C:\Users\Guillaume\Desktop\Microsimulation\%scenario_name\parameters\
parameters\lfp_imput.csv",param.lfp_imput);
%import
("C:\Users\Guillaume\Desktop\Microsimulation\%scenario_name\parameters\
parameters\formal_imput.csv",param.formal_imput);
```

4.2 Labour Force Participation Module

Labour force participation rates (P) are estimated from a logit regression model. Logit models can be estimated with SAS using the LOGISTIC procedure.[1] When modeling a binary variable such as labour force participation, logit models are preferred over linear models, as the predicted value of a logit model can only range from 0 to 1. If the predicted outcome has more than two categories, multinomial or ordered logit may

[1] The documentation for this procedure can be consulted here: https://support.sas.com/documentation/onlinedoc/stat/131/logistic.pdf.

be more appropriate. The logit model used in the labour force participation module is based on data from the National Sample Survey on Employment and Unemployment 2017/2018 (population aged 15–74; n = 323,092). The model is described in Eq. 4.1:

$$\begin{aligned}\text{logit}(P) = {} & \beta_{s,0} + \beta_{s,1} AGEGR + \beta_{s,2} AGEGR^2 + \beta_{s,3} EDU + \\ & \beta_{s,4} REGION + \beta_{s=F,5} YOUNG\_KID + \beta_{s=F,6} POSTSEC * YOUNG\_KID + \beta_{s,7} EDU * \\ & AGEGR + \beta_{s,8} EDU * AGEGR^2 \end{aligned} \tag{4.1}$$

The logit of a probability corresponds to the natural logarithm of its odds. Therefore, the logit of the participation rate (P) is log(P/(1 − P)), and the rate P can be calculated from the parameters, such as:

$$\text{P} = \frac{\begin{array}{c}\exp(\beta_{s,0} + \beta_{s,1} AGEGR + \beta_{s,2} AGEGR^2 + \beta_{s,3} EDU + \beta_{s,4} REGION + \beta_{s=F,5} YOUNG\_KID + \\ \beta_{s=F,6} POSTSEC * YOUNG\_KID + \beta_{s,7} EDU * AGEGR + \beta_{s,8} EDU * AGEGR^2)\end{array}}{\begin{array}{c}1 + \exp(\beta_{s,0} + \beta_{s,1} AGEGR + \beta_{s,2} AGEGR^2 + \beta_{s,3} EDU + \beta_{s,4} REGION + \beta_{s=F,5} YOUNG\_KID + \\ \beta_{s=F,6} POSTSEC * YOUNG\_KID + \beta_{s,7} EDU * AGEGR + \beta_{s,8} EDU * AGEGR^2)\end{array}}) \tag{4.2}$$

Each sex has its own set of parameters and its own intercept. The slope for age, education, and having a young kid is thus assumed to be the same in all regions. The age group is included with a quadratic function, allowing it to be modelled with a reverse U-shape with lower participation rates for younger adults still in school and the elderly. The interaction of age and education allows the model to take into account that the age pattern in labour force participation varies by educational attainment. It was not possible to have region-specific parameters because of the small number of respondents in many categories (such as highly educated people in a specific age range in smaller regions). However, regions have their own gradients.

The max-rescaled R-Square is 0.5342 for the males' model (percent concordant = 91.2) and 0.2374 for the females' one (percent concordant = 76.6). Complete parameters can be found in the parameters file lfp. In Fig. 4.1, we showed the predicted rates from the model by age and education for both males and females (with no young kid). For males, rates are very high for everyone between 25 and 59. The education gap concerns mainly young and older adults, with lower rates for higher educated ones. In other words, more educated men enter later in the labour market, since they stay at school longer, and they also retire earlier, probably because they have better jobs during their working lives and can afford an earlier retirement. For females, the pattern is very different. For all education categories and at any age groups, rates are much lower than for men, generally more than twice as low. Furthermore, the effect of education seems to follow a U-shape, with higher rates for both the highest and lowest categories. These trends are similar to those observed by Kapsos et al. (2014).

The parameter for the variable young_kid is −0.3236, implying that women who gave birth within the last 5 years are much less likely to work. This parameter would thus reduce by about 8.5 percentage points a participation rate that would otherwise have been 42%. The negative impact of having a young_kid is moreover much larger

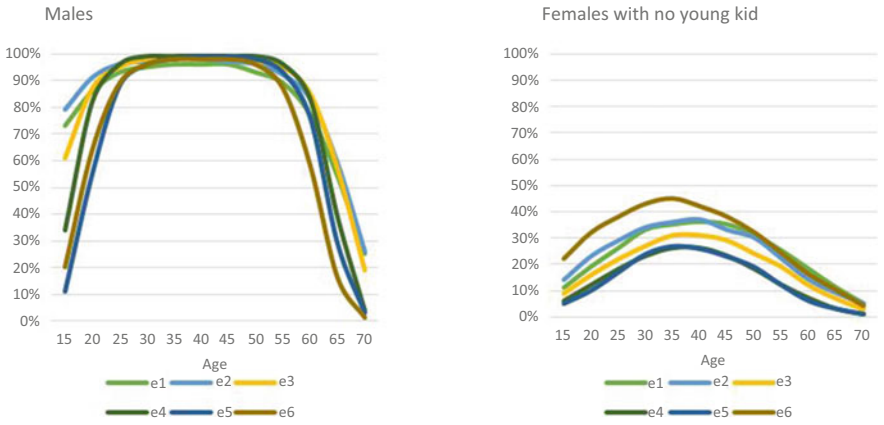

Fig. 4.1 Predicted labour force participation rates from Eq. 4.1 by age and education, India

for women with a postsecondary education than for other women, as the parameter for the interaction between these two variables is −0.3895. Finally, parameters show strong heterogeneity among regions, and also higher participation rates in rural areas than in urban areas of the same region.

For other modules, assumptions are implemented directly as rates that were merged to individuals according to their characteristics. For the labour force participation module, we use regression parameters and therefore, the implementation method is different. Variable-specific parameters will first be merged one by one to the corresponding population. Then, using those parameters, we will calculate the individual probability of participating in the labour force.

To merge the parameter file to the population file, we need to structure it in a particular way, as shown in Fig. 4.2. Each discrete variable needs to have its own column with specific categories on different rows. Parameters corresponding to these categories are on another column, under the label 'variable name'_p. Reference categories (such as edu = e3) also need to be included with a parameter of 0. Otherwise, a missing value would be used in the calculation of the rate, which would result in an error. For continuous variables as well as for intercepts, since they are applied in the calculation of the labour force participation rate for each individual, each is implemented under a specific column, such as agegr_p and agegr2_p for the two parameters of the quadratic form of age, and agegr_edu_p and agegr_edu_p for the quadratic form of the interaction of age with education.

The labour force participation module is implemented once the demographic events are completed, when the age and year are those corresponding to the end of the period. The population file to be used is thus work.pop2 and the module needs to be written right after the "time module" and before cleaning the population file for the next period.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	sex	edu	young_kid	region	intercept	edu_p	agegr_p	agegr2_p	agegr_edu_p	agegr2_edu_p	young_kid_p	young_kid_edu_p	region_p
2	1				-4.4198								
3	1		0								0		
4	1		1								-0.3236		
5	1	e1				0.3929							
6	1	e2				0.9207							
7	1	e3				0							
8	1	e4				-1.5251							
9	1	e5				-2.3395							
10	1	e6				1.0876							
11	1						0.1847						
12	1							-0.00248					
13	1	e1							0.000397				
14	1	e2							-0.0266				
15	1	e3							0				
16	1	e4							0.0685				
17	1	e5							0.1092				
18	1	e6							-0.00883				
19	1	e1								0.000007265			
20	1	e2								0.000251			
21	1	e3								0			
22	1	e4								-0.00087			
23	1	e5								-0.00133			
24	1	e6								0.000007983			
25	1	e1	0									0	
26	1	e2	0									0	
27	1	e3	0									0	
28	1	e4	0									0	
29	1	e5	0									0	
30	1	e6	0									0	
31	1	e1	1									0	
32	1	e2	1									0	
33	1	e3	1									0	
34	1	e4	1									0	
35	1	e5	1									0	
36	1	e6	1									-0.3895	

Fig. 4.2 Screenshot of the parameter file lfp.csv (opened with Excel)

First, we need to merge the parameters file (param.lfp) with the population file (pop2). Using the merge statement as in previous modules would not be optimal, since it would require a specific merging for each variable from the logit model. We thus use a command in Structured Query Language (SQL),[2] which is supported by SAS. We create a new population file (pop_lfp1) that links parameters from the file lfp to individuals for the last population file (pop2, that we select under p.*). Parameters are selected one by one, with the appropriate variables (under t1 to t9). Parameters in each set are joined by their specific correspondent variables. For instance, parameters for education are joined both by sex and education, while parameters for the presence of kids and its interaction with education are joined by sex, education, and presence of kids, and so on for other sets of parameters. In the code, we also specify "where not missing ('name of the parameter)" to join only the values of parameters, as we don't want missing cells to be imported.

[2] For more information about using SQL in SAS, see: https://support.sas.com/resources/papers/proceedings/proceedings/sugi30/257-30.pdf.

```
/*Labor force participation*/
    /*Implementing parameters*/
    proc sql;
      create table pop_lfp1 as
        select
          p.*,
          t1.intercept, t2.edu_p, t3.agegr_p, t4.agegr2_p, t5.agegr_edu_p,
          t6.agegr2_edu_p, t7.young_kid_p, t8.young_kid_edu_p, t9.region_p

        from
          pop2 p

          left join
          ( select sex, intercept
            from param.lfp
            where not missing(intercept)
          ) t1
          on p.sex=t1.sex

          left join
          ( select sex, edu, edu_p
            from param.lfp
            where not missing(edu_p)
          )t2
          on p.sex=t2.sex and p.edu=t2.edu

          left join
          ( select sex, agegr_p
            from param.lfp
            where not missing(agegr_p)
          )t3
          on p.sex=t3.sex

          left join
          ( select sex, agegr2_p
            from param.lfp
            where not missing(agegr2_p)
          )t4
          on p.sex=t4.sex
          left join
          ( select sex, edu, agegr_edu_p
            from param.lfp
            where not missing(agegr_edu_p)
          )t5
          on p.sex=t5.sex and p.edu=t5.edu

          left join
          ( select sex, edu, agegr2_edu_p
            from param.lfp
            where not missing(agegr2_edu_p)
          )t6
          on p.sex=t6.sex and p.edu=t6.edu

          left join
          ( select sex, young_kid, young_kid_p
            from param.lfp
            where not missing(young_kid_p)
          )t7
          on p.sex=t7.sex and p.young_kid=t7.young_kid
```

```
      left join
      ( select sex, edu, young_kid, young_kid_edu_p
        from param.lfp
        where not missing(young_kid_edu_p)
      )t8
      on p.sex=t8.sex and p.edu=t8.edu and p.young_kid=t8.young_kid

     left join
      ( select sex, region, region_p
        from param.lfp
        where not missing(region_p)
      )t9
      on p.sex=t9.sex and p.region=t9.region;
quit;
```

The population file pop_lfp1 now includes individual-specific parameters for the labour force participation module. Starting from this file, we create a new one (pop_lfp2) in which the labour force participation event occurs. For each step of the projection, the labour force variable is first reset to 0 (out of the labour force) for all individuals (labour = 0). We then calculate the individual-specific labour force participation rate for the population affected by the event (those aged between 15 and 74). In our example, we use logit regression parameters. The rate thus corresponds to the exponential of the sum of parameters (multiplied by the value of the variable in the case of continuous variables) divided by 1 + the exponential of the sum parameters.

```
        /*Labour force participation event*/
        data work.pop_lfp2;
        set work.pop_lfp1;

        labour=0;
        if 15<=agegr<74 then do;
        exp_lab = exp(intercept + agegr_p*agegr + agegr2_p*agegr*agegr + edu_p
    + agegr_edu_p*agegr + agegr2_edu_p*agegr*agegr
        + region_p + young_kid_p + young_kid_edu_p);
        prob_lab = exp_lab/(1+exp_lab);
     (…)
```

Once each individual has a specific probability of participating in the labour force, we can proceed to the simulation of the event with the Monte Carlo method. When the rate is higher than the random number, we switch the labour force variable to 1. Finally, we drop parameters for labour force participation from the population file.

```
     (…)
        if rand('uniform')<prob_lab then labour=1;
        end;

        drop intercept agegr_p edu_p agegr_edu_p agegr2_p agegr2_edu_p
     region_p young_kid_p young_kid_edu_p exp_lab prob_lab;
        run;
```

4.3 Sector of Activity Module

In India, as in many developing countries, the informal sector (jobs that are not regulated or monitored by the government, including unpaid jobs) represents a large part of the economy. With the modernisation of the economy, urbanisation, globalisation, the demographic transition, and the expansion of the educational attainment, the informal sector is likely to shrink and be replaced by formal jobs (Cáceres-Delpiano 2012; McCaig and Pavcnik 2015; Siggel 2010).

The sector of activity module is implemented in the same way as the labour force participation module, with logit regression parameters. However, covariates and their interactions differ. More than age, the cohort of birth has a major influence on whether or not an individual is likely to work in the formal sector (McCaig and Pavcnik 2015). Thus, the formalisation of an economy occurs in large part by the replacement of generations, through the mechanism of demographic metabolism (Lutz 2013). Accordingly, the modelling of the sector of activity (S) uses the cohort of birth as an individual determinant, while the age dimension is dismissed. Equation 4.3 describes the model:

$$\begin{aligned}\text{logit}(S) = SEX * (\beta_0 + \beta_1 COHORT + \beta_2 EDU + \beta_3 REGION \ + \ \beta_4 YOUNG\_KID + \\ \beta_5 POST\_SEC * YOUNG\_KID + \beta_6 REGION * COHORT)\end{aligned} \tag{4.3}$$

The model is applied only to the active population of the National Sample Survey on Employment and Unemployment 2017/2018. The max-rescaled R-Square is 0.3080 for the males' model (percent concordant = 78.3) and 0.5235 for the females' model (percent concordant = 88.4). Complete parameters can be found in the parameters file *formal*. As for the labour force participation model, each sex has its own set of parameters and its own intercept.

The cohort is implemented as a continuous variable taking the value of 0 for the cohort born in 1940–1944, 1 for those born in 1945–1949 and so on. The cohort parameter thus captures the secular trend that can be extrapolated for future cohorts entering the labour market. The model also includes an interaction between the cohort and the region in order to take into account the regional disparity in the pace of development. To avoid inconsistencies for regions that already have a very high proportion of the active population working in the formal sector, we added the constraint that region-specific cohort trends need to be positive or equal to 0 ($\beta_1 + \beta_6 \geq 0$).

In Fig. 4.3, we present the extrapolation for future cohorts of the arithmetic average (not weighted by region population) of region-specific cohort parameters. It shows that for both sexes, there is a sharply increasing trend in the proportion of workers in the formal sector. A bit more than 20% of cohorts born in the 50s work in the formal sector, compared to half of cohorts born in the late 90s. When extrapolating trends, the proportions will exceed 60% for cohorts born after 2025. Despite having much lower labour force participation rates, women are slightly more likely to work in the formal sector, but the gap will gradually close. The model also accounts for strong

regional differentials (not shown in the figure). Rates are in general much lower in rural regions than in urban ones, but the difference shrinks gradually over cohorts.

In addition to cohort and region, the model also includes education, a parameter for women that have a young child (having given birth in the last 5 years) and its interaction with the education. As shown in Table 4.1, presenting odds ratios for the educational attainment, education emerges as a key determinant of having a formal job for both males and females, as a steep gradient in parameters is observed between the lowest degree and the highest. Active men with a postsecondary education are about 10 times more likely to work in the formal sector than men with no education (6.699/0.544). This ratio is above 25 for women (17.921/0.689).

Finally, the model includes the negative effect of having a young kid at home for women on the probability of working in the formal sector (−0.845). The positive parameter (0.799) for the interaction of the variables YOUNG_KID and POST_SEC

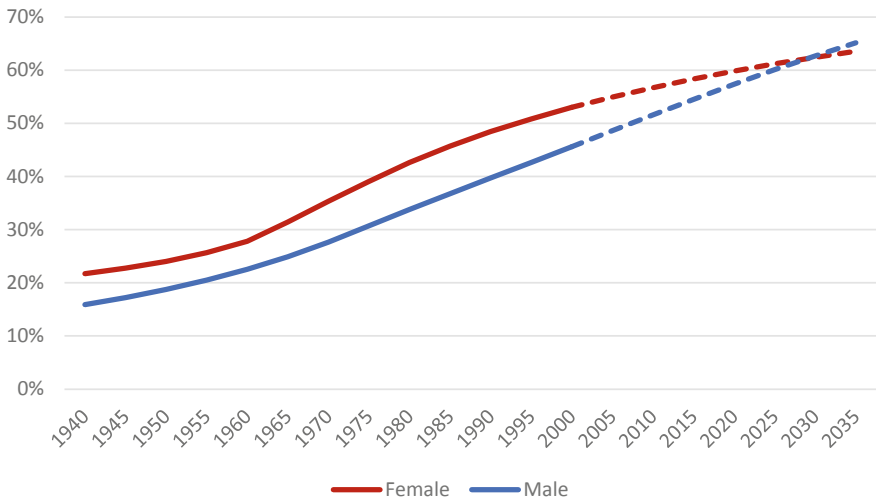

Fig. 4.3 Arithmetic average of region-specific cohort parameters for the sector of activity converted into rate (education = complete primary; no birth in the last 5 years)

Table 4.1 Odds of working in the formal sector by level of educational attainment (exp(β_2) from Eq. 4.3)

Educational attainment	Males	Females
e1—No education	0.654[a]	0.689[a]
e2—Incomplete primary	0.823[a]	0.885
e3—Complete primary	1.000	1.000
e4—Lower secondary	1.370[a]	1.603[a]
e5—Upper secondary	2.168[a]	3.881[a]
e6—Postsecondary	6.699[a]	17.921[a]

[a]<0.0001

Source Authors' calculations from the National Sample Survey on Employment and Unemployment 2017/2018

however suggests that this effect is much less for women with a postsecondary education.

As the sector of activity module also uses regression coefficients, the parameter file "formal" has the same format as the parameter "lfp file" (one different column for each variable and one different column for each set of parameters), as shown in Fig. 4.4. As a reminder, the cohort is implemented as a continuous variable and therefore does not require a category to link it to the corresponding population.

The code to implement parameters is also similar to the one used for the labour force participation module. Using a SQL command, in a new population file called "pop_formal1", we join to the last population file ("pop_lfp2") the parameters from the parameters file "formal" (which is stored in the library *param*), using the appropriate set of variables for each parameter.

	A	B	C	D	E	F	G	H	I	J	K
1	sex	edu	young_kid	region	intercept	edu_p	cohort_p	young_kid_p	young_kid_edu_p	region_p	cohort_region_p
2	1				-0.2422						
3	1		0					0			
4	1		1					-0.845			
5	1	e1				-0.373					
6	1	e2				-0.1217					
7	1	e3				0					
8	1	e4				0.4722					
9	1	e5				1.3562					
10	1	e6				2.8855					
11	1	e1	0						0		
12	1	e2	0						0		
13	1	e3	0						0		
14	1	e4	0						0		
15	1	e5	0						0		
16	1	e6	0						0		
17	1	e1	1						0		
18	1	e2	1						0		
19	1	e3	1						0		
20	1	e4	1						0		
21	1	e5	1						0		
22	1	e6	1						0.7992		
23	1						0.00458				
24	1			AD_rural						-2.3663	
25	1			AD_urban						-0.2655	
26	1			AN_rural						-2.0301	
27	1			AN_urban						-0.622	
28	1			AR_rural						-4.1565	

Fig. 4.4 Screenshot of the parameter file formal.csv (opened with Excel)

```
/*Sector of activity*/
    /*Implementing parameters*/
    proc sql;
      create table pop_formal1 as
        select
          p.*,
          t1.intercept, t2.edu_p, t3.cohort_p, t4.cohort_region_p,
          t5.young_kid_p, t6.young_kid_edu_p, t7.region_p

        from
          pop_lfp2 p

          left join
          ( select sex, intercept
            from param.formal
            where not missing(intercept)
          ) t1
          on p.sex=t1.sex

          left join
          ( select sex, edu, edu_p
            from param.formal
            where not missing(edu_p)
          )t2
          on p.sex=t2.sex and p.edu=t2.edu

          left join
          ( select sex, cohort_p
            from param.formal
            where not missing(cohort_p)
          )t3
          on p.sex=t3.sex


       left join
       ( select sex, region, cohort_region_p
         from param.formal
         where not missing(cohort_region_p)
       )t4
       on p.sex=t4.sex and p.region=t4.region

       left join
       ( select sex, young_kid, young_kid_p
         from param.formal
         where not missing(young_kid_p)
       )t5
       on p.sex=t5.sex and p.young_kid=t5.young_kid

       left join
       ( select sex, edu, young_kid, young_kid_edu_p
         from param.formal
         where not missing(young_kid_edu_p)
       )t6
       on p.sex=t6.sex and p.edu=t6.edu and p.young_kid=t6.young_kid

      left join
       ( select sex, region, region_p
         from param.formal
         where not missing(region_p)
       )t7
       on p.sex=t7.sex and p.region=t7.region;
quit;
```

In a new population file (pop_formal2), we can now simulate the event, which will split workers between the formal and the informal sector. First, in a temporary variable “cohort2”, we need to transform the cohort variable to make it correspond to the one used in the regression model. As a reminder, the cohort born in 1940–1944 has the value 0, while the cohort born in 1945–1949 has the value 1, and so on. Therefore,

the cohort variable used in the sector of activity event should be (cohort—1940)/5. Someone born in 2020 would thus have a value of 16.

```
/*Formal - Informal event*/
data work.pop_formal2;
set work.pop_formal1;

cohort2=(cohort-1940)/5;
(…)
```

Because the sector of activity is modelled using a cross-sectional approach, we reset the variable to 0 (formal = 0, which corresponds to being out of the labour force). For those aged 15 to 74 and in the labour force (labour = 1, which is the outcome of the labour force module of the previous section), we set by default the variable formal to 1, signifying working in the informal sector. We then use parameters to calculate the probability of working in the formal sector (prob_form) and we proceed to the Monte Carlo experiment to select those who work in the formal sector (formal = 2). Finally, we drop parameters and temporary variables.

```
(…)

    formal=0;
    if 15<=agegr<74 and labour=1 then do;
    formal=1;
    if (cohort_region_p+cohort_p)<0 then do; cohort_region_p=0; cohort_p=0;end;
    exp_form = exp(intercept + edu_p + cohort2*cohort_p + cohort2*cohort_region_
p + region_p + young_kid_p + young_kid_edu_p);
    prob_form = exp_form/(1+exp_form);
    if rand('uniform')<prob_form then formal=2;
    end;

    drop intercept cohort_p edu_p cohort_region_p region_p young_kid_p young
_kid_edu_p exp_form prob_form cohort2
    run;
```

Now, the last population file is work.pop_formal2. In the section for cleaning the population file for the next period, we thus need to replace pop2 (which was the last population file in Chap. 3) with this.

```
/*Cleaning the population file for next period*/
data pop.pop_&endyr;
set work.pop_formal2;
(…)
```

4.4 Including the New Dimensions in the Outputs

The population file pop_&endyr (pop_2015 for the first step of the projection) now includes the projected status of the labour force and the projected sector of activity. We now need to modify the code that generates the projection outputs to include these dimensions. First, in the code generating the population by some characteristics, we add the variable “formal” to the table.

```
/*Generating outputs*/
    /*Population per age sex region education and lfp*/
    proc freq data=pop.pop_&endyr noprint;
    table year*agegr*sex*edu*region*formal/list norow nocol nopercent nocum
out=work.outputpop(rename=(count=pop) drop=percent);
    weight weight;
    run;
```

The variable for labour force participation (labour) doesn't need to be included, since it can be rebuilt from the variable "formal" (summing up those working in the formal sector and those working in the informal sector gives the active population, while the inactive have their own category).

After adding this new dimension to the outputpop table, each set of age-sex-region-education group is now divided into three categories, "inactive" (formal = 0), "working in the informal sector" (formal = 1) and "working in the formal sector" (formal = 2), as illustrated in Fig. 4.5, showing a screenshot of the file.

Before merging the population count with the components of the growth, we need to transpose the output file "outputpop" that we just created to have the variable "pop" in three columns, one for each category of the variable "formal". We use the transpose procedure for this purpose. The variable "pop" is selected following the statement var which selects the variable to transpose. The "by" statement identifies the group of variables in columns in the new dataset. We specify the variable "formal" under the "id" statement to have one column for each category of this variable. Since a column cannot have numerical label, an underscore is added, so the category 0 is labelled as _0, 1 as _1 and 2 as _2. In the options of the out statement, we rename the new columns by the name of the category, "inactive" for _0, "informal" for _1, and "formal" for _2.

	YEAR	agegr	SEX	edu	region	formal	Frequency Count
1	2015	0	0	e1	AD_rural	0	2235137.529
2	2015	5	0	e1	AD_rural	0	2038141.1602
3	2015	10	0	e1	AD_rural	0	2475631.3676
4	2015	15	0	e1	AD_rural	0	58334.388955
5	2015	15	0	e1	AD_rural	1	104373.21069
6	2015	15	0	e1	AD_rural	2	9647.7334567
7	2015	20	0	e1	AD_rural	0	53090.154745
8	2015	20	0	e1	AD_rural	1	197626.53589
9	2015	20	0	e1	AD_rural	2	36418.005977
10	2015	25	0	e1	AD_rural	0	38742.808237
11	2015	25	0	e1	AD_rural	1	418318.91639
12	2015	25	0	e1	AD_rural	2	32774.619467
13	2015	30	0	e1	AD_rural	0	30559.077893
14	2015	30	0	e1	AD_rural	1	548744.2201
15	2015	30	0	e1	AD_rural	2	68537.10778

Fig. 4.5 Screenshot of outputpop before the transpose procedure

	YEAR	agegr	SEX	edu	region	inactive	informal	formal
1	2015	0	0	e1	AD_rural	2235137.529	.	.
2	2015	5	0	e1	AD_rural	2038141.1602	.	.
3	2015	10	0	e1	AD_rural	2475631.3676	.	.
4	2015	15	0	e1	AD_rural	58334.388955	104373.21069	9647.7334567
5	2015	20	0	e1	AD_rural	53090.154745	197626.53589	36418.005977
6	2015	25	0	e1	AD_rural	38742.808237	418318.91639	32774.619467
7	2015	30	0	e1	AD_rural	30559.077893	548744.2201	68537.10778
8	2015	35	0	e1	AD_rural	10000	638533.45225	56356.786126
9	2015	40	0	e1	AD_rural	6127.869963	716163.72669	74615.186103
10	2015	45	0	e1	AD_rural	36495.024978	740170.11632	95762.205042
11	2015	50	0	e1	AD_rural	48183.353386	713048.16198	27001.847059
12	2015	55	0	e1	AD_rural	69613.374308	550573.27157	25498.987941
13	2015	60	0	e1	AD_rural	150245.30838	312995.70141	14254.91057
14	2015	65	0	e1	AD_rural	291316.30947	271673.92933	22035.514926
15	2015	70	0	e1	AD_rural	234398.90857	84293.580827	473.68696612
16	2015	75	0	e1	AD_rural	287001.29061	.	.

Fig. 4.6 Screenshot of outputpop after the transpose procedure

```
proc transpose data=work.outputpop out=work.outputpop (rename=(
_0=inactive _1=informal _2=formal)drop=_name__label_);
var pop;
by year agegr sex edu region;
id formal;
run;
```

An excerpt of the resulting dataset is shown in Fig. 4.6. Values are indeed missing in the formal and informal columns for the age group 0–14 and 75+, as by default in the modelling, they are all inactive.

The merger with the components of growth outputs can then proceed. However, the population is still split among the inactive, the informal workers, and the formal workers. In the code that creates the final output file of the period (output_&endyr), we can rebuild the total population and the active population, right after changing the missing values into 0 and the rounding of outcomes. As highlighted in yellow in the code below, the active population thus corresponds to the sum of the formal and informal workers, while the total population (pop) corresponds to the sum of the inactive and active populations.

```
/*Merging the population count and components of growth*/
    (…)
    data results.output&endyr;
    merge work.outputpop work.birth work.death work.inflow work.outflow;
    by year agegr sex edu region;

    array change _numeric_;
      do over change;
          if change=. then change=0;
          change=round(change);
      end;

    active=formal+informal;
    pop=inactive+active;
    run;
```

Up to this point, the labour force participation and the sector of activity have been projected from 2015 to 2060, but they are not included in the initial population of

2010. Since we want to be able to generate trends, we need to incorporate those variables in the initial population. Ideally, we would use real values from a survey, such as the National Sample Survey on Employment and Unemployment 2009, but the variable suffered from methodological problems in this wave and is therefore not comparable (Kapsos et al. 2014). We will thus input those variables in the initial population in a way similar to what we did for the forecasted years, using regression parameters from the National Sample Survey on Employment and Unemployment 2017/2018.

Because the initial population does not include a variable on the presence of a child in the model, we need to re-estimate the logit models without this variable. Those parameters are included in the parameter files lfp_imput.csv and formal_imput.csv, which were imported and converted already in Chap. 2. Parameters for men are exactly those used for the simulation, while those for women differ slightly because of the omission of the presence of a child at home in the model. The structure of these files is the same as those used for the simulation, with each variable having its own column with their specific categories on different rows and parameters corresponding to these categories in another column.

We impute the labour force and sector of activity to the base population of 2010 the same way we did for the simulation, with SQL commands that first merge individuals to parameters corresponding to their characteristics. For the labour force, this is done in a temporary population dataset "lfp_imput".

```
/*Imputing the labour force participation and the sector of activity for 2010*/
   /*Labour force participation*/
   /*Implementing parameters*/

    proc sql;
   create table work.lfp_imput  as
     select
       p.*,
       t1.intercept, t2.edu_p, t3.agegr_p, t4.agegr2_p, t5.agegr_edu_p,
       t6.agegr2_edu_p,t7.region_p

     from
      pop.pop_2010 p

       left join
       ( select sex, intercept
         from param.lfp_imput
         where not missing(intercept)
       ) t1
       on p.sex=t1.sex

       left join
       ( select sex, edu, edu_p
         from param.lfp_imput
         where not missing(edu_p)
       )t2
       on p.sex=t2.sex and p.edu=t2.edu

       left join
       ( select sex, agegr_p
         from param.lfp
         where not missing(agegr_p)
       )t3
       on p.sex=t3.sex
```

```
      left join
      ( select sex, agegr2_p
        from param.lfp
        where not missing(agegr2_p)
      )t4
      on p.sex=t4.sex

      left join
      ( select sex, edu, agegr_edu_p
        from param.lfp_imput
        where not missing(agegr_edu_p)
      )t5
      on p.sex=t5.sex and p.edu=t5.edu

      left join
      ( select sex, edu, agegr2_edu_p
        from param.lfp_imput
        where not missing(agegr2_edu_p)
      )t6
      on p.sex=t6.sex and p.edu=t6.edu

     left join
      ( select sex, region, region_p
        from param.lfp_imput
        where not missing(region_p)
      )t7
      on p.sex=t7.sex and p.region=t7.region;
quit;
```

From this, the imputation is then done with a random experiment in a data step creating a new population dataset “lfp_imput2”.

```
      /*Labour force participation imputation*/
   data work.lfp_imput2;
      set work.lfp_imput;

      labour=0;
      if 15<=agegr<74 then do;
      exp_lab = exp(intercept + agegr_p*agegr + agegr2_p*agegr*agegr + edu_p +
   agegr_edu_p*agegr + agegr2_edu_p*agegr*agegr
      + region_p);
      prob_lab = exp_lab/(1+exp_lab);

      if rand('uniform')<prob_lab then labour=1;
      end;

      drop intercept agegr_p edu_p agegr_edu_p agegr2_p agegr2_edu_p region_p
           exp_lab prob_lab;
      run;
```

The same is then done for the sector of activity. The resulting dataset “formal_imput2” includes the base population of 2010 with their imputed labour force participation and sector of activity.

```
/*Sector of activity*/
/*Implementing parameters*/
proc sql;
  create table work.formal_imput as
    select
      p.*,
      t1.intercept, t2.edu_p, t3.cohort_p, t4.cohort_region_p,
      t5.region_p

    from
      work.lfp_imput2 p

      left join
      ( select sex, intercept
        from param.formal_imput
        where not missing(intercept)
      ) t1
      on p.sex=t1.sex

      left join
      ( select sex, edu, edu_p
        from param.formal_imput
        where not missing(edu_p)
      )t2
      on p.sex=t2.sex and p.edu=t2.edu

      left join
      ( select sex, cohort_p
        from param.formal_imput
        where not missing(cohort_p)
      )t3
      on p.sex=t3.sex



       left join
       ( select sex, region, cohort_region_p
         from param.formal_imput
         where not missing(cohort_region_p)
       )t4
       on p.sex=t4.sex and p.region=t4.region

      left join
       ( select sex, region, region_p
         from param.formal_imput
         where not missing(region_p)
       )t5
       on p.sex=t5.sex and p.region=t5.region;
 quit;

 /*Formal - Informal imputation*/
 data work.formal_imput2;
 set work.formal_imput;

 cohort2=(cohort-1940)/5;

 formal=0;
 if 15<=agegr<74 and labour=1 then do;
 formal=1;
 if (cohort_region_p+cohort_p)<0 then do; cohort_region_p=0; cohort_p=0;end;
 exp_form = exp(intercept + edu_p + cohort2*cohort_p + cohort2*cohort_region_p
 + region_p);prob_form = exp_form/(1+exp_form);
 if rand('uniform')<prob_form then formal=2;
 end;

 drop intercept cohort_p edu_p cohort_region_p region_p exp_form prob_form cohort2;
 run;
```

Before concatenating the output files of the different periods together, we apply the same addition (highlighted in yellow) in the code of the imputed population of 2010 as we did for the simulated population, in order to have five columns for population count in the output: the total population, the active one, those working in the formal sector, those working in the informal one and those who are inactive.

```
*Population by age sex region education for 2010*/
proc freq data=pop.formal_imput2 noprint;
table  year*agegr*sex*edu*region*formal/list norow nocol nopercent nocum
out=work.output2010(rename=(count=pop) drop=percent);
weight weight;
run;


proc transpose data=work.output2010 out=work.output2010(rename=(_0=inactive
=inactive _1=informal _2=formal) drop=_name_ _label_);
var pop;
by year agegr sex edu region;
id formal;
run;


data results.output2010;
set work.output2010;

array change _numeric_;
   do over change;
      if change=. then change=0;
      change=round(change);
   end;

active=formal+informal;
pop=inactive+active;

run;
```

The final output file exported in CSV (outputTotal.csv) now includes the population by age, sex, education, region, labour force status and sector of activity.

4.5 Overview of Results

The scenario produced in this chapter assumes constant parameters for labour force participation and sector of activity. At aggregated levels, this means that any change in those dimensions comes from changes in the population composition. In Fig. 4.7, we show the projection outcomes by labour force status and sector of activity.

The total population of India is projected to grow by a bit more than 500 M from 2010 to 2060. About 40% (210 M) of this growth will be among the active population (formal + informal), which is likely to stabilize around 2045, passing from about 415 M in 2010 to 625 M. Accordingly, the labour force dependency ratio (the inactive population divided by the active one) will not change much over the next decades. According to this scenario, a small decline may first be seen as a result of the demographic dividend. The ratio will thus pass from about 1.93 in 2010 to

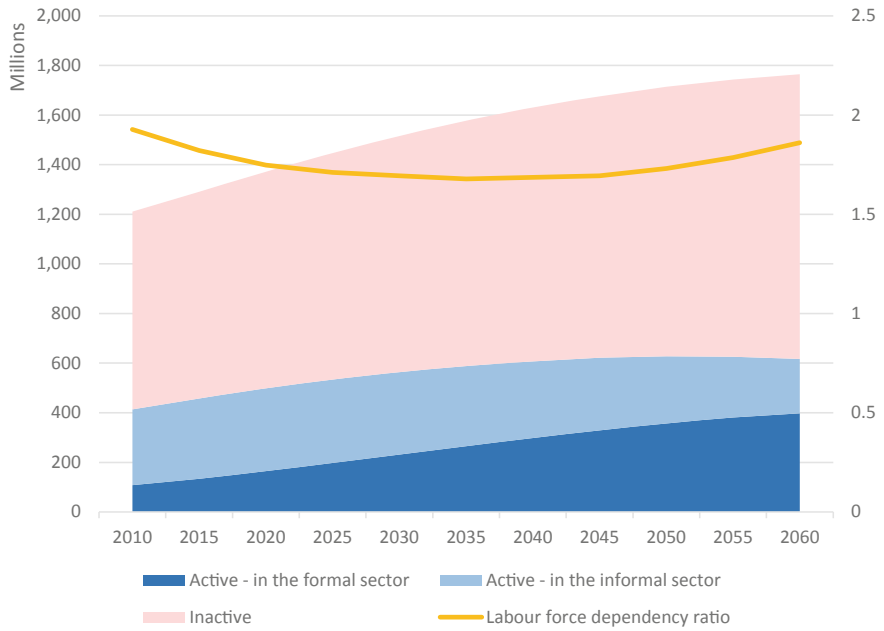

Fig. 4.7 Projected population by labour force status and sector of activity (left) and labour force dependency ratio (right), India, 2010–2060

1.68 in 2040. Because of the population ageing that will further increase the share of elderly that are inactive, the ratio will then increase slightly, reaching 1.86 in 2060.

The composition of the labour force will however change drastically. In 2010, about one quarter of workers worked in the formal sector. This proportion is likely to grow to 65% in 2060. This change is caused by a cohort effect. By demographic metabolism, the younger cohorts that are already much more likely to work in the formal sector will gradually replace the older ones.

As shown in Fig. 4.8, the education composition of the labour force is also projected to change drastically. The proportion of workers with a high school degree or above is likely to double, passing from about one-third in 2010 to about two-third in 2060. The proportion of women among workers, however, remains very low (about 20%). This is because of our assumption of constant parameters. This means that India doesn't use a large portion of its potential workforce, and therefore, the projected labour force size and the labour force dependency ratio could be much better. In the next chapter, we will build an alternative scenario showing what India might gain from greater participation of women in the labour force.

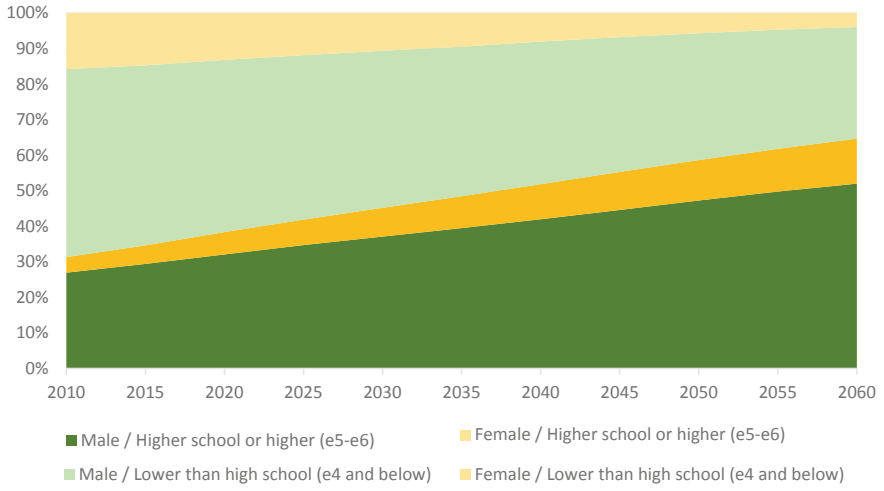

Fig. 4.8 Projected change in the sex and education composition of the labour force, India, 2010–2060

References

Cáceres-Delpiano, J. (2012). Can we still learn something from the relationship between fertility and mother's employment? Evidence from developing countries. *Demography, 49*, 151–174. https://doi.org/10.1007/s13524-011-0076-6

Kapsos, S., Silberman, A., & Bourmpoula, E. (2014). Why is female labour force participation declining so sharply in India. International Labour Office, Geneva, Switzerland.

Lutz, W. (2013). Demographic metabolism: A predictive theory of socioeconomic change. *Population and Development Review, 38*, 283–301. https://doi.org/10.1111/j.1728-4457.2013.00564.x

McCaig, B., & Pavcnik, N. (2015). Informal employment in a growing and globalizing low-income country. *American Economic Review, 105*, 545–550. https://doi.org/10.1257/aer.p20151051

Siggel, E. (2010). The Indian informal sector: The impact of globalization and reform. *International Labour Review, 149*, 93–105. https://doi.org/10.1111/j.1564-913X.2010.00077.x