**The socio-economic status component for the PROTON simulation**

5 October 2018

# 1. Relevant variables

1. Age. This attribute still does not have an operational translation, as the temporal dimension of the simulation, i.e. the relation between simulation ticks and real-world years has not been determined yet. For the time being, therefore, the temporal dimension is being described in “regular” years, as is age.

2. Sex. A number of SES distributions differ between the sexes, so keeping them separate increases model verisimilitude.

2. Wealth level (**w**). Agents initially receive their parents’ **w** when they are born. Their **w** is then updated in function of their work status **ws**. As agents can change their **ws** score across their work life, we expect **w** to be updated as a result. However, we imagine **w** to be more “sticky” than **ws**, as it represents the material and financial assets agents’ families possess at a certain point in time, w. w has five possible modalities (1 to 5, based on quintiles of the wealth distribution).

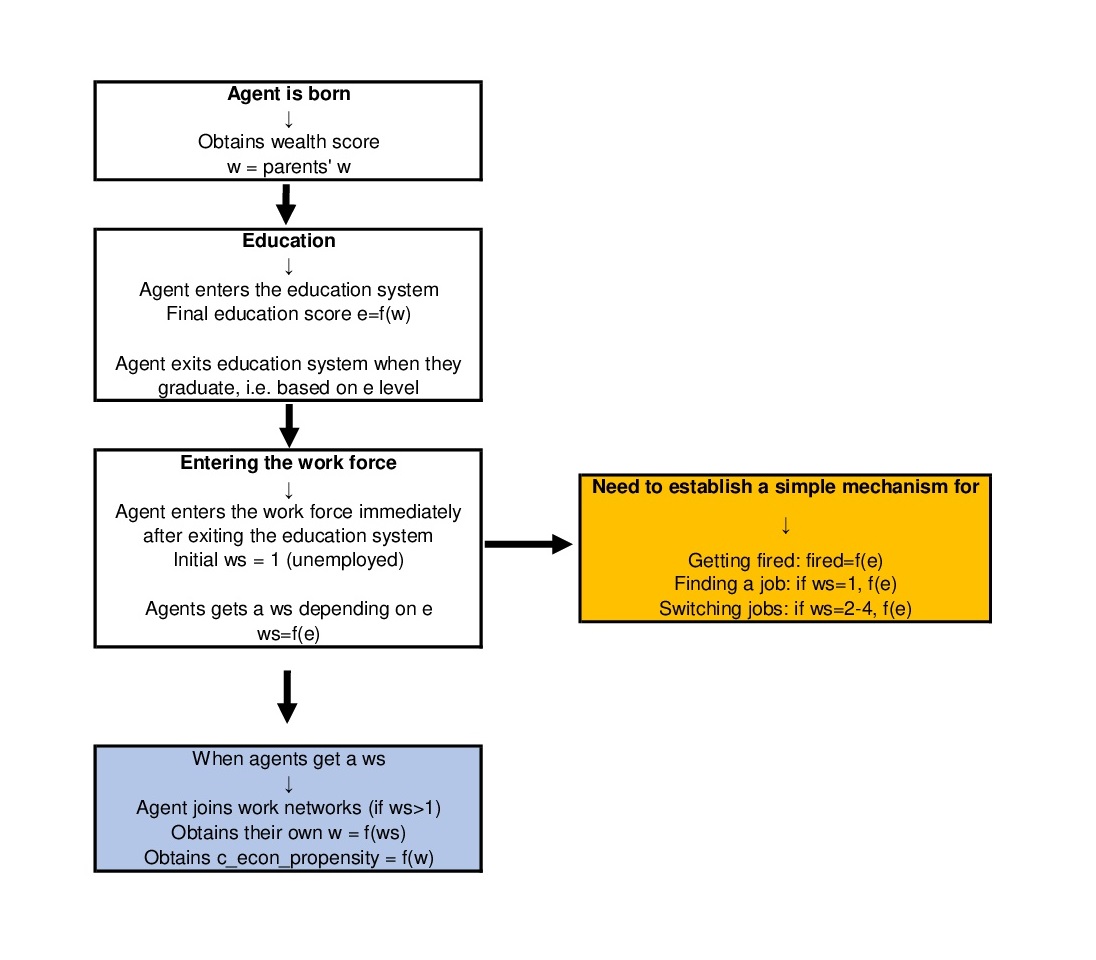
3. Education score (**e**). It is the education attainment of an agent throughout its life. The attribute has four possible modalities (1 = primary education; 2 = lower secondary education; 3 = higher secondary education; 4 = tertiary education). *Note that given current education distributions, e=1 will likely disappear as the model runs.*

5. Work status (**ws**). The agent’s actual occupational level. It also describes work positions within work networks (e.g. a network may comprise one manager position and two white collar workers). It has four modalities (1 = unemployed/inactive; 2 = blue collar worker; 3 = white collar worker; 4 = manager). When the agent enters the work force, it equals 1. It is then updated when the agent joins a work network,is fired and/or changes job.

6. SES-related criminal propensity (**c\_econ\_propensity**). This represents a general evaluation of agents’ economic stability and satisfaction. We assume less economically-stable agents to be more open to criminal activities. Note, however, that **c\_econ\_propensity** is only one of the elements of the agent’s criminal propensity score **c** (not described in this document). **c\_econ\_propensity** has four modalities. The first three (1 to 3) represent decreasing degrees of perceived economic instability; the last (4) represents perceived economic stability. **c\_econ\_propensity** is a function of the agent’s **w**, and therefore it initially depends on the agent’s parents’ economic status. When **w** is updated as a consequence of changes in **ws**, **c\_econ\_propensity** is updated as well.

# 2. Structure of the SES component

The following diagram summarizes the dynamics of the SES component:



Agents’ wealth score at birth equal their parents’. Parents’ education and wealth determines the agents’ educational attainment across their life (in accordance with the probability distributions available in the Excel annex). If e=2 the agents exits the education system at age [16-18]; e=3 the agent exits the education system at 19-20; e= the agent exits the system at age [23-27].

When agents exit the education system, they enter the workforce. They are initially unemployed (**ws** = 1). Depending of their achieved education score, agents will have different probabilities of finding jobs within a given ws categories (in accordance with to the probability distributions available in the Excel annex).

Agents stay in the workforce until age [60-65], with increasing probabilities to retire. Once retired, the agents will keep their last w and **c\_econ\_propensity**.

**4. Data and transformations**

Data for the SES component comes from Istat, for the work network structure, and from a yearly Bank of Italy survey on Italian families’ income and expenditures, for individual attributes’ distribution (available at <https://www.bancaditalia.it/pubblicazioni/indagine-famiglie/index.html>). In the next paragraphs we focus on the latter.

The BoI survey contains individual-level information on Italian responders. This includes geographic origin, age, education level, occupation status and a number of measures of income and wealth. Thanks to these we created distributions tables matching sets of two variables.

Only data from the most recent survey (2016) was selected. We limited the search to responders from Sicily (IREG = 19)[[1]](#footnote-2) and to individuals aged 16 – 65 (ETA>=16 ∧ ETA<=65). This limited our set to n = 858 responders. We excluded the retired in the age bracket (QUAL = 6, “retired”, n = 71). The resulting distribution is:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **QUAL = 1**  **“blue-collar workers”** | **QUAL = 2**  **“white-collar workers”** | **QUAL = 3**  **“management”** | **QUAL = 4**  **“entrepreneur /private practitioner”** | **QUAL = 5**  **“other self-employed”** | **QUAL = 7**  **“other unemployed”** | **TOT** |
| n = 132 (.168) | n = 126 (.160) | n = 18 (.023) | n = 29 (.037) | n = 39 (.050) | n = 443 (.563) | **n = 787 (1.000)** |

At this point we created a set of Excel sheets (2A to 2D) with all the measures from the survey necessary for the SES component. They contain the following variables:

|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Modalities** |
| **nquest** | Questionnaire number (one for each family) | [six-digit code] |
| **nord** | Individual interviewed within each family | [one-digit code; range = 1 – 8] |
| **NCOMP** | Number of members in the family | [1-8] |
| **sex** | Sex of responder | 1 = man2= woman |
| **studio** | Education level of responder | 1 = “no education”; 2 = “completed primary education”; 3 = “completed lower secondary education”; 4 = “completed 3-year professional diploma [normally completed at 16 - 17]”; 5 = “graduated high school”; 6 = “university bachelor’s degree”; 7 = “university master’s degree”; 8 = “post-degree specialization” |
| **ETA** | Age of responder | 16 - 65 |
| **QUAL** | Employment status of responder | 1 = “blue-collar worker”; 2= “white-collar worker” ; 3 = “manager” ; 4 = “entrepreneur / private practitioner” ; 5 = “other self-employed” ; 6 = “retired” [excluded from the sample]; 7 = “other unemployed” |
| **W** | Family wealth | [euros] |
| **CLW2** | Wealth quintile the family belongs to among responder families | [1-5] |
| **condgen** | Family head’s perception of own family’s economic stability (same for each family member) | "Your family income allows your family to reach the end of the month:"  1 = "with significant difficulties"; 2 = "wiith difficulties"; 3 = "with some difficulties"; 4 = "fairly easily"; 5 = "easily"; 6 = "very easily" |

These variables were adapted to become model attributes, thus creating the final dataset used to model the variable distributions (sheet “3. Data\_clean”, n = 787)[[2]](#footnote-3).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **nquest+nord** | **sex1** | **e** | **age** | **ws** | **W/NCOMP** | **CLW2** | **c\_econ\_propensity** |
| **Description** | Agent’s ID | Agent’s sex | Agent’s education level | In years | Agent’s work status | Agent’s individual wealth, i.e. family wealth divided by number of family members | quintile the responder’s family belongs to in the survey distribution | Individual’s criminal propensity due to economic factors |
| **Transformations** | nquest + nord | M <- sex=1  F <- sex=2 | 1 <- studio = 1, 2  2 <- studio = 3  3 <- studio = 4, 5  4 <- studio = 6, 7, 8 | Age = ETA | 1 <- QUAL = 7  2 <- QUAL = 1, 5  3 <- QUAL = 2  4 <- QUAL = 3, 4 | W/NCOMP |  | 1 <- condgen = 1  2 <- condgen = 2  3 <- condgen = 3  4 <- condgen = 4, 5, 6 |
| **Modalities** | [seven-digit code] | M = man  F = woman | 1 = primary education  2 = lower secondary education  3 = higher secondary education (incl. 3-year professional diploma  4 = tertiary education | 16 - 65 | 1 = unemployed  2 = blue-collar worker  3 = white-collar worker  4 = manager | [euros] | 1 - 5 | "Your family income allows your family to reach the end of the month:"  1 = "with significant difficulties"  2 = "wiith difficulties"  3 = "with some difficulties"  4 = "fairly easily", "easily" or "very easily" |

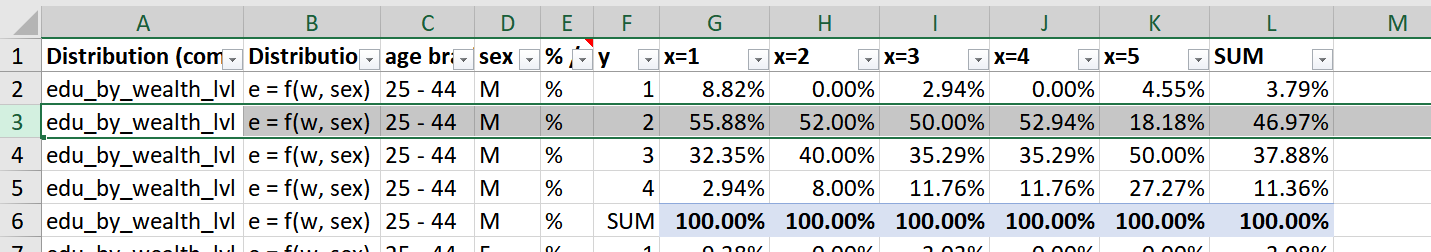
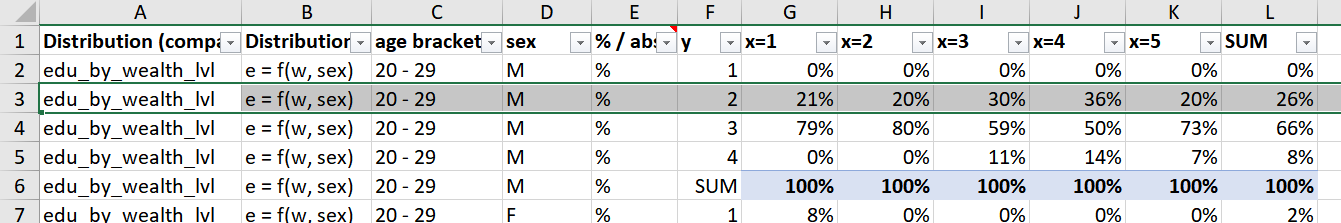
# 3. SES-related probabilities

The values of the relevant variables will be distributed according to the observed probabilities in the BoI dataset. We created the distribution matrix sheet (“4. Distribution\_matrices). The structure is the following:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A | B | C | D | E | F | G | H | I | J | K | L |
| Distribution (compact name) | Distribution (function) | age bracket | sex | % / abs\_val | y | x=1 | x=2 | x=3 | x=4 | x=5 | SUM |

Columns A and B describe the distribution of interest (e.g. education level by family wealth), both in alphanumeric and mathematical terms. Column C and D describe respectively the age bracket and agent sex(es) the distribution refers to. Column E tells whether the distribution as contained in columns G-L is in absolute terms, i.e. whether it refers to the absolute number of survey responders in the category, or to their column percentage, i.e. the fraction of all responders in the same age, sex and independent variable categories. Column F contains the output variable modality the distribution in columns G-L refers to (including the sum of all modalities). Columns G-K contain modalities 1 to 5 of the independent variable in the output variable category described by columns C, D and F. Column L contains the sum of the content of columns G-K.

We can better illustrate the previous paragraph with an example:



Let’s focus on row 3. First of all, the distribution of interest is education level by family wealth, as shown in columns A and B. That is, the independent variable x is wealth, while the output variable y is education level. The distribution refers to male responders aged 25-44, as per columns C and D. The distribution is displayed in percentages, as shown in column E. Column F tells us we are looking at the distribution of responders whose education level y equals 2. Columns G to L tell us the wealth distribution of responders whose y=2. For instance, looking at cell J3 we can say about 53% of the male survey responders aged 25-44 who belong to a family in the fourth wealth quintile have a lower secondary education.

From a programming standpoint, it is useful to look at the table by independent variable (x) categories. In this example, taking cells J2 to J6 (all of which refer to males aged 25-44) we know male agents in the simulation who belong to the fourth wealth quintile will have a probability of (roughly) 53% of reaching an education level =2, a (roughly) 35% of reaching an education level = 3 and a (roughly) 12% probability of reaching an education level = 4. In other terms, about half of fourth-quintile males will finish middle school, about a third will graduate from high school and those who are left will get a university degree.

1. We were able to select only residents of the Palermo municipality, but that created a responder sample too small for accurate distributions. [↑](#footnote-ref-2)
2. Note that we incorporated categories QUAL = 4 (“entrepreneur/private practitioner”, n = 29) and QUAL = 5 (“other self-employed”, including craftsmen, owners or partners of family businesses and "sporadic workers", n = 39) respectively into “manager” and “blue-collar” work status categories respectively. This is because QUAL = 4 includes independent contractors, freelancers, private practitioners, individual entrepreneurs and company shareholders, all of whom we assume to have manager characteristics. QUAL = 5 includes craftsmen, owners or partners of family businesses and "sporadic workers", all of whom we assume to have blue-worker characteristics. [↑](#footnote-ref-3)