Adjusting the ASEC Tax Filing Status Variable

Johannes Fleck*

June 14, 2020

1 Tax Filing Status in the ASEC Dataset

As explained in the IPUMS documentation[1] the CPS ASEC variable FILESTAT reports the federal income tax filing status. It provides six different tax filing status categories:

Code	Label
0	No data
1	Joint, both less than 65
2	Joint, one less than 65 and one 65+
3	Joint, both 65+
4	Head of household
5	Single
6	Nonfiler

Note that FILESTAT never assumes the "No data" category so its values are strictly positive.

1.1 Time Comparability

As for other ASEC tax-related variables, values of FILESTAT are imputed by the Census Bureau's tax model, i.e. they are not provided by survey respondents. The IPUMS documentation mentions that comparability of FILESTAT might have been affected by the introduction of a new Census Bureau tax model in 2004.

Figure 1 shows the distributions of tax filer categories for 2004 and adjacent years. As the plots illustrate, the share of nonfilers appears to be much larger in 2004 and 2005. The reverse applies to the share of joint filers below 65 while the shares of head of household and single filers seem comparable across years.

To provide more details on these discrepancies, table 1 displays the relative frequencies of the FILESTAT categories over a longer time period. Comparing the numbers of 2004 and 2005 with other years corroborates the earlier impression – the discrepancies affect the joint filer and nonfiler categories only. Moreover, compared to other years, the share of non filers is about 20 percentage points larger while the share of all joint filers is about 20 percentage points lower. Lastly, the shares of singles and nonfilers are very similar to those of other years.

FILESTAT	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
1	33.43	33.84	33.66	33.63	**16.64**	**16.49**	34.78	35.41	35.21	35.17	34.14	34.02	33.4	33.15	32.98	32.67
2	1.89	1.68	1.67	1.6	**0.73**	**0.73**	0.84	0.88	0.97	1.02	0.97	1.01	1.02	1.06	1.12	1.13
3	4.21	3.08	3.05	3.03	**0.98**	**0.98**	1.54	1.6	1.62	1.61	1.63	1.7	1.83	1.9	2.04	2.05
4	4.28	4.76	4.83	4.87	4.56	4.59	4.61	4.24	4.25	4.16	4.21	4.19	4.24	4.14	4.11	4.09
5	20.59	18.43	18.12	17.69	17.09	16.83	18.1	18.27	18.72	18.44	18.5	18.3	18.43	18.79	18.86	19.06
6	35.6	38.2	38.66	39.19	**60.0**	**60.38**	40.12	39.6	39.23	39.6	40.54	40.78	41.07	40.96	40.89	40.99

Table 1: Relative frequencies of FILESTAT in selected years (in %)

*This report describes the algorithm I developed to adjust the ASEC variable FILESTAT for the years 2004 and 2005. I welcome comments and questions at Johannes.Fleck@eui.eu. Julia code generating the figures and tables shown in this report is available here: https://github.com/Jo-Fleck/ASEC_FILESTAT_adjustment

[1] https://cps.ipums.org/cps-action/variables/FILESTAT#description_section

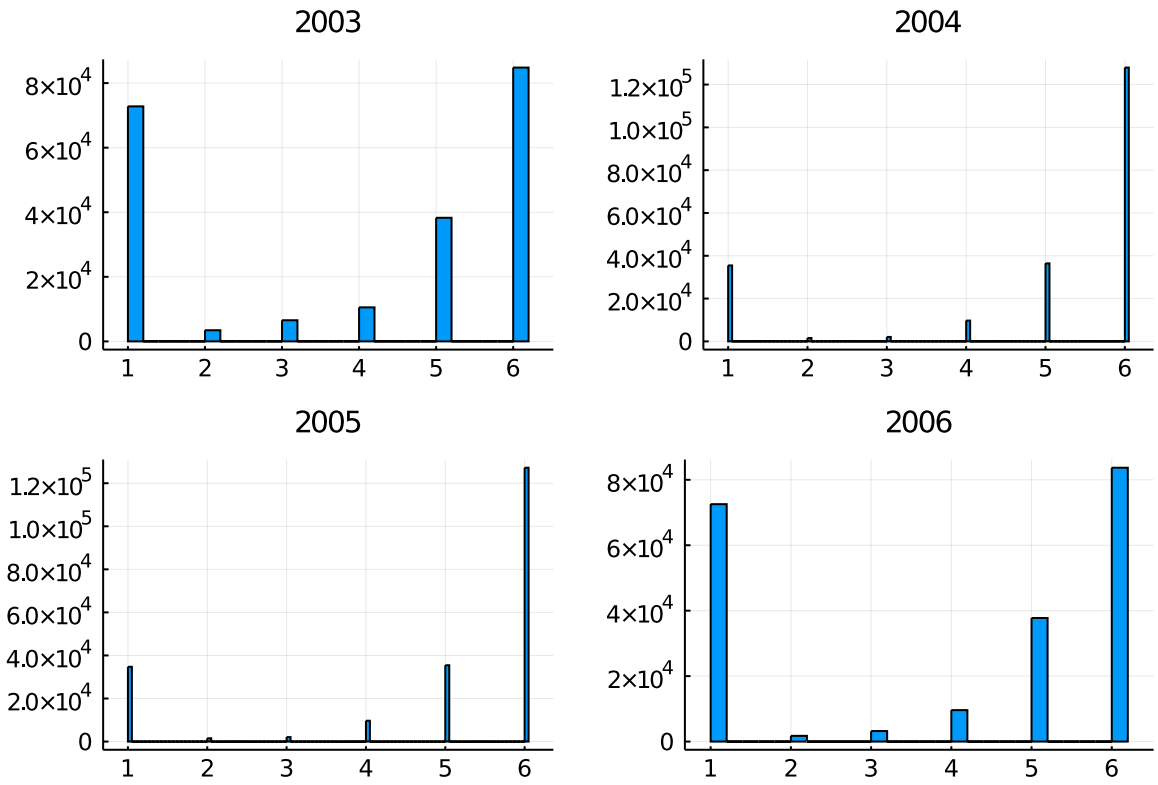

Figure 1: Histograms of FILESTAT in years close to the introduction of the new tax model

2 The Adjustment Algorithm

In this section, I first provide a brief description of the algorithm I developed to adjust the FILESTAT variable for 2004 and 2005 and show its implementation in Julia (for 2004). Next, I demonstrate that it reproduces the values of other years remarkably well. Finally, I apply the algorithm to 2004 and 2005 and present the resulting adjusted FILESTAT distributions.

2.1 Description

Based on the distributions of FILESTAT values in different years presented above, I formed the following hypothesis; in years other than 2004 and 2005, the imputation procedure generating FILESTAT values assigned persons who filed jointly the same category. In 2004 and 2005, however, it assigned one of them the joint filer status and nonfiler status to the other. My adjustment algorithm corrects the discrepancies on the basis of this hypothesis. Its structure is as follows:

- Does a household qualify as a joint filer?
  - If no: keep the original value of FILESTAT.
  - If yes:
    - Check the age of the household head and the spouse and assign both of them the applicable joint filer status.
      - If both have zero adjusted gross income: assign non filer status to both.
    - For the remaining household members, keep the original FILESTAT value.

2.2 JULIA IMPLEMENTATION FOR 2004

The Julia file `FILESTAT_adj_2004.jl` printed below is available in this repository:
`https://github.com/Jo-Fleck/ASEC_FILESTAT_adjustment`

```
# Prepare 2004 data
df_2004 = select!(df_ASEC_2004,[:SERIAL, :RELATE, :AGE, :ADJGINC, :FILESTAT, :FILESTAT_adj]);
df_2004[!, :num] = 1:(size(df_2004,1));
hhs_2004 = unique(df_2004.SERIAL);

for k in hhs_2004

    df_tmp = df_2004[df_2004.SERIAL .== k, :]
    RELATE_vec = unique(df_tmp.RELATE)

    if ~(201 in RELATE_vec)
         continue                                    # keep FILESTAT categories as they are
    else

        num_vec = unique(df_tmp.num)

        age_101 = df_tmp[df_tmp.RELATE .== 101, :AGE][1]
        age_201 = df_tmp[df_tmp.RELATE .== 201, :AGE][1]
        if age_101 < 65 && age_201 < 65              # Both below 65
            df_2004[num_vec[1], :FILESTAT_adj] = 1
            df_2004[num_vec[2], :FILESTAT_adj] = 1
        elseif age_101 >= 65 && age_201 >= 65        # Both above 65
            df_2004[num_vec[1], :FILESTAT_adj] = 3
            df_2004[num_vec[2], :FILESTAT_adj] = 3
        else                                         # One above, one below 65
            df_2004[num_vec[1], :FILESTAT_adj] = 2
            df_2004[num_vec[2], :FILESTAT_adj] = 2
        end

        # hhs with agi income == 0 do not need to file
        adjginc_101 = df_tmp[df_tmp.RELATE .== 101, :ADJGINC][1]
        adjginc_201 = df_tmp[df_tmp.RELATE .== 201, :ADJGINC][1]
        if adjginc_101 == 0 && adjginc_201 == 0
            df_2004[num_vec[1], :FILESTAT_adj] = 6
            df_2004[num_vec[2], :FILESTAT_adj] = 6
        end

        # keep FILESTAT categories as they are for remaining hh members
        if length(num_vec) > 2
            for l = 3:length(num_vec)
                df_2004[num_vec[l], :FILESTAT_adj] = df_2004[num_vec[l], :FILESTAT]
            end
        end
    end
end
```

2.3 REPLICATING FILESTAT

Table 2 shows the number of observations (in % of total) for which the adjustment algorithm produces the same values as the original `FILESTAT` variable. The time variation indicates there was a structural change in the imputation of `FILESTAT` in the years 2004 and 2005; the algorithm produces (almost)[2] identical values prior to 2004 and a high rate of identical values after 2005.

2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
100.0	100.0	100.0	100.0	82.04	82.18	97.48	97.29	97.36	97.33	97.36	97.23	97.03	96.91	96.72	96.72

Table 2: Observations classified identically as `FILESTAT` by the adjustment algorithm (in %)

[2] Before rounding, the percentages are 99.99%.

2.4 Adjusted FILESTAT for 2004 and 2005

Table 3 displays the relative frequencies of the adjusted and unadjusted FILESTAT categories for 2004 and 2005 (in blue and red, respectively). The adjusted values line up well with the unadjusted ones of other years. This impression is also supported by figure 2. Hence, the adjustment algorithm helps to reduce the FILESTAT discrepancies in 2004 and 2005.

FILESTAT	2003	2004	2004	2005	2005	2006
1	33.63	**16.64**	**32.92**	**16.49**	**32.63**	34.78
2	1.6	**0.73**	**1.45**	**0.73**	**1.44**	0.84
3	3.03	**0.98**	**1.94**	**0.98**	**1.95**	1.54
4	4.87	**4.56**	**4.56**	**4.59**	**4.59**	4.61
5	17.69	**17.09**	**17.09**	**16.83**	**16.83**	18.1
6	39.19	**60.0**	**42.04**	**60.38**	**42.57**	40.12

Table 3: Relative frequencies of FILESTAT original (red) and adjusted (blue) in 2004 and 2005 (in %)

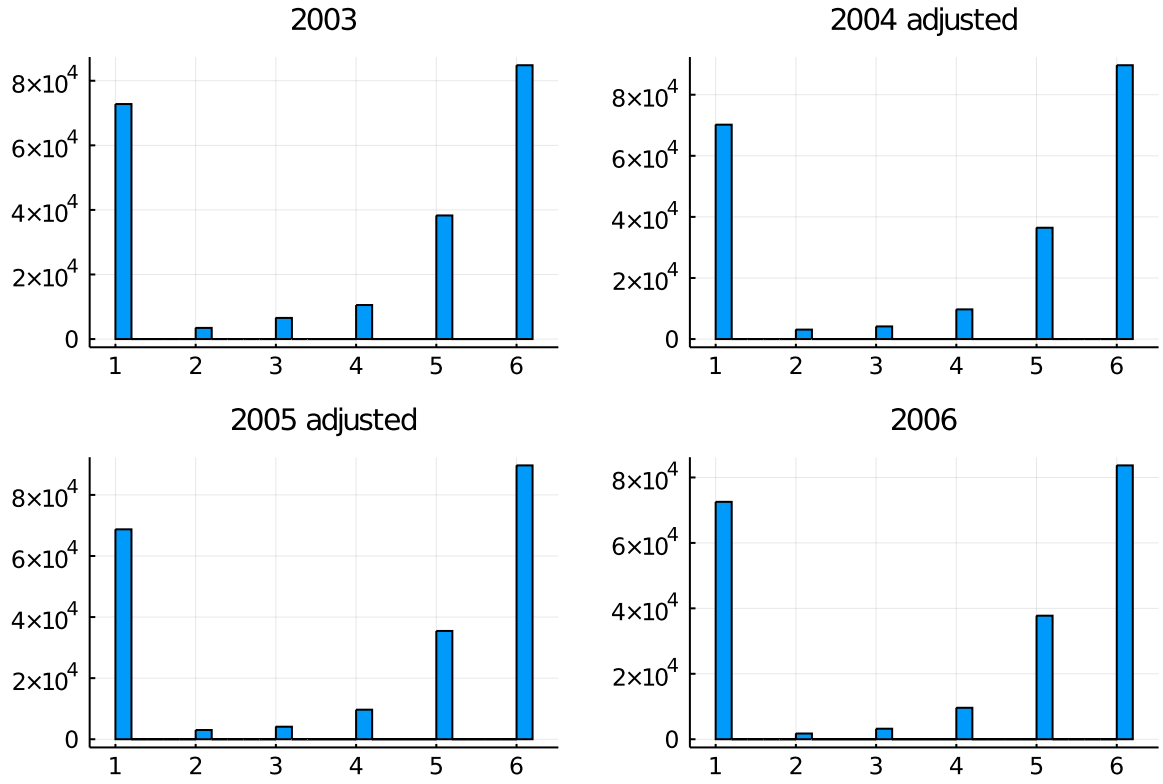

Figure 2: Histograms of adjusted FILESTAT in years close to the introduction of the new tax model