Migrating historical AEA supplements - DRAFT

Since July 16, 2019, the American Economic Association has used the **AEA Data and Code Repository** at **openICPSR** as the default archive for its supplements. This archive serves a dual purpose: to share data with the AEA Data Editor prior to being published, and as a publication outlet for supplements to articles in AEA journals.

At the time, the AEA also announced that it would migrate the historical supplements, hitherto stored as ZIP files on the AEA website, into the AEA Data and Code Repository.

On Oct 1, 2019, openICPSR had 867 deposits, which covered 94 deposits in the DataLumos archive, 46 in the AERA archive, and 13 in the PSID archive. The **AEA Data and Code Repository** contained at the time 93 deposits, of which 5 were public, the others awaiting publication of the associated article.

Between Oct 11 and Oct 13, 2019, the staff at openICPSR ingested 2,552 historical supplements, increasing the size of the openICPSR repository **by a factor of 3**, to 3,461. This was only the first part of the migration, as there are about 1,000 more archives that need to be migrated.

Increased findability

The migrated archives are now available through the openICPSR search interface, the general ICPSR search interface, as well as through a variety of federated search interfaces such as Google Dataset Search. For instance, the current AER Editor's supplements can be found here, here and here, with increasing generality.

Characteristics of AEA supplement data

We can describe this subset of the historical supplements in a variety of ways.

Time coverage

This is only a subset of all supplements, so what years are covered?

year	count
1999	2
2000	2
2004	24
2005	280
2006	7
2007	625
2008	1010
2009	3918
2010	6063
2011	6198
2012	7394
2013	6923
2014	11813
2015	7511
2016	10037
2017	11323
2018	13574
2019	6147
NA	1865

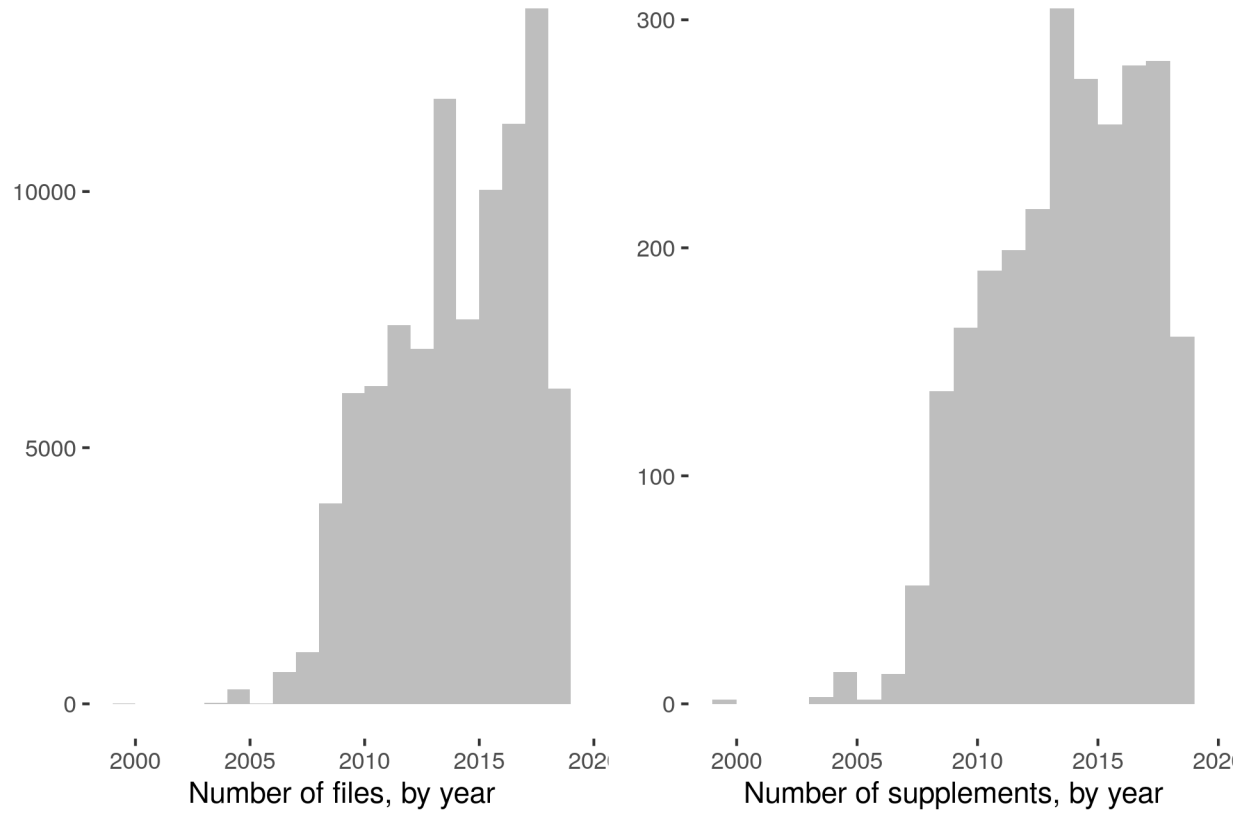

Figure 1: Distribution across years

Number of files per supplement and size of supplement

doi	size	count
10.1257/pandp.20181045	32782817227	795
10.1257/pol.20150168	27251086584	691
10.1257/app.20170080	19924915397	236
10.1257/aer.20131496	15688843069	186
10.1257/app.20160510	14419727617	465
10.1257/app.6.3.206	12190878093	45
10.1257/aer.102.2.994	10789428265	36
10.1257/aer.20121662	10334181612	622
10.1257/mic.20130164	8042445763	20
10.1257/aer.20141374	6804364827	140

The 2,552 supplements contain a total of 94,465 files - programs, documents, datasets. The largest supplement within this group in terms of file count has 795 files, summing to 30.5 Gb (Armour, Button, and Hollands, 2018). Note however that among the remaining non-migrated supplements are very large packages: the largest we have identified has 201,972 files.

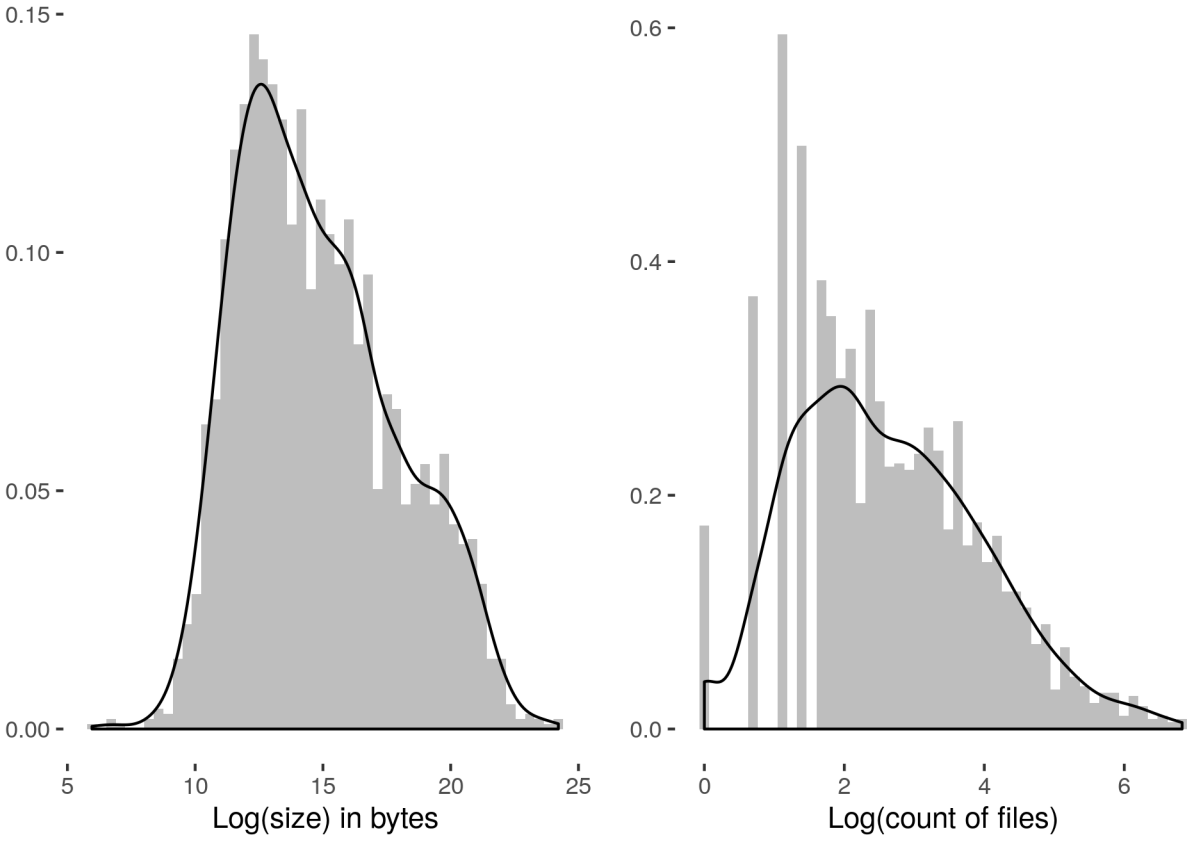

Figure 2: Distribution of filesizes and filecounts

Distribution overall

Stats by journal

We can look at the size of the supplements globally by journal. The following table shows cumulative and median size and number of files.

Journal	Articles	Median Size (Mb)	Cumulative Size (Mb)	Median no. of
American Economic Review	1,238	1.4	165,552.8	
American Economic Journal: Applied Economics	363	3.1	101,969.8	
American Economic Journal: Economic Policy	351	4.2	98,574.0	
AEA Papers and Proceedings	109	0.9	44,497.3	
American Economic Journal: Macroeconomics	259	1.4	14,943.5	
American Economic Journal: Microeconomics	115	1.1	12,899.3	
Journal of Economic Perspectives	115	0.8	12,755.7	
Journal of Economic Literature	7	17.8	432.1	
American Economic Review: Insights	3	8.2	74.7	

Distribution across JEL codes

The top 10 JEL codes associated with supplements are:

Number of packages	Pct	JEL	Description
263	10.31	E32	Business Fluctuations; Cycles
245	9.60	J24	Human Capital; Skills; Occupational Choice; Labor Productivity
217	8.50	O15	Economic Development: Human Resources; Human Development; Income Distributi
214	8.39	D12	Consumer Economics: Empirical Analysis
207	8.11	J31	Wage Level and Structure; Wage Differentials
191	7.48	J16	Economics of Gender; Non-labor Discrimination
183	7.17	J13	Fertility; Family Planning; Child Care; Children; Youth
176	6.90	I21	Analysis of Education
162	6.35	D72	Political Processes: Rent-seeking, Lobbying, Elections, Legislatures, and Voting Beh
157	6.15	D83	Search; Learning; Information and Knowledge; Communication; Belief

Note:
A supplement can be associated with multiple JEL codes.

Software used

To identify software usage and data formats, we (manually) mapped file extensions into known software packages, and classified the file type into a set of categories:

File type	Number of extensions
Program	66
Document	36
Data	26
Junk	14
Archive	7
Unknown	3
Logfile	2

The table below shows the top ten software, by frequency of program files:

Number of files	Pct	Software
19676	48.37	Stata
15244	37.47	Matlab
1750	4.30	Fortran
1084	2.66	SAS
659	1.62	R
454	1.12	General
416	1.02	C
291	0.72	Unknown
258	0.63	None
207	0.51	Python

The top software with respect to number of files is **Stata**. Note that there are 258 supplements that do not contain files that we have identified as program files ("None").

More interesting is how many supplements use one or more software:

Number of Software	N	Percent
1	2058	80.64
2	397	15.56
3+	105	4.11

with a maximum of 7 different software packages used in any one of the supplements. In turn, the number of supplements in which software is used at least once is reflected in the next table (restricted to at least 10 mentions):

Name of Software	Usages	Percent
Stata	1862	72.96
Matlab	573	22.45
None	258	10.11
SAS	111	4.35
R	97	3.80
Fortran	64	2.51
Python	54	2.12
Unknown	37	1.45
C	34	1.33
General	29	1.14
Shell	29	1.14
Windows	24	0.94

Note:
Percentage sum to more than 100 percent, since a supplement can use multiple software packages.

Clearly, **Stata** is the most popular statistical software in the journals of the AEA, followed by **Matlab**. Note again the 258 supplements that do not contain files that we have identified as program files ("None").

Data formats

It is somewhat more ambiguous identifying data files, as they come in a large variety of formats. Furthermore, data might be compressed. In the following table, we tabulate data files and archives, by the software package associated with their extension. The data type "General" encompasses formats like "tsv" or "csv" that are

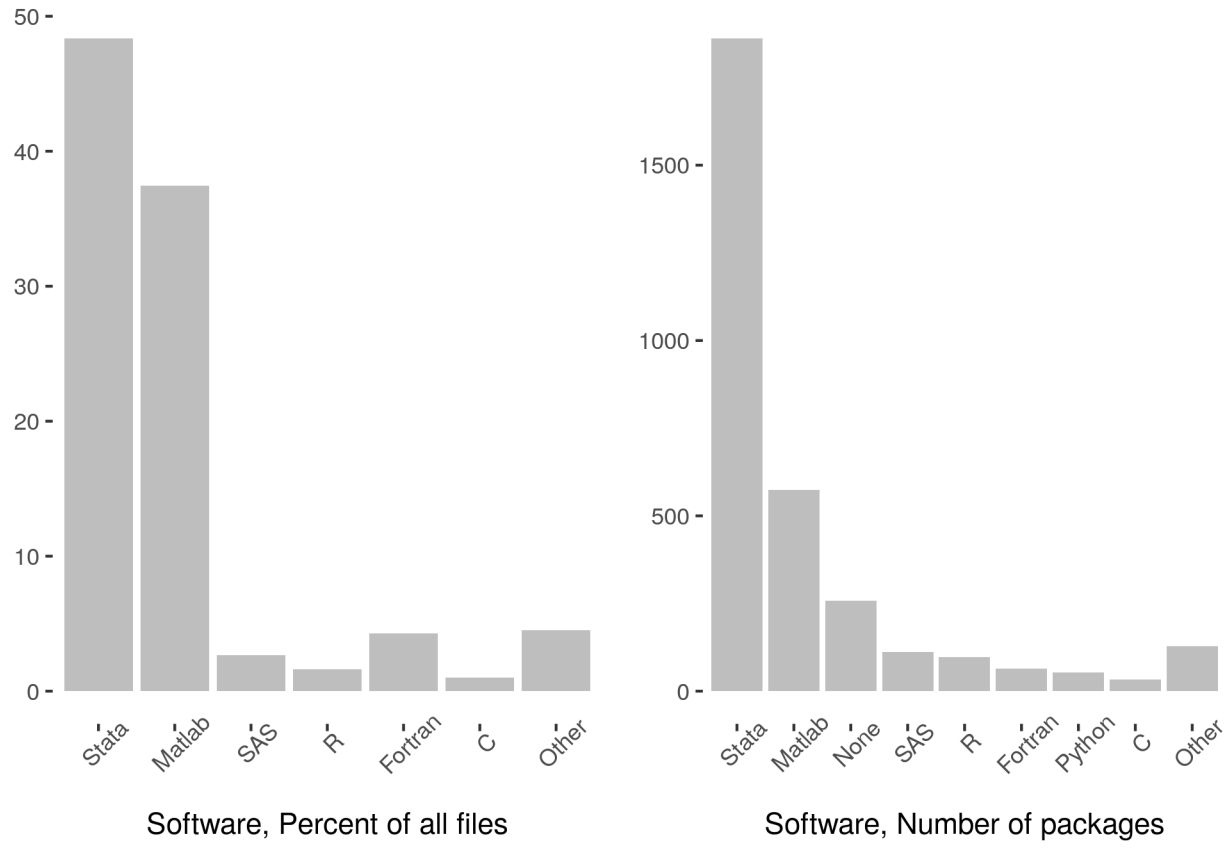

Figure 3: Software usage

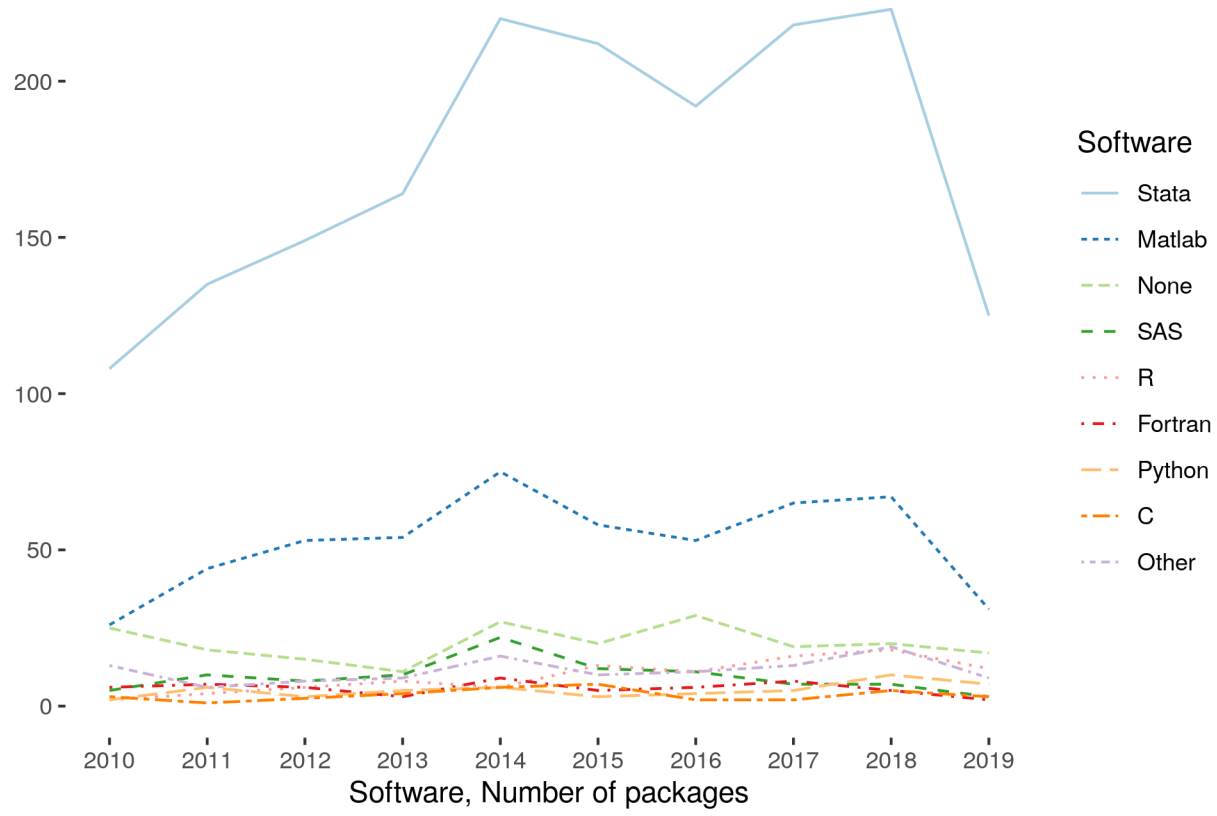

Figure 4: Software usage over time, number of supplements

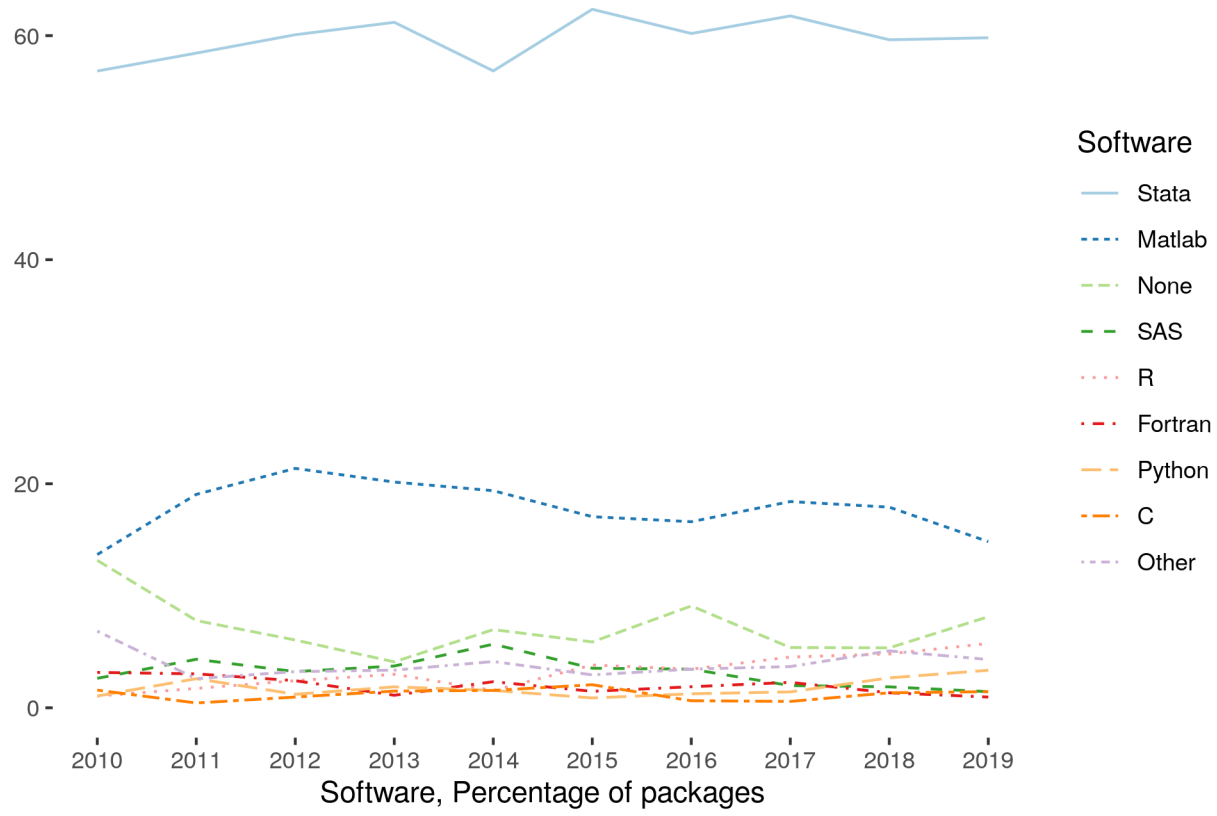

Figure 5: Software usage over time, in percent

not associated with any particular software, but are nevertheless clearly identifiable as data files (full list available). We restrict ourselves to the number of supplements which contain files with such extensions.

Name of Software	Usages	Percent
Stata	1392	44.90
Excel	682	22.00
General	547	17.65
Matlab	242	7.81
Archive	179	5.77
SAS	38	1.23
R	9	0.29
OpenOffice	5	0.16
Unknown	3	0.10
SPSS	2	0.06
Julia	1	0.03

Note:
Percentage sum to more than 100 percent, since a supplement can use multiple software packages.

Metadata

When planning the migration, the preservation of existing metadata - the information about the data and code - was important. The AEA Data Editor worked with the openICPSR staff to enhance the data infrastructure, adding the capability to store and display JEL codes in addition to subject terms. Going forward, in addition to adding the JEL codes that also describe the linked article, authors can add metadata such as *geographic coverage*, *funding sources*, *time periods*, *geographic units* as well as *units of observation*, greatly enhancing the ability of researchers to find data through the openICPSR search interface.

Two important caveats apply, however. First, none of the additional metadata exists for the historical archives. Second, the openICPSR search interface only allows to search for these in an implicit way, i.e., one can search for "J31" because it is unlikely to appear as anything else, but there is no selection by specific JEL codes currently possible. The ability to do so is planned for a later implementation.

Data Availability

The input data to this paper are available at (OPENICPSR TBD). The tables presented in this paper, and the data underlying the figures, are available at https://github.com/AEADataEditor/aea-supplement-migration/data/generated.

Code Availability

The code underlying this analysis can be downloaded at https://github.com/AEADataEditor/aea-supplement-migration.

References

Armour, Phillip, Patrick Button, and Simon Hollands. 2018. "Disability Saliency and Discrimination in Hiring." *AEA Papers and Proceedings* 108. American Economic Association: 262–66. https://doi.org/10.1257/pandp.20181045.