

Volume 7: DAF18 Development History and Construction Methods

July 2020

Submitted to:

Social Security Administration
Office of Retirement and Disability Policy
Office of Research, Demonstration, and Employment Support
Washington, DC 20024-2796
Project Officers: Paul O’Leary and Debra Tidwell-Peters
Contract Number: SS00-16-60003

Submitted by:

Mathematica
1100 1st Street, NE
12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763
Project Director: Jody Schimmel Hyde
Reference Number: 50214.Y3.T05.530.360

Suggested Citation: “Disability Analysis File 2018 (DAF18) Documentation: Data from January 1994 through December 2018.” Washington, DC: Mathematica, July 2020.

This page has been left blank for double-sided copying.

CONTENTS

GLOSSARY	vii
OVERVIEW OF DAF DOCUMENTATION.....	xi
I. THE EVOLUTION OF THE DAF OVER TIME.....	1
II. OVERVIEW OF DAF CONSTRUCTION TASKS	5
A. Task 1: Assemble and combine DBAD files	5
B. Task 2: Assemble and combine CER.....	6
C. Task 3: Create finder files.....	6
D. Task 4: Submit finder files	7
E. Task 5: Process 831 & 832/833 data	8
F. Task 6: NUMIDENT processing	8
G. Task 7: Process SSI-LF data.....	9
H. Task 8: Process MBR data.....	9
I. Task 9: DMG Pre-processing	10
J. Task 10: Create DAF.DMG component	11
K. Task 11: Create DAF.Ticket component.....	11
L. Task 12: Annuals Pre-processing	13
M. Task 13: Create DAF.Annual data	14
N. Task 14: Building STWs and BFWs	14
O. Task 15: Create payments component (EN payments data)	15
P. Task 16: Create payments component (VRRMS data).....	15
Q. Task 17: Create DAF-RSA files.....	17
R. Task 18: Create LAUS/SAIPE formats.....	17
III. CHANGES IN BENEFICIARY SELECTION CRITERIA ACROSS DAF VERSIONS.....	19
IV. CHANGES IN THE DAF SOURCE DATA OVER TIME	21
V. LEGACY PROCESSING ERROR AFFECTING SSI EARNINGS AND BENEFITS DUE DISCOVERED AND CORRECTED IN DAF17	23
A. Description of the error	23
B. Variables affected by the error	25
C. Implications of the error for snapshot statistics	26
D. Magnitude of the effect of the error	27

1. Directly-affected variables	27
2. Indirectly-affected variables	31
E. Characteristics of beneficiaries with affected records	37
F. Discussion	44
VI. ASSESSING DATA NEEDS FOR DAF18	45
A. Overview of the process	45
B. Process for assessing data needs	45
1. Review ad hoc requests	45
2. Survey DAF users	45
3. Timeline	46
4. Prepare recommendations for SSA	47
APPENDIX A MONTHLY GRAPHS OF MEASURES OF CHANGE DUE TO THE SSI REPROCESSING ERROR IDENTIFIED AND CORRECTED IN DAF17	A.1

TABLES

VII.1.	Example of combining SSI-LF records in the old (incorrect) and new (corrected) DAF processing algorithm	25
VII.2.	Percentage distribution of STWSSI values before and after correcting the SSI-LF processing error	36
VII.3.	Comparison of beneficiary characteristics for those whose FAMT records were or were not affected by the SSI-LF processing error	40
VII.4.	Comparison of beneficiary characteristics for those whose EICM records were or were not affected by the SSI-LF processing error	42

FIGURES

VII.1.	Percentage increase in the number of beneficiaries with a positive value after correcting the SSI-LF processing error, as a share of beneficiaries with a positive value before correcting the error	28
VII.2.	Percentage-point increase in the share of SSI beneficiaries with positive values after correcting the SSI-LF processing error	29
VII.3.	Percentage change in the average dollar value among observations with positive values after correcting the SSI-LF processing error	31
VII.4.	Percentage increase in the number of beneficiaries with value equal to 1 after the SSI-LF processing error was corrected in DAF17, as a share of beneficiaries with a positive value before the error was corrected	32
VII.5.	Percentage-point increase in the share of non-terminated SSI beneficiaries with value equal to 1 after correcting the SSI-LF processing error	33
VII.6.	Percentage increase in the number of beneficiaries with positive BFWSSI after correcting the SSI-LF processing error	37
VII.7.	Percentage change in the average dollar value of BFWSSI among positive observations after correcting the SSI-LF processing error	37
A.1.	EICMyymm: Number of beneficiaries with value > 0	A.3
A.2.	EICMyymm: Percent of SSI beneficiaries with value > 0	A.4
A.3.	EICMyymm: Average value among beneficiaries with value > 0	A.5
A.4.	UINCyymm: Number of beneficiaries with value > 0	A.6
A.5.	UINCyymm: Percent of SSI beneficiaries with value > 0	A.7
A.6.	UINCyymm: Average value among beneficiaries with value > 0	A.8
A.7.	FAMTyymm: Number of beneficiaries with value > 0	A.9
A.8.	FAMTyymm: Percent of SSI beneficiaries with value > 0	A.10
A.9.	FAMTyymm: Average value among beneficiaries with value > 0	A.11

A.10.	SAMTyymm: Number of beneficiaries with value > 0	A.12
A.11.	SAMTyymm: Percent of SSI beneficiaries with value > 0	A.13
A.12.	SAMTyymm: Average value among beneficiaries with value > 0	A.14
A.13.	DUESyymm: Number of beneficiaries with value > 0	A.15
A.14.	DUESyymm: Percent of SSI beneficiaries with value > 0	A.16
A.15.	DUESyymm: Average value among beneficiaries with value > 0	A.17
A.16.	CONCyymm: Number of beneficiaries with value = 1	A.18
A.17.	CONCyymm: Percent of SSI beneficiaries with value = 1	A.19
A.18.	CONCyymm: Among beneficiaries with populated values, percent with value = 1	A.20
A.19.	PROAyymm: Number of beneficiaries with value = 1	A.21
A.20.	PROAyymm: Percent of SSI beneficiaries with value = 1	A.22
A.21.	PROAyymm: Among beneficiaries with populated values, percent with value = 1	A.23
A.22.	PROByymm: Number of beneficiaries with value = 1	A.24
A.23.	PROByymm: Percent of SSI beneficiaries with value = 1	A.25
A.24.	PROByymm: Among beneficiaries with populated values, percent with value= 1	A.26
A.25.	STWSSlyymm: Number of beneficiaries where STWSSI = 0	A.27
A.26.	STWSSlyymm: Number of beneficiaries where STWSSI = 1, 2, or 3	A.28
A.27.	STWSSlyymm: Number of beneficiaries where STWSSI = 4	A.29
A.28.	STWSSlyymm: Number of beneficiaries where STWSSI = 8	A.30
A.29.	STWSSlyymm: Number of beneficiaries where STWSSI = 9	A.31
A.30.	STWCMyymm: Number of beneficiaries where STWCM = 0	A.32
A.31.	STWCMyymm: Number of beneficiaries where STWCM = 1, 2, or 3	A.33
A.32.	STWCMyymm: Number of beneficiaries where STWCM = 8	A.34
A.33.	STWCMyymm: Number of beneficiaries where STWCM = 9	A.35
A.34.	BFWSSI_DRAFTyymm: Number of beneficiaries with values > 0	A.36
A.35.	BFWSSI_DRAFTyymm: Among beneficiaries with STW = 0, 1, 2, or 3, share (%) of beneficiaries with values > 0	A.37
A.36.	BFWSSI_DRAFTyymm: Average of values > 0	A.38
A.37.	BFWCM_DRAFTyymm: Number of beneficiaries with values > 0	A.39
A.38.	BFWCM_DRAFTyymm: Among beneficiaries with STW = 0, 1, 2, or 3, share (%) of beneficiaries with values > 0	A.40
A.39.	BFWCM_DRAFTyymm: Average of values > 0	A.41

GLOSSARY

AB	Accelerated Benefits Demonstration
ADM	Awardee Data Mart
AIME	Average Indexed Monthly Earnings
B.E.S.T.	Benefits Entitlement Services Team
BFW	Benefits forgone due to work
BIC	Beneficiary Identification Code
BMF	Budget Month Factor
BOAN	Beneficiary's Own Account Number
BOND	Benefit Offset National Demonstration
BOPD	Benefit Offset Pilot Demonstration
CAN	Claim Account Number
CDR	Continuing Disability Review
CDRCF	CDR Control File
CER	Characteristics Extract Record 100% Field File
COLA	Cost-of-Living Adjustment
DAC	Disabled Adult Child
DAF	Disability Analysis File (previously known as TRF)
DBAD	Disabled Beneficiary and Dependents Extract
DCF	Disability Control File
DDS	Disability Determination Services
DER	Detailed Earnings Record
DI	Disability Insurance, also referred to as SSDI
DMG	Demographic component of the DAF
DSN	Dataset names

DWB	Disabled Widow Beneficiaries
EN	Employment Network (also called a TTW provider)
EPE	Extended Period of Eligibility
EVS	Enumeration Verification System
EXR	Expedited Reinstatement
FBR	Federal Benefit Rate
FCI	Federal Countable Income
FIPS	Federal Information Processing Standards (in reference to U.S. Census standardized codes for uniform identification of geographic entities)
FRA	Full Retirement Age
HI	Hospital Insurance (Medicare Part A)
HOPE	Homeless Outreach Projects and Evaluation Demonstration
HUN	Housed Under Number
ICD-9	International Classification of Diseases Coding Scheme
IPE	Individualized Plan for Employment, developed by SVR Agency
IRS	Internal Revenue Service
IRWE	Impairment-Related Work Expense
LAF	Ledger Account File
LAUS	Local Area Unemployment Statistics
LRF	Longitudinal Record Format
MBR	Master Beneficiary Record
MBR810	MBR extract, version number 810
MBR814	MBR extract, version number 814
MEF	Master Earnings File
MHTS	Mental Health Treatment Study
MIE	Medical Improvement Expected

MO	Milestone + Outcomes payment system
MPR-EVS	Mathematica's EVS
NBS	National Beneficiary Survey
NSCF	National Survey of SSI Children and Families
NUMIDENT	Numerical Identification File
OIM	Office of Information Management
OO	Outcomes-Only payment system
PAN	Person's Account Number
PASS	Program to Achieve Self-Support
PHUS	Payment History Update System
PIA	Primary Insurance Amount
PIN	Personal Identification Number
POD	Promoting Opportunity Demonstration
POMS	SSA's Program Operations Manual System
PROMISE	Promoting Readiness of Minors in SSI
Provider	Service provider under TTW (also called an EN)
PUF	Public Use File
REMICS	Revised Management Information Counts System
RIB	Retirement Insurance Benefits
RMA	Retrospective Monthly Accounting
RSA	Rehabilitation Services Administration
RSA-911	RSA Case Service Report
SAIPE	Small Area Income and Poverty Estimates
SAS	Statistical Analysis Software, used to produce the DAF
SCWF	Standalone Companion Work File

SED	Supported Employment Demonstration
SER	Summary Earnings Record
SGA	Substantial Gainful Activity
SMI	Supplemental Medical Insurance (Medicare Part B)
SNAP	Supplemental Nutrition Assistance Program
SSN	Social Security Number
SSA	Social Security Administration
SSDI	Social Security Disability Insurance (also referred to as DI)
SSI	Supplemental Security Income
SSI-LF	SSI - Longitudinal File
SSR	Supplemental Security Record
STW	Suspension or termination of cash benefits for work
SVR Agency	State Vocational Rehabilitation Agency
T2	Title II, the SSDI Program
T16	Title XVI, the SSI Program
TANF	Temporary Assistance for Needy Families
TCNEI	Total countable non-earned income
TKT	DAF component containing data related to TTW participation
TRF	Ticket Research File, now called the DAF
TTW	Ticket to Work
TWP	Trial Work Period
VR	Federal/State Vocational Rehabilitation program
VRRMS	Vocational Rehabilitation Reimbursement Management System; data from this system is contained in the Payments component
YTD	Youth Transition Demonstration

OVERVIEW OF DAF DOCUMENTATION

The documentation for the DAF consists of the eleven volumes described below. Questions about these documents should be directed to ORDES.DAF@ssa.gov. All of these documents are available at <https://www.ssa.gov/disabilityresearch/daf.html>.

- **Volume 1: Getting Started with the DAF18.** Provides an overview of the structure and contents of the DAF and related linkable files.
- **Volume 2: Working with the DAF18.** Contains practical suggestions such as how to extract data and interpret blank or missing variables as well as more detailed information on DAF data marts and linkable files.
- **Volume 3: Tips for Conducting Analysis with the DAF18.** Contains suggestions for working with common research concepts in the DAF such as program participation, benefits paid versus benefits due, and constructed measures related to beneficiary work activity resulting in the loss of cash benefits.
- **Volume 4: Lists of DAF18 Variables.** Contains lists of new, changed, and deleted variables, as well as lists of variables by DAF component and analytic category.
- **Volume 5: DAF Variable Detail Pages.** Contains specifications for each DAF variable, including name, definition, data format, identification of the DAF component to which it belongs, data source, availability, and (where applicable) SAS code used to construct the variable.
- **Volume 6: Validating the DAF18 Against Other Sources.** Provides an explanation of validation methods and summary of validation results.
- **Volume 7: DAF18 Development History and Construction Methods.** Describes key changes in DAF construction methodology over time as well as a description of each step in the current year DAF construction process.
- **Volume 8: DAF18 Construction Workflow Charts and Task Tables.** Provides detailed information in both chart and table format on each step in the current year DAF construction process.
- **Volume 9: DAF18 Source File Descriptions.** Describes the administrative source files used to construct the DAF.
- **Volume 10: SSA Administrative Source File Documentation.** Contains documentation from SSA on the administrative source files described in Volume 9.
- **Volume 11: DAF18 Construction Code.** Contains all SAS code used to construct the DAF.
- **Volume 12: RSA Administrative Source File Documentation.** Contains a description of the processing of Rehabilitation Services Administration (RSA) data for linkage to the DAF, along with documentation from RSA on the RSA-911 files.

The following table provides specific locations for common research-related questions and issues.

In order to ...	Refer to ...
Get started with a research task	Volume 2, "Working with the DAF18," for information about selecting beneficiaries using finder files versus selection criteria
Identify what's changed in the latest version of the DAF	Volume 1, "Getting Started with the DAF18"
View lists of DAF variables	Volume 4, "Lists of DAF18 Variables"
Understand individual variable definitions, specifications, and value ranges	Volume 5, "DAF Variable Detail Pages"
Understand the structure of the DAF data files at a high level	Volume 1, "Getting Started with the DAF18"
Identify variables for a specific research task	Volume 4, "Lists of DAF18 Variables," for a list of variables contained within each DAF file and by analytic category
Understand the beneficiaries for which the DAF does and does not contain data	Volume 1, "Getting Started with the DAF18"
Identify administrative data sources for the DAF	Volume 9, "DAF18 Source File Descriptions"
Understand the linkage of the DAF to RSA-911 data and contents of the RSA files	Volume 12, "RSA Administrative Source File Documentation"
Generate ideas for using the DAF more efficiently	Volume 1, "Getting Started with the DAF18" and Volume 2, "Working with the DAF18"
Find suggested ways to identify common research concepts in the DAF, such as calculating age of retirement, or disability title	Volume 3, "Tips for Conducting Analysis with the DAF18"
Understand what variables have changed in the most recent DAF	Volume 4, "Lists of DAF18 Variables"
Read about how information in the DAF is validated against other sources	Volume 6, "Validating the DAF18 Against Other Sources"

I. THE EVOLUTION OF THE DAF OVER TIME

In recent years, the Disability Analysis File (DAF) database has been constructed by Mathematica on an annual update cycle. Each year, the DAF is rebuilt from scratch, so that records for beneficiaries already in the DAF are updated while records for new beneficiaries who first participated in Social Security Disability Insurance (SSDI) or Supplemental Security Income (SSI) during the most recent year are added. As records are updated with more recent information, a beneficiary's specific variable values may differ between versions of the DAF. For example, in the DAF15 the DUES1503 value—the SSI benefit due in March 2015—might be \$500 while the DUES1503 value in the DAF16 might be \$0. This type of change would imply that Social Security Administration (SSA) revised their determination as to the benefit due in March 2015 sometime after the data were pulled for DAF15 but before the data were pulled for the DAF16.

To date, the following versions of the DAF database have been constructed:

1. The first version, Ticket Research File (TRF.1), was completed in the spring of 2004 and contained data on working-age disabled beneficiaries who participated in SSI or SSDI in one or more months between March 1996 and August 2003. The file included monthly data for these participants beginning in January 1994 and ending in August 2003.
2. The second version, TRF.2, was completed in July 2005. It built on the existing database by incorporating all beneficiaries already included in TRF.1 and expanding it to include new beneficiaries who entered the SSI or SSDI programs between September 2003 and September 2004. Monthly participation data extended to December 2004 for all included beneficiaries.
3. The third version was named TRF05 to indicate that its contents included monthly data through 2005. TRF05, which was completed in July 2006, built on the existing database by incorporating all beneficiaries included in TRF.2 and expanding it to include new beneficiaries who entered the SSI or SSDI programs between October 2004 and December 2005. Additionally, the selection criteria were revised to include children age 10 or older who participated in the SSI program in at least one month between January and December 2005.¹

¹ See Chapter III, “Changes in Beneficiary Selection Criteria Across DAF Versions” for further details and the current selection criteria.

4. The fourth version, named TRF06 to indicate that its data extends to December 2006, was completed in 2007. It built on the existing database by incorporating all beneficiaries included in TRF05 and expanding it to include new beneficiaries who entered the SSI or SSDI programs between January 2006 and December 2006. However, unlike earlier versions, all annual files in TRF06 were reconstructed from scratch to smooth out data inconsistencies resulting from earlier construction activities.
5. The fifth version, named TRF07 to indicate that its data extends to December 2007, was completed in early 2009. It built on the existing database by retaining all beneficiaries included in TRF06 and expanding it to include new beneficiaries who entered the SSI or SSDI programs between January 2007 and December 2007. In this year, the selection criteria for new beneficiaries were expanded to include those between ages 65 and Full Retirement Age (FRA).² A new component, Payments, containing data relating to payments to Employment Networks (ENs) was also added.
6. The sixth version, named TRF08 to indicate that its data extends to December 2008, was completed in late 2009. It built on the existing database by incorporating all beneficiaries already included in TRF07 and expanding it to include new beneficiaries who entered the SSI or SSDI programs from January 2008 through December 2008.
7. The seventh version, named TRF09 to indicate that its data extends to December 2009, was completed in 2010. It built on the existing database by incorporating all beneficiaries included in TRF08 and expanding it to include new beneficiaries who entered the SSI or SSDI programs between January 2009 and December 2009.
8. The eighth version, named TRF10 to indicate that its data extends to December 2010, was completed in 2012. It built on the existing database by incorporating all beneficiaries included in TRF09 and expanding it to include new beneficiaries who entered the SSI or SSDI programs between January 2010 and December 2010.
9. The ninth version, named DAF11 to indicate that the name of the database had changed to the DAF and that its data extend to December 2011, was completed in 2013. It built on the existing database by incorporating all beneficiaries included in TRF10 and expanding it to include new beneficiaries who entered the SSI or SSDI programs between January 2011 and December 2011.
10. The tenth version, named DAF12 to indicate that its data extends to December 2012, was completed in 2013. It built on the existing database by incorporating all beneficiaries included in DAF11 and expanding it to include new beneficiaries who entered the SSI or SSDI programs between January 2012 and December 2012. Deceased beneficiaries were dropped from the Annual files beginning with the year after their death. As a result, each Annual file on the DAF12 contains a different number of records than are present in the other three core components (demographic component of the DAF [DMG], Ticket, and Payments).
11. The eleventh version was named DAF13 to indicate that its data extended through December 2013; the core components of the DAF were completed in 2014 and revised in 2015. It built on the existing database by incorporating all beneficiaries included in DAF12

² Until DAF16, beneficiaries between 65 and FRA were included in 2007 onward. In DAF16, we applied these selection criteria across all years, adding in beneficiaries between 65 and FRA from 1996 onward.

- and expanding it to include new beneficiaries who entered the SSI or SSDI programs between January 2013 and December 2013.
12. The twelfth version was named DAF14 to indicate that its data extended through December 2014; the core components of the DAF were completed in late 2015 and revised in 2016. It built on the existing database by incorporating all beneficiaries included in DAF13 and expanding it to include new beneficiaries who entered the SSI or SSDI programs between January 2014 and December 2014.
 13. The thirteenth version was named DAF15 to indicate that its data extended through December 2015; the core components of the DAF were completed in late 2016. It built on the existing database by incorporating all beneficiaries included in DAF14 and expanding it to include new beneficiaries who entered the SSI or SSDI programs between January 2015 and December 2015.
 14. The fourteenth version was named DAF16 to indicate that its data extended through December 2016; the core components of the DAF were completed in spring 2018. It built on the existing database by incorporating all beneficiaries included in DAF15 and expanding it to include new beneficiaries who entered the SSI or SSDI programs between January 2016 and December 2016. It also expanded on the earlier selection criteria to include SSI youth (ages 0-9) from 2005 – 2016. In DAF16, we also obtained data from 1996 through 2006 of beneficiaries between the ages of 65 and FRA.³ In DAF16, we switched the source file for SSI earnings variables (those indicated with a T16 in their name) from the DCF to the SSR. We did this after SSA discovered that some earnings that were recorded in the SSR were not being transferred to the DCF. As a result, the earnings values in the DAF prior to this version that were sourced from the SCF were underestimates of earnings values. An investigation at the time suggested that the DCF values were about 5 percent lower than those from the SSR.
 15. The fifteenth version was named DAF17 to indicate that its data extended through December 2017. The core components of the DAF were completed in early 2019, but were rebuilt in August 2019 to account for a legacy error that affected SSI earnings and benefits paid variables since the inception of DAF. This issue and implications for research are contained in Section V. It built on the existing database by incorporating all beneficiaries included in DAF16 and expanding it to include new beneficiaries who entered the SSI or SSDI programs between January 2017 and December 2017. It also expanded on the earlier selection criteria to include SSI youth (ages 0-9) from 1996 – 2004. As a result, the DAF contained all SSDI beneficiaries who received at least one month of benefits from March 1996 through December 2017 and who were ages 18 to FRA. It also included SSI beneficiaries who received at least one month of benefits while under FRA during that same time span.
 16. The current version is the sixteenth iteration of the DAF, named DAF18 to indicate that its data extend through December 2018. The core components were completed in January 2020. It built on the existing database by incorporating all beneficiaries included in DAF17 and expanding it to include new beneficiaries who entered the SSI or SSDI programs between

³ Note that by virtue of the selection criteria of beneficiaries up to age 65 in the years from 1996 through 2006 and through FRA in 2007 onward, this change was quite small, capturing only beneficiaries who *only* received SSDI or SSI between 65 and FRA and did not receive at any younger age in another year.

January 2018 and December 2018. The selection criteria for the DAF remained the same as in the previous version.

Beginning with the DAF13 and continuing in subsequent versions of the database, beneficiaries with no data in any of the twelve monthly occurrences of ten key variables were excluded from that year's Annual file. Those ten key variables are: LAFyymm, PSTAyymm, PAYDyymm, PAYSyymm, STWDIyymm, STWSSIyymm, STWCMyymm, BFWDIyymm, BFWSSI_DRAFTyymm and BFWCM_DRAFTyymm. In most cases a lack of data for all twelve occurrences of these ten variables means that the beneficiary had either not yet entered the disability rolls by December of that year or died prior to January of that year. Additionally, any beneficiary who reached FRA on or before January 1 of the year is excluded from that year's Annual file, regardless of the data in the ten key variables. In DAF13 through DAF15 those beneficiaries were included on a separate companion Non-Enrolled Annual file (described in more detail in Volume 1). This change was made to reduce file size and make processing more efficient. Beginning in DAF16, these Non-Enrolled Annual files are no longer available, as they proved to be of limited research use but were increasing in size.

As a result of keeping individuals who meet the age selection criteria during the year and have populated information on at least one of the ten variables, each Annual file contains a different number of records, reflecting a changing composition of beneficiaries meeting the selection criteria. In general, beneficiaries whose entitlement to SSDI and SSI terminates are removed from Annual files beginning with the year after their termination. An exception, however, is beneficiaries who terminated as a result of work. Because these beneficiaries continue to have valid suspension or termination of cash benefits for work (STW) and benefits forgone due to work (BFW) variable values, they do not meet the criteria for removal even though their entitlement has ceased.

II. OVERVIEW OF DAF CONSTRUCTION TASKS

This section provides a general description of the steps to build the DAF. Additional detail on construction methods is available in Volume 8, which contains workflow charts and task tables for each of the tasks outlined here. However, because this description is general, task names and numbers here may not precisely match the names and titles in Volume 8.

Over time, the criteria used to select beneficiaries for the DAF have changed, and are described more fully in Chapter III. The steps described in this section presume knowledge of the source files used in constructing the DAF; these are described in more detail in Volume 9. Changes in the source files over time are also described in Volume 9.

A. Task 1: Assemble and combine DBAD files

The first step is to identify the new beneficiaries who entered the SSDI program during the year to be added to the database; this is accomplished by processing the twelve monthly Disabled Beneficiary and Dependents Extract (DBAD) files for that year. For DAF18, we processed the twelve DBAD files that contained data for January 2018 to December 2018. We convert each DBAD file to Statistical Analysis Software (SAS) format and select records for beneficiaries participating in the SSDI program. Selection criteria are based on the program participation variables Beneficiary Identification Code (BIC), Ledger Account File (LAF) (Status), Type of Claim (TOC), and beneficiary's age. For records of primary beneficiaries (BIC = "A") we use the Claim Account Number (CAN) to populate the Social Security Number (SSN) field in the DAF. For records for auxiliary beneficiaries (where BIC begins with "C" or "W"), we use the Beneficiary's Own Account Number (BOAN) to populate the SSN field, but we also keep the CAN to facilitate later linking with other SSA administrative data. We delete any records with blank CANs because these cannot be matched to other SSA administrative files. We then de-

duplicate on SSN combined with BIC, rather than on SSN alone, which allows us to retain multiple records for dual-eligible beneficiaries, i.e. primary beneficiaries who are entitled to benefits from their own account but are also entitled to dependent benefits from the account of another primary beneficiary. At the end of this step, we have twelve files containing records for selected SSDI beneficiaries from the year being added to the database.

Next, we merge the twelve files of records selected and processed from the DBAD files to create the SSDI finder file. We also keep selected variables for later processing to build longitudinal variables for the DAF Annual files.

B. Task 2: Assemble and combine CER

We then identify the new beneficiaries who entered the SSI program during the year to add to the database. This is accomplished by processing the twelve monthly Characteristics Extract Record 100% Field File (CER) that represent the selected year. For DAF18, the twelve selected CER contained data for January 2018 to December 2018. We convert each CER to SAS format and select records for beneficiaries participating in the SSI program. Selection criteria are based on the variables PSTAT (Payment Status), MFT (Master File Type), DENCDE (Denial Code), and age. Because each SSI record is listed under the beneficiary's own SSN (PAN), we set SSN to PAN as the identifier for SSI beneficiaries in the DAF. At the end of this step, we have twelve files containing records for selected SSI beneficiaries from the year being added to the database.

Next, we combine the twelve files of records selected and processed from the CER to create the SSI finder file. We also keep selected variables for later processing to build longitudinal variables for the DAF Annual files.

C. Task 3: Create finder files

We combine the SSNs from the previous version of the DAF with the lists of SSNs (derived from the CANs, BOANs, and PANs) from the DBAD and CER in the steps above to build a

finder file of all SSNs for inclusion in the new DAF based on the selection criteria used in the relevant version of the DAF. Then we identify the type of beneficiary for each SSN: SSDI, SSI, or concurrent, and output the SSNs into separate finder files for SSI and SSDI. Both finder files contain the SSNs for the concurrent beneficiaries. We create combined files of all SSNs (BOANs for SSDI and PANs for SSI), to be used as finders for the NUMIDENT and earnings data. We also create a combined file of SSNs (CANs for SSDI and PANs for SSI) to be used when extracting records from 831 & 832/833 files.

D. Task 4: Submit finder files

We submit the SSDI finder through SSA's Master Beneficiary Record (MBR) finder process and request two output components: 1) selected MBR variables, as specified by the custom output layout from the prior year, and 2) all Payment History Update System (PHUS) variables. This routine captures all the MBR records, both primary and auxiliary, associated with each CAN. If needed, new MBR output variables can be requested by adding them to the previous year's custom output layout.

We submit the SSI finder to SSA's SSI Longitudinal File (SSI-LF) finder process. Data are returned from the finder process in one large file, which must be temporarily divided into multiple segments in order to be read into SAS. Starting in DAF14 we began receiving a custom extract of select Supplemental Security Record (SSR) variables related to SSI payment calculation, not available on the SSI-LF. In DAF16, we began receiving an additional custom extract of select SSR variables related to beneficiary earnings; we use these variables for constructing STW and BFW for SSI beneficiaries.

We submit the combined finder (for all SSDI and SSI beneficiaries) to the Numerical Identification File (NUMIDENT) finder process and also submit it to obtain earnings from 1990 onwards from the Master Earnings File (MEF). Because access to the MEF data is available only

to select SSA staff, the finder results from the MEF will be stored in a location accessible only to those staff.

E. Task 5: Process 831 & 832/833 data

Using the list of SSNs created above, we extract and combine records from the 831 & 832/833 files, which contain data for both SSDI and SSI beneficiaries. SSDI records in these files are identified only by CAN/BIC, which means that the record for an auxiliary SSDI beneficiary contains the CAN of the primary, not the BOAN of the auxiliary, which can hinder proper linking with other files. Therefore, we create a linking file for these records using the crosswalk generated during MBR processing to attach BOANs to the CAN/BICs. For SSI beneficiaries, we build histories of stop and start dates and set them aside to be added to the DAF DMG component. For both SSDI and SSI beneficiaries, we build longitudinal variables for disability adjudication, diagnosis codes, Medical Improvement Expected (MIE) indicators, and levels of education, and set those aside to be added to the Annual files.

F. Task 6: NUMIDENT processing

The NUMIDENT records are returned from the finder process as one large flat file with several records for each submitted SSN. To facilitate processing, this file is broken into record segments. During the conversion from flat to SAS format, selected variables needed for DAF are conditionally read by record type. These segments are then combined and collapsed to a one-record-per-beneficiary format by selecting the most recently populated version of a variable. The result is a file in DAF format (one record per beneficiary) that contains variables drawn from multiple NUMIDENT record types. This file is then set aside to be used in the construction of the DMG.

G. Task 7: Process SSI-LF data

The first step of this task is to load the SSI-LF returned records into SAS, excluding certain records missing key data and splitting into record segments to facilitate processing. The segments are then combined and processed to build a one record per SSN file with monthly data occurrences. We extract demographic and non-monthly data such as birthdate and SSI application dates for eventual storage in the DAF DMG component. We also extract monthly data such as living arrangements, benefit paid amounts, and payment status codes for 1994 through the year being added to the database, determining the month and year the data belongs to, then storing it in longitudinal fields named accordingly.

Starting in DAF14, we also receive a custom extract of select SSR variables, not available on the SSI-LF. This file is returned with records at the SSN, Records Establishment Date, and benefit computation month level. We process these records to combine this data into a one record per SSN format with monthly variables and then combine with the processed SSI-LF. This file is then set aside to eventually be used in the construction of the DAF DMG and Annual components in a later task.

Starting in DAF16, we also began receiving a second custom extract of SSR SSI earnings variables not available on the SSI-LF. The earnings extract is at the SSN, Record Number, and Record Establishment Date level. Both custom extracts are processed into a one record per SSN format and incorporated into the SSI-LF file. This file is set aside to eventually be used in the construction of the DAF Annual component in a later task.

H. Task 8: Process MBR data

Mathematica processes the MBR files returned from the MBR finder process, converting them to SAS. These files are in the form of a custom extract that includes only variables

requested by Mathematica for DAF purposes. Because this conversion process is complex, we specify the steps involved below:

- **Step 1:** We obtain monthly auxiliary data by sorting the returned MBR records by CAN and BIC. For each primary beneficiary, we sum the benefit amounts due to their auxiliary/dependent beneficiaries. A record for each primary beneficiary is output, containing variables reflecting the total benefits paid and the amount due to their auxiliaries as well as the number of auxiliaries. Although we have a record for each primary beneficiary the file from this step only contains summarized auxiliary data related to that primary.
- **Step 2:** We process the PHUS data in a similar fashion as above but are processing the benefit amounts paid instead.
- **Step 3:** We read the returned MBR records a second time to obtain data for primary beneficiaries by processing the BOANs, this time extracting demographic and time-invariant data, such as birthdate, as well as longitudinal data, such as SSDI application dates and PIA amounts, for eventual storage in the DAF DMG component. We also extract longitudinal data such as monthly payment status and benefit amount due for 1994 through the year being added to the database, determining the month and year the longitudinal data belongs to, then storing it in longitudinal fields named accordingly. We process the PHUS data in a similar fashion and combine it with the longitudinal data from the MBR. The file is sorted on the SSN/BIC combination, then de-duplicated by SSN. For records with multiple SSN/BIC combinations, we keep the BIC from the first occurrence.
- **Step 4:** Next we process the variables for dual-eligible SSDI beneficiaries, i.e. for a beneficiary entitled to benefits based on their own primary SSDI record as well as benefits from another primary SSDI beneficiary, such as a spouse or parent's SSDI record, and add it to the file we produce in Step 3 above.
- **Step 5:** We attach the dependent/auxiliary amounts from Steps 1 and 2 to the records in the file produced by Steps 3 and 4. The resulting data are set aside for later inclusion in the Annual components.
- **Step 6:** In the final step, we create a crosswalk of SSNs, CANS, BICs to facilitate later processing of the 831 & 832/833 data.

I. Task 9: DMG Pre-processing

Before creating the DMG component, we develop the demographic and time-invariant data for beneficiaries using the Disability Control File (DCF) Claims and Medical data. In order to do this, the first step is to create SAS extracts of the claim group data, historical claim data, claim medical history data, and the historical medical group data. To generate the DCF claim table, which provides the monthly trial work period completion information, the claim data is merged

with the master list of SSNs to retain only those SSNs that appear in the DAF18. To generate the DCF Medical table, the blind date information is obtained then merged onto the claim medical data and the master list of SSNs to retain only those SSNs that appear in DAF18. After the merge, we develop the monthly improvement variables (MEDEXyymm) then collapse the dataset to one record per SSN then split into yearly files. Lastly, we create a table of diagnosis variables using the MBR and 831 data, subsequently split it into yearly files.

J. Task 10: Create DAF.DMG component

The DMG component file is created from scratch for each DAF version by combining the demographic and time-invariant data for SSDI and SSI beneficiaries that was processed in earlier steps, creating a single record for each beneficiary. For an SSDI-only beneficiary, the SSI fields in the beneficiary record of the DMG component will be blank, while for an SSI-only beneficiary, the SSDI fields will be similarly blank, and for a concurrent beneficiary, all fields will be populated. For some fields with multiple sources of data, such as birthdate, which is available from both the MBR and the SSR, the DMG component will contain all versions of the variable, even if there are discrepancies. For these variables, we employ algorithms to determine the “best” field, which is captured as an additional variable in the DMG component.

K. Task 11: Create DAF.Ticket component

The Ticket data for the Ticket component is processed separately from the data for the DAF DMG component and Annual files. The first step is to SAS load the various DCF DB2 tables that hold Ticket data. For each DCF table, the programs loop through multiple Ticket records for each beneficiary, such as records for Ticket mailing and assignment dates, then build a single record for each SSN with monthly flags for Ticket events, including Ticket mailing dates, Ticket assignment dates, and types of providers of employment services. Using the DCF table of monthly Ticket data, we construct each beneficiary's monthly program participation status (SSDI

or SSI) within Ticket to Work (TTW) from the effective date of the TTW program participation until a recorded program participation change. These constructed indicators apply only within the context of the TTW program as it relates to provider reimbursement; they should not be understood as proxies for the beneficiary's program participation within the broader context of the SSI or SSDI programs.

Note that not all SSNs in the Ticket data will match to the DAF DMG component and Annual components, and vice versa, for three main reasons. First, we usually process the DCF DB2 tables (used when creating the DAF Ticket components) several months after records are extracted from the DBAD and CER (used to build DAF.DMG). Therefore, records added recently to the DCF may not have a counterpart in the DBAD or CER. For example, a beneficiary who began SSDI participation in February 2019 and had a Ticket mailing date in March 2019 will not have records in the DBAD files for 2018 and therefore will have a record in the Ticket component of DAF18 but will not have a corresponding record in the DMG component. Second, a beneficiary who participated in either SSI or SSDI prior to 2002 but not after will appear in the DAF.DMG component but not in the Ticket data as the Ticket program did not begin until 2002. Third, while all Ticket participants in the DCF are selected for inclusion in the Ticket data, records from the DBAD and CER are selected only if the beneficiary meets certain age and program participation criteria. The number of beneficiaries for whom this occurs is quite small and consists only of those who began participating in SSI or SSDI between January 1 of the year following the last year covered by the DAF (i.e., January 1, 2019 for the DAF189) and the time of Ticket data extraction.

Once all the Ticket data is generated, we separate it into a series of files to make each the files a more manageable size. We create a Ticket Base that contains Ticket event data such as

Ticket mail and assignment dates along with a series of Ticket Annual files that contain series of monthly flags for a given year that indicate participation in the Ticket program, for instance, whether a Ticket participant had a Ticket assigned to a provider in April 2004.

Note about dropped Ticket records: Some Ticket records that were in previous versions of the DAF database do not appear in the current version. The reason is that the DCF, from which the Ticket data is derived, is periodically cleaned and some records are removed from the DCF for various reasons. An example is the deletion of a Ticket record that was generated but was later found to have been generated in error because the intended Ticket participant did not meet all the Ticket eligibility criteria. Such a record would exist in earlier DAF versions but will not appear in DAF versions that were built after the DCF was cleaned and that record was deleted from the DCF.

L. Task 12: Annuals Pre-processing

Before building the Annuals File, the first step is to extract earnings data from the DCF tables and reformat it into monthly variables, which are used later during Annuals processing. The next step is to create SAS versions of the DCF SSDI Earnings Table, DCF SSDI Work Detail Table, and DCF Alleged Earnings Table, then individually reformat the tables into monthly variables. In addition, we create monthly SSDI earnings indicator variables using the corresponding earnings data. We also use the monthly SSI earnings variables processed in Task 7 in order to create monthly SSI earnings indicator variables. Afterwards, we combine all SSI and SSDI monthly earning variables into one dataset which is then split into annual files for later use in developing the Annual files in Task 13. The annual historical SEIE files are read into SAS and then combined into annual files. Finally, the Federal Information Processing Standards (FIPS) flat file is uploaded to the mainframe and transformed into a format library for later use.

M. Task 13: Create DAF.Annual data

Every Annual file is created from scratch for each DAF version by combining the longitudinal data built from the MBR, SSR, DCF, CER, DBAD, and 831 & 832/833 files processed in the preceding steps. We create a new Annual file for each year from 1994 to the year being added to the database, each file containing one record for each beneficiary and combining SSDI and SSI data.⁴ For a beneficiary who has only SSDI data, in other words never had an SSI-LF record, the SSI variables are blank or missing on each Annual file. For a beneficiary who has only SSI data, in other words never had an MBR record, the SSDI variables will similarly be blank or missing. For a beneficiary with both SSDI and SSI data, in other words had records on both the MBR and the SSI-LF, both SSI and SSDI variables will be populated in the Annual files. We also add the monthly STW and BFW variables. Finally, we identify and build the most recent state of residence for each beneficiary as of December 2017 for use in finalizing the DMG component. Each record in each Annual component has a one-to-one match to a record in the DAF DMG component. However, the number of records in each Annual component is not constant because of the dual-file structure described above, under which beneficiaries who lacked program participation, payment, or eligibility data are removed to a companion Non-Enrolled Annual file.

N. Task 14: Building STWs and BFWs

Next we build the monthly flags for STW, creating separate series of indicators for STW status in SSDI and SSI programs based on monthly program participation. From the SSDI and SSI indicators, we also create a series of “combined” STW flags which indicate STW status in both programs simultaneously. We then use the STW flags to construct variables providing an

⁴ Note that in the current version of the DAF (and starting with DAF14), the beneficiaries who died prior to 1996 are removed from the Annual file.

estimate BFW for all months in which the STW flag indicates suspense or termination for work, as well as months in which countable earned income reduced but did not eliminate an SSI benefit. We set these data aside for subsequent merging with Annual files. Additionally, we produce an alternative set of these measures for SSI beneficiaries that use a different algorithm and variables. These measures reside on a standalone DAF linkable file; from DAF15 to DAF17, this file was called the SSI Companion Work File (SCWF), in DAF18, the set of alternate measures was swapped so that the SCWF measures became the core version and the old core version moved to the standalone file. For more details on the construction of the STW flags and BFW variables, consult Volume 3.

O. Task 15: Create payments component (EN payments data)

The Payments component consists of data relating to payments made to ENs and State Vocational Rehabilitation (SVR) Agencies acting as ENs under the Ticket program, as well as traditional reimbursement to SVR Agencies outside the TTW program. The EN payment data are initially supplied as an Excel spreadsheet containing multiple records for beneficiaries and recording Milestone + Outcomes (MO) payments made to TTW providers. We convert the data to SAS and perform some basic data cleaning which results in the Vertical file, structured at the payment level. We then take this cleaned data and create the Horizontal file by combining multiple records for a single beneficiary into a single record that preserves the information for multiple events such as dates, amounts, and types of payment, e.g. MO. Extraneous Excel data, such as header and summary rows, is removed.

P. Task 16: Create payments component (VRRMS data)

Payments made to SVR Agencies under the traditional, non-Ticket reimbursement system are contained in the Vocational Rehabilitation Reimbursement Management System (VRRMS) which are stored in two different formats, one that covers the period from January 1994 to

February 2017 and one for the period from March 2017 onward. For both formats, processing this data and placing it on the DAF begins with converting the data to SAS format and collapsing the payment data to the one-record-per-SSN DAF structure. This process differs between the two formats.

For the VRRMS format that covers the period from January 1994 to February 2017 the records are organized into categories describing the status of each VRRMS claim, e.g. “allowed”, “denied”, etc. Many SSNs have multiple VRRMS claims records; for multiple claims records, we determine if the claims history is complete and discard it if not.⁵ For complete histories, we roll all claims records on a spell up to a single record per SSN, summing the payment amounts. For records with only a single claim, no summing of payments is needed. We output the SSN-level record keeping all Vocational Rehabilitation (VR) spells, a maximum of 9 in DAF18, in *n*-suffixed variables, following traditional DAF naming conventions, and a set of variables related to the most recent VR spell only.

For the VRRMS format covering the period from March 2017 onward, the records are kept for all claims for which a reimbursement payment was made. Multiple reimbursement payments made on a claim are rolled up to the claim level (this is expected to be very rare and in fact, there were no such occurrences in the data used in DAF18 processing). We output an SSN-level record keeping all claims in *n*-suffixed variables, following traditional DAF naming conventions, and a set of variables related to the most recent claim only.

⁵ Incomplete claims are identified by examining the earliest available claim detail record within a claim (lowest value of C_DETL_CNTR within a claim). If that claim detail record does not indicate that it is the first such record for the claim (C_DETL_CNTR ne 1) then we delete the entire claim. If a claim is incomplete we will not be able to determine the appropriate payment amounts after adjustments, corrections, and re-computes.

Q. Task 17: Create DAF-RSA files

The RSA data are initially supplied by the Department of Education's Rehabilitation Services Administration (RSA) as text files, one for each year from 1998 through the most recent fiscal year. We convert the RSA data to SAS format and then verify the data against the NUMIDENT in a special Mathematica Enumeration Verification System (MPR-EVS) process.⁶ Next we create all of the DAF-RSA files described in Volume 2.

R. Task 18: Create LAUS/SAIPE formats

In this step, we create SAS format libraries for county-level LAUS (unemployment) and SAIPE (poverty level) statistics; these are described in more detail in Volume 2. When a state and county FIPS code is entered into these formats, they return the corresponding LAUS and SAIPE data.

⁶ Since the RSA-911 data comes from outside SSA, the SSNs contained there need to be verified before being combined with SSA data. Usually this process involves submitting the RSA-911 SSNs to SSA's EVS process, but that is not possible with these data because the source does not include the RSA-911 participant name. With special permission from SSA, we validate with a less robust method we developed, described in more detail in Volume 2.

This page has been left blank for double-sided copying.

III. CHANGES IN BENEFICIARY SELECTION CRITERIA ACROSS DAF VERSIONS

The criteria for selecting beneficiaries to be included in the DAF have evolved over time as researchers have used the DAF to analyze an increasing number of topics. The DAF was initially constructed for an evaluation of the TTW program and, as noted above, includes information on the entire working-age population eligible for SSI or SSDI disability benefits at any point from 1996 onwards. Since initial construction, the DAF has expanded to facilitate a range of disability research that far surpasses TTW. For example, as described above, the DAF began including information on SSI children age 10 and older starting in October 2004.⁷ Understanding this evolution may help researchers understand certain aspects of the DAF's construction and achieve greater utility.

In the first version of the DAF (TRF.1), beneficiaries who participated in SSI or SSDI as early as 1996 were included in order to compare beneficiaries in the pre-TTW period (prior to TTW roll-out in 2002) with beneficiaries in the post-TTW period. The criteria were sufficiently broad to encompass all beneficiaries between the ages of 18 and 65 who were receiving a disability-based SSI or SSDI benefit regardless of meeting over future TTW criteria.⁸

For each DAF version, new beneficiaries were selected if they met eligibility criteria during a specified time period and were combined with the existing beneficiaries from the previous DAF.⁹ We use a wide range of pay status codes for both SSI and SSDI beneficiaries in order to select as many new beneficiaries as possible who might have been in current pay status during

⁷ There are some known problems with the beneficiary selection methodology; these are described in more detail in Volume 1.

⁸ Prior to July 2008, individuals who met MIE standards were not eligible to participate in the TTW program nor were adult SSI beneficiaries who had not had their age 18 redetermination.

⁹ See Chapter I, "The Evolution of DAF Over Time" for the exact months used for selection of new beneficiaries for each DAF version.

the relevant year. For example, if a beneficiary was in suspension in one month, he or she may have returned to full eligibility in a subsequent month, so we want to ensure that we do not omit cases not in current pay status at a given point in time.

The beneficiary selection criteria related to age expanded over time, reflecting the changing uses of the DAF. Beginning in 2005, SSA lowered the age cutoff for selecting new SSI beneficiaries for inclusion in the DAF from 18 to 10 (for SSDI beneficiaries it remained at 18). In DAF16, we expanded the age range to capture children from birth to age 9 from 2005 onward. As we made these changes to expand the criteria, we did not affect the selection criteria that had occurred in prior waves, meaning that children who had received benefits under age 9 only would still not have appeared on the DAF during the earlier iterations. In DAF17, we further expanded the age range to capture children from birth to age 18 from 1996 to 2004. As of DAF17, records of children 0-18 who received benefits in any month from March 1996 through December 2017 are included in the DAF.

Similarly, we extended beneficiaries included in the DAF on the upper end over time as well. Beginning with TRF07, we extended the upper end of the age cutoff used in selecting new beneficiaries for inclusion in the DAF from age 65 to FRA. FRA varies according to each beneficiary's birth date; see Volume 3 for a description of the calculation of FRA. Later versions of the DAF continued to add new beneficiaries using the extended age criteria, but the selection process was not re-done for the earlier DAF years. In DAF16, we applied the selection criteria through FRA to all years of the DAF, meaning that we went back to years before TRF07 to include a small number of additional beneficiaries who had only received benefits between 65 and FRA during those years. As of DAF16, all adult beneficiaries through FRA have been included in the DAF.

IV. CHANGES IN THE DAF SOURCE DATA OVER TIME

Over time, the source files used to build the DAF have been modified; any changes of substance have been reflected in each year's file construction. For example, the SSI monthly extracts currently used in construction of the DAF are the CER; previous versions were known as the Revised Management Information Counts System (REMICS) and SORD¹⁰ files. These files provide a snapshot of the beneficiaries participating in the SSI program during a given month. For TRF.1, REMICS files were used for each month beginning with March 1996 through August 2003. Because there was not a REMICS file available for January or February 1996, any beneficiaries who were on the SSI rolls during those months but who had left the program by March 1996 are not included in the DAF. For TRF.2, SORD files spanning the months from September 2003 to September 2004 were used to select SSI beneficiaries to be added to the DAF. For TRF05, SORD files spanning the months from October 2004 to December 2005 were used to select SSI beneficiaries. For TRF06 through the current DAF, CER files spanning the full year were used to select beneficiaries.

The SSDI monthly extracts currently in use are the DBAD files; the previous files used were the ZIP extracts, which were replaced by the DBAD files. Like their SSI counterparts, the DBAD files provide a snapshot of the beneficiaries participating in the SSDI program during a given month. For TRF.1, ZIP extracts were snapshots of the MBR available only every six months between June 1996 and December 1998, after which they became available quarterly. Beginning with January 2001, the monthly DBAD files were used to select DAF beneficiaries up through August 2003. As with the SSI beneficiaries, it is possible that some beneficiaries who were on the SSDI rolls during the early months of 1996 were not selected for inclusion in TRF.1. For

¹⁰ Unofficially, "Son of REMICS Data."

example, an SSDI beneficiary who entered the program in July 1996 but left the rolls in November 1996 would not be included in either of the ZIP extracts for 1996 and would therefore not be in the TRF. For TRF.2, DBAD files spanning the months from September 2003 to September 2004 were used to select SSDI beneficiaries to be added to the DAF, while for TRF05, DBAD files from October 2004 to December 2005 were used. Finally, for TRF06 through the current DAF, DBAD files spanning the full year were used to select beneficiaries.

V. LEGACY PROCESSING ERROR AFFECTING SSI EARNINGS AND BENEFITS DUE DISCOVERED AND CORRECTED IN DAF17

In this section, we document an error that we discovered during processing the 2017 version of the Disability Analysis File (DAF17). This error was present in all earlier versions of the DAF and its predecessor, the Ticket Research File (TRF). As a result of the error, measures of earnings and benefits due for SSI beneficiaries in DAF16 and earlier versions of the DAF were too low, with the effect being larger for historical data than for more recent data. Variables derived from earnings, including measures of suspense or termination for work (STW) and benefits forgone due to work (BFW) were also affected, also with larger effects on historical data.

In what follows, we outline the nature of the error, the variables it affected, and how those variables differ under the old (incorrect) and new (correct) processing.

A. Description of the error

The coding error affected records in the DAF that were derived from the Supplemental Security Income (SSI) Longitudinal file (SSI-LF), during the process of converting multiple SSI records into the one-record-per-beneficiary structure of the DAF. Each record in the SSI-LF contains monthly data on benefits due, benefits paid, and earnings information (used to calculate SSI benefits due) covering a certain number of months. While each record can contain data for many months, in certain instances, SSI beneficiaries may have more than one record. A beneficiary can have multiple records for several reasons. For example, a new record is often established if it is a new period of SSI eligibility, if a child SSI recipient becomes an adult SSI recipient, if there are changes in the individual's status that affect benefits, if the record becomes too "complicated" over time due to recording of earnings or benefits data, or if the record simply

runs out of space after many months of benefit receipt. Because of the various reasons for which a new SSI record can be established, more than half of SSI beneficiaries in the DAF have more than one SSI record.

Each SSI record has an “establishment date” that indicates when it was started. Though one might expect that multiple SSI records for the same person would not contain overlapping data for the same calendar months, this is not always the case in practice. In fact, it is common for a record to have data populated in months that precede the record establishment date, and thus, overlap with monthly data from an earlier record. Moreover, data for the same month from two different records do not always align. In some cases, the later-established record may have populated data from months before the establishment date with zeros, rather than leaving the values missing or populating the record with the same information as the earlier record. To address this, we designed the intended DAF construction logic to allow us to create a single DAF record from a combination of multiple SSI-LF records for cases with multiple SSI records with inconsistent values for a given month is as follows:

- The monthly value from the latest established record will overlay the value from the earlier-established record, provided that the later-established value is not zero or missing.
- If the value on the later-established record is zero or missing then the value from the earlier-established record will be used.

The processing error represented a departure from the intended logic. Specifically, in some instances, the erroneous processing allowed zero values after the first record to overwrite non-zero values in the first established record. Affected records include those for which the first established record had a populated monthly value, but *all subsequent records* had a zero amount for that same month. In those cases, the zero value incorrectly replaced a valid, populated value from the first established record. In practice, this meant that older positive benefits due and earnings values were set to zero based on newer records when they should not have been.

To demonstrate this, we show a hypothetical example in Table 1. In this example the values for January, February, and April of 2005 (0501, 0502, and 0504) were incorrect using the old DAF processing because these records had only zero values for these months in all records after the first record. Though March 2005 (0503) had a 0 in the third record, it was correctly processed because there was a populated value of \$25 on the second record.

Table VII.1. Example of combining SSI-LF records in the old (incorrect) and new (corrected) DAF processing algorithm

	Record establishment date	EICM 0501	EICM 0502	EICM 0503	EICM 0504	EICM 0505	EICM0 506
Raw data from the SSI-LF							
Record 1	January 2002	150	50	200	100	.	.
Record 2	April 2005	0	0	25	0	0	0
Record 3	June 2005	0	0	0	0	50	100
Old DAF processing	--	0	0	25	0	50	100
New DAF processing	--	150	50	25	100	50	100

B. Variables affected by the error

The error affected variables related to SSI benefits due and earnings; it did not affect any character variables. It also did not affect benefits paid variables, as those data were correctly processed by summing together information from all established SSI records. It also did not affect any variables specific to the SSDI program. The variables that were directly affected by the error are:

- Countable earned income, *EICMyymm*
- Unearned income, *UINCyymm*
- Federal benefit amount due, *FAMTyymm*
- State benefit amount due, *SAMTyymm*
- Total amount due (the sum of *FAMT* and *SAMT*), *DUESyymm*

Several additional variables were indirectly affected by the error because they were derived from the variables listed above. These are:

- An indicator for 1619a status, *PROAyymm*

- An indicator for 1619b status, *PROByymm*
- An indicator for concurrently being due an SSI and Social Security Disability Insurance (SSDI) benefit, *CONCyymm*
- Variables indicating suspense or termination of cash benefits for work; *STWSSIymm* for SSI alone, and *STWCMymm* that considers both SSI and SSDI
- Variables indicating the dollar value of cash benefits suspended for work; *BFWSSI_DRAFTymm* for SSI alone, and *BFWCMymm* for combined SSI and SSDI

C. Implications of the error for snapshot statistics

There are two general implications of the error on benefits due and earnings. First, because the error caused a subset of earnings (or benefits due) records with positive values to be replaced with zeros, correcting the error resulted in the incorrect observations being changed from zeros to positive values. Therefore, aggregate measures of earnings (or benefits due) derived from the affected variables are expected to increase as a result of the correction. Similarly, measures of the number of beneficiaries with non-zero earnings (or benefits due) based on the affected variables are expected to increase. However, average values calculated over the set of non-zero earnings (or benefits due) values could increase, decrease, or stay the same. The direction of the change depends on the distribution of the incorrectly omitted non-zero observations compared to the distribution of the remaining observations. For example, if the correct values of the observations that were incorrectly set to zero tend to be higher than the unaffected observations, after the correction, the average observed non-zero value for an affected variable will increase.

The second implication is that the effect on current year values is expected to be small or non-existent, because the probability of multiple records in a single year is small, but the magnitude of the effect of the error will increase the farther back in time (relative to the current DAF year) one observes. This is because only beneficiaries with multiple SSI record establishments could have overwritten values, and the likelihood of that is greater the farther back in time one moves from the DAF year. For example, in the DAF17, the error would affect a

larger share of observations in 2010 than in 2017. Furthermore, we would expect that the error in the 2010 observations in the DAF17 would be of a larger magnitude than the error in the same observations in the DAF13.

D. Magnitude of the effect of the error

1. Directly-affected variables

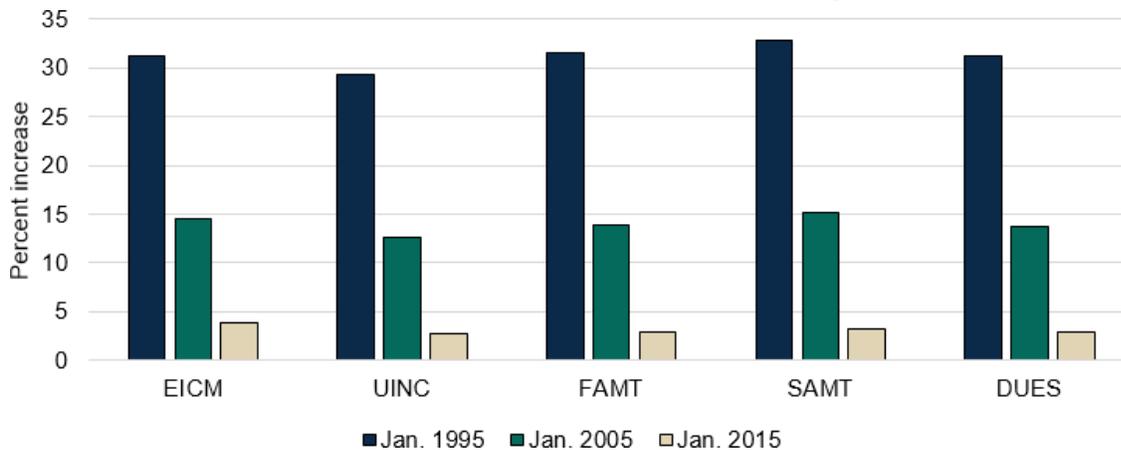
Figures VII.1, VII.2, and VII.3 show summary measures of the magnitude of the effect of error at three points in time for the directly-affected variables. The three points in time, January of 1995,¹¹ 2005, and 2015; are arbitrary but provide a snapshot at evenly-spaced intervals over the time period covered by the DAF17. The appendix to this memo contains graphs of these measures that span all months in DAF17 (January 1994 through December 2017). In the remainder of the memo, we discuss variables using only their prefix, omitting the *yymm* suffix of the variable name; we instead reference the calendar month of interest.

As discussed above, correcting the processing error resulted in an increase in populated positive values in almost all months, with the change getting larger for months farther in the past. Figure VII.1 shows the percentage increase in the number of beneficiaries with a positive value as a result of the correction.¹² For all of the directly-affected variables, the correction resulted in an increase in positive values of approximately 30 percent for observations in January 1995. Positive observations in January 2005 increased by about 15 percent, and in January 2015, the increase was about 3 percent.

¹¹ DAF contains data on beneficiaries who participated in SSI in one or more months starting in March 1996. The earliest observations for these beneficiaries go back to January of 1994. Therefore, observations from January 1995 do not include beneficiaries who had stopped participating before March 1996. In contrast, the January 2005 and 2015 observations include all SSI participants in those months. We confirmed that the patterns we show for 1995 look similar to those in 1996 and 1997, so differences across time are not related to DAF selection criteria.

¹² In what follows, we consider positive values versus zero values because we expect that the variables of interest should be positive (earnings, benefits due). In actuality, correcting the processing error also could have replaced zero values with negative values, had those been in the data.

Figure VII.1. Percentage increase in the number of beneficiaries with a positive value after correcting the SSI-LF processing error, as a share of beneficiaries with a positive value before correcting the error



Source: DAF17, using the corrected method of processing SSI-LF records compared to the method in place in DAF16 and earlier.

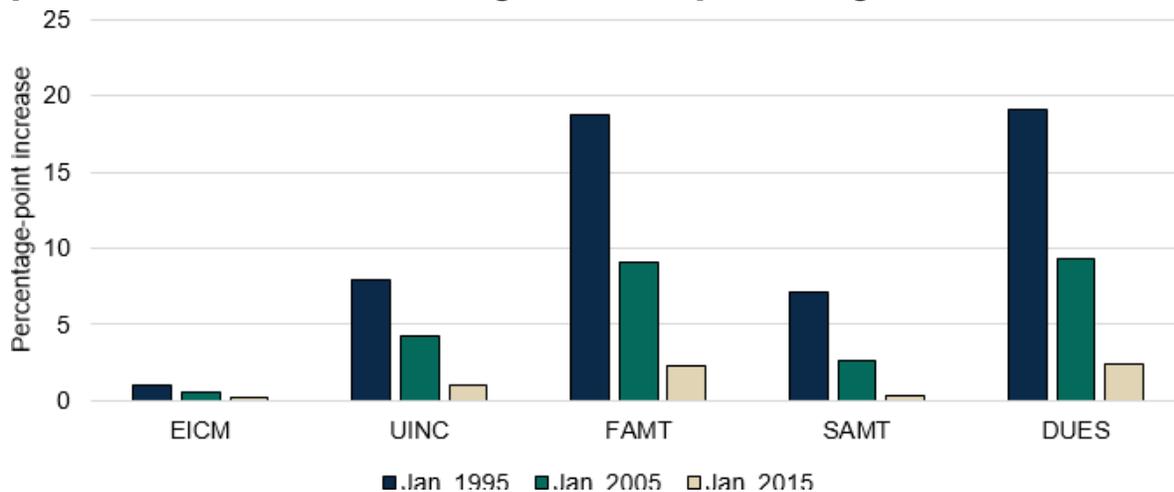
Note: Positive values summed across all beneficiaries in the DAF in the month shown, with the percentage change calculated by comparing the DAF17 corrected value to the share with positive values before reprocessing.

The records that changed as a share of positive records before the correction, shown in Figure VII.1, may not give a complete picture of the magnitude of the effect on beneficiary-level statistics, as some variables in the administrative records apply to virtually all beneficiaries, while others apply to very few. To consider this, we calculated the number of beneficiaries whose value on a given variable changed from a zero to a positive value after reprocessing, as a share of the number of beneficiaries receiving benefits in a given month (Figure VII.2). We calculated this value before and after reprocessing. We present the difference in those values—the percentage-point change—that results from the correction in the share of beneficiaries with a positive value among those who are also in non-terminated SSI payment status (*PSTAYymm*). For convenience we will refer to the set of beneficiaries with non-terminated SSI payment status as “SSI beneficiaries.”

Although the increase in the percentage increase in positive values does not vary greatly across affected variables (Figure VII.1), the change in the share of beneficiaries varies widely, as

shown in Figure VII.2. This variation is a function of the number of SSI beneficiaries who had a positive value to begin with. For example, relatively few beneficiaries have positive values of EICM, so a 31 percent increase in the number of records with a positive value translates to a one percentage-point increase in the share of SSI beneficiaries with a positive value (from 3 to 4 percent). In contrast, the 32 percent increase in the number of records with a positive value of FAMT in January 1995 represents a 19 percentage-point increase in the share of SSI beneficiaries with a positive value (from 59 to 78 percent). The correction also resulted in a similar increase in SSI beneficiaries with a positive value for DUES in January 1995 (consistent with DUES being the sum of FAMT and SAMT). In January 2005, the increases for these variables were about nine percent, while in January of 2015 the increase was two percent as a share of all SSI beneficiaries. The percentage-point increases for UNIC and SAMT were approximately eight, four and one in January 1995, 2005, and 2015, respectively.

Figure VII.2. Percentage-point increase in the share of SSI beneficiaries with positive values after correcting the SSI-LF processing error



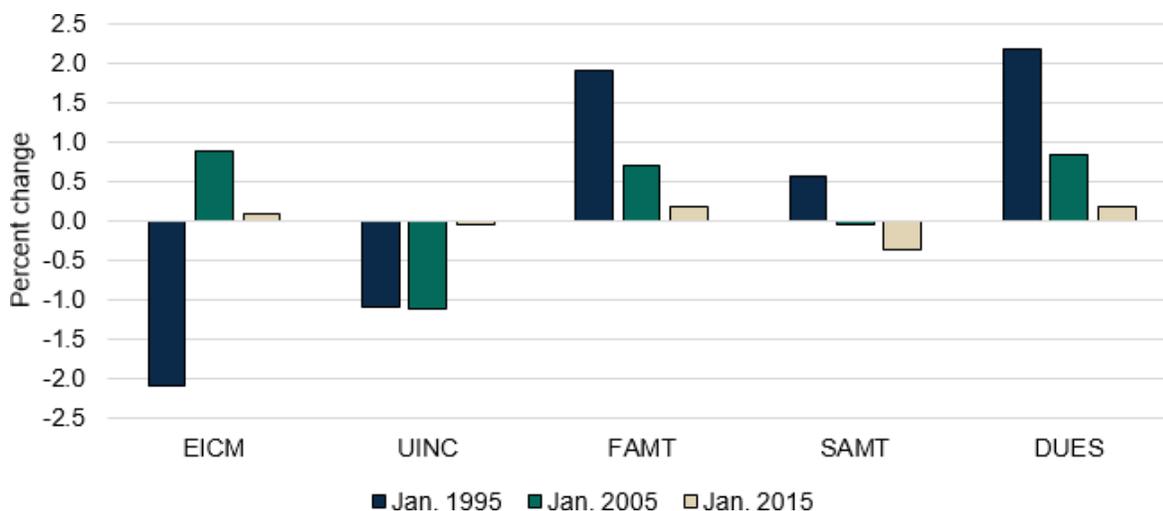
Source: DAF17, using the corrected method of processing SSI-LF records compared to the method in place in DAF16 and earlier.

Note: Percentage point change is calculated by subtracting the share of SSI beneficiaries with a positive value before the correction from the share of SSI beneficiaries with a positive value after the correction.

Average dollar values did not change much after correcting the processing error. Figure VII.3 shows the percent change in the average dollar value of positive observations as a result of the correction. The largest change in the average positive value of the directly-affected variables across all months of the DAF17 (January 1994 through December 2017) was less than 3 percent. In January 1995, the average positive value of the income variables (EICM and UINC) decreased slightly and the average positive value of the benefits due variables (FAMT, SAMT, DUES) increased slightly. These changes indicate that the beneficiaries whose records were affected by the processing error had slightly lower earned and unearned income in January 1995 than beneficiaries whose records were unaffected. Given the inverse relationship between earnings and SSI benefits due, this pattern is consistent with an increased average value of benefits due in the corresponding months.

Figure VII.3 also shows how average values for the affected variables changed in January 2005 and January 2015; the changes are not necessarily uniform over time. In January of 2005, the average positive value of EICM increased by just under 1 percent after the correction whereas the average positive value of UINC decreased by just over 1 percent. The average positive value of FAMT and DUES increased by about 0.7 percent while the average positive value of SAMT was essentially unchanged. The percentage change was less than one-half of a percent for all of the earnings and benefits due variables in January 2015.

Figure VII.3. Percentage change in the average dollar value among observations with positive values after correcting the SSI-LF processing error



Source: DAF17, using the corrected method of processing SSI-LF records compared to the method in place in DAF16 and earlier.

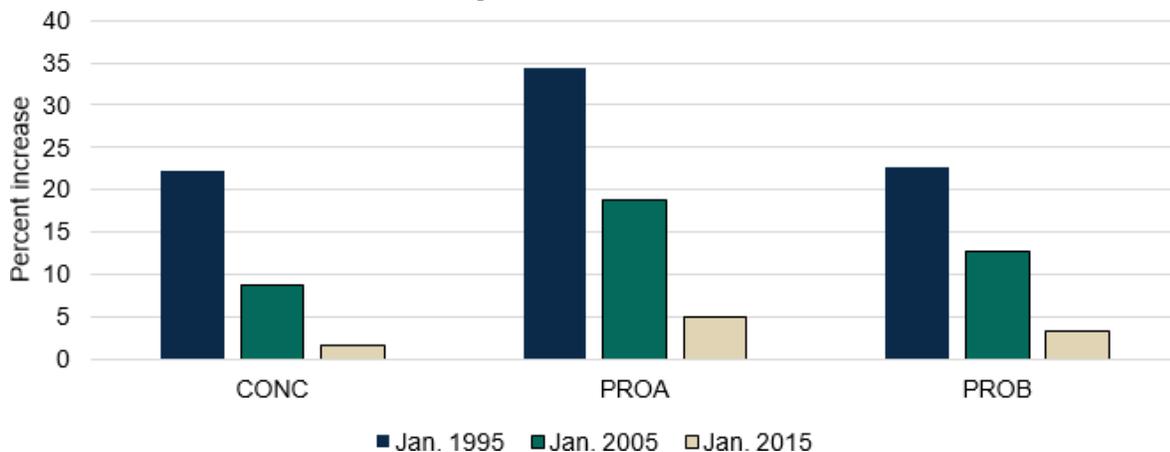
2. Indirectly-affected variables

We use a similar set of figures as those shown in the previous section to describe the effects of the error on CONC, PROA and PROB, before moving in a slightly different direction to discuss the effects on STWSSI, STWCM, BFWSSI and BFWCM.

The share of records affected by the error for indirectly affected variables is similar to the share of core variables that were affected directly. Figure VII.4 shows the percent increase in positive records for the binary variables CONC, PROA, and PROB. In each case, a positive record indicates that the observation has a value of 1 instead of 0. This means that the beneficiary was in the status in a given month—for example, concurrently received SSI and SSDI payments in the month (CONC), was in 1619a status (PROA), or was in 1619b status (PROB). The percentage increase in each time period for the variables shown is of a similar magnitude as the increases shown for the directly-affected variables in Figure VII.1 from which they were constructed. In January 1995 the increase was 22 percent, 34 percent, and 23 percent for CONC, PROA, and PROB, respectively. In January 2005 the increase was about half of the magnitude in

the period ten years prior. By January 2015, the magnitude of the increase was no greater than 5 percent.

Figure VII.4. Percentage increase in the number of beneficiaries with value equal to 1 after the SSI-LF processing error was corrected in DAF17, as a share of beneficiaries with a positive value before the error was corrected

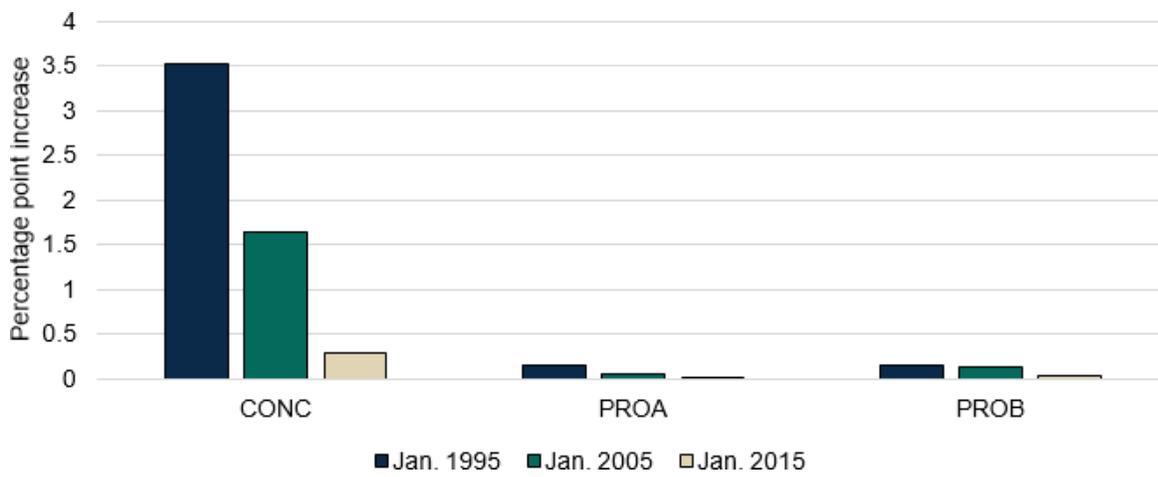


Source: DAF17, using the corrected method of processing SSI-LF records compared to the method in place in DAF16 and earlier.

Note: Positive values summed across all beneficiaries in the DAF in the month shown, with the percentage change calculated by comparing the DAF17 corrected value to the share with positive values before reprocessing.

Figure VII.5 shows the percentage-point change that results from the correction in the share of beneficiaries with a value equal to 1 among those who are also in non-terminated SSI payment status (PSTAyymm). The share of SSI beneficiaries flagged as being concurrently due SSI and SSDI benefits (CONC equal to 1) increased by 3.5 percentage points in January 1995 as a result of the correction. In January 2005, the increase was about 1.5 percentage points and in January 2015, the increase was about one-third of a percentage point. Relatively few beneficiaries have 1619a (PROA equal to 1) or 1619b (PROB equal to 1) status. As a result, even though there was a non-negligible percentage increase in the number of beneficiaries flagged as having either 1619a or 1619b status (Figure VII.4), the change in the share of SSI beneficiaries flagged as 1619a or 1619b is less than two tenths of a percentage point.

Figure VII.5. Percentage-point increase in the share of non-terminated SSI beneficiaries with value equal to 1 after correcting the SSI-LF processing error



Source: DAF17, using the corrected method of processing SSI-LF records compared to the method in place in DAF16 and earlier.

Note: Percentages calculated by dividing the number of SSI beneficiaries with a positive value by the number of beneficiaries in non-terminated SSI payment status (*PSTAyy mm*) in the month shown.

We now describe the effects of reprocessing on variables that measure the effects of beneficiary work activity—STW and BFW. Because SSI benefit amounts are affected by the amount of earned and unearned income that are recorded, the corrected algorithm for processing SSI-LF data resulted in changes to these variables. For users unfamiliar with the STW and BFW concepts and measures, we suggest reviewing Volume 3- Tips for Conducting Analysis with the DAF. In the case of SSI, there are a number of interactions between earned and unearned income that make it difficult to predict the effect of the error on the STW and BFW measures.

In Table VII.2, we show the distributions in each month under the old and corrected processing methods. Because most SSI beneficiaries do not have earnings that would result in STW, the overall distribution of STWSSI looks quite similar before and after the correction. Yet, comparing the change between the old (incorrect) and new (correct) processing algorithms highlights patterns that echo the earlier findings: the affected values are among those who had earned or unearned income (STWSSI=1,2,3,4) and the magnitude of the change increases the

farther back in time one goes. These statistics also show that at each point, correcting the error resulted in an increase in the number of SSI beneficiaries with an STWSSI value indicating suspense or termination of cash benefits due to work (STWSSI=1,2,3), an increase in the share with earnings, but whose unearned income alone led to benefit suspense or termination (STWSSI=4), and a decrease in the share who were in suspense or termination status for a reason not determined to be work (STW=8).

We present the effects on BFWSSI in Figures VII.6 and VII.7, following a similar structure to earlier figures. Because the error affected earnings and SSI benefits are reduced by \$1 for every \$2 of earnings above an income disregard, we would expect to see BFWSSI increase with the corrected algorithm. However, because STWSSI=4 cases (countable unearned income was sufficient to cause suspension) also increased, and because those cases do not accrue BFWSSI, we would not expect the magnitude to be as large as for earnings alone.

In Figure VII.6, we show the percentage increase in the number of beneficiaries with a positive BFWSSI value after the correction and the percentage increase when limiting to beneficiaries in STWSSI=0,1,2,3 (the only groups who can accrue BFWSSI).¹³ As a share of the beneficiaries with positive BFW in a month, correcting the error increased positive BFWSSI by 33 percent in January 1995, 15 percent in January 2005, and 4 percent in January 2015. Relative to the number of SSI beneficiaries in current pay status or with benefits suspended or terminated for work in a given month, however, the magnitude of the change was 1 percent or less, reflecting the fact that relatively few SSI beneficiaries accrue BFWSSI in a given month.

¹³ Because the STWSSI values and BFWSSI values changed as a result of correcting the error, we used a denominator for this calculation that limited to beneficiaries who had STWSSI=0,1,2,3 under the old processing method. As shown in Table 2, the share with STW=0,1,2,3 increased overall, so had we instead used the corrected STWSSI values as a denominator, this share would have been smaller.

Figure VII.7 shows that average BFWSSI increased by 0.5 to 3.5 percent in the months shown, without a clear pattern over time. This suggests that the BFWSSI that was erroneously omitted tended to be slightly higher than the BFWSSI for beneficiaries whose values were included.

In the appendix, we present statistics for both STWSSI and BFWSSI, and the combined version of both variables (STWCM and BFWCM). The magnitude of the change for the latter is smaller than for the former, which is expected based on their construction. The combined variables take into account beneficiaries' status in both programs, when relevant. In the case of STW, STWCM errs toward current pay status if a beneficiary is in current pay status in one program and not the other. In the case of BFWCM, BFWSSI and BFWDI are summed together, with the former generally representing much smaller dollar values than the latter, per program rules.

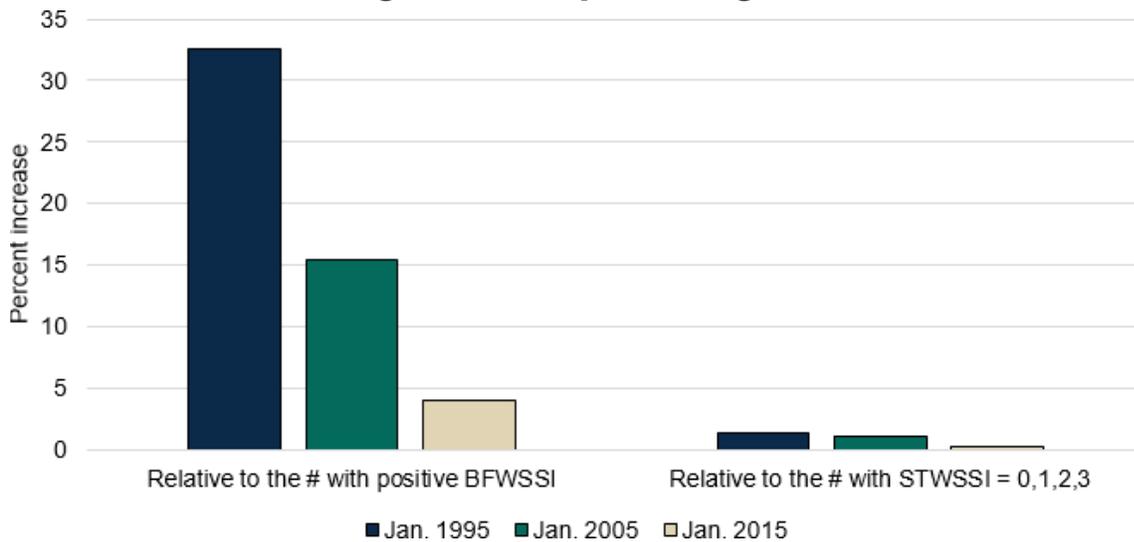
Table VII.2. Percentage distribution of STWSSI values before and after correcting the SSI-LF processing error

	January 1995- Old	January 1995- New	Percent change (New vs. Old).	January 2005- Old	January 2005- New	Percent change (New vs. Old).	January 2015- Old	January 2015- New	Percent change (New vs. Old).
STWSSI=0	74.5	74.5	0%	54.6	54.6	0%	49.9	49.9	0%
STWSSI=1,2,3	0.7	0.9	32%	1.5	1.7	18%	1.3	1.4	4%
STWSSI=4	0.1	0.2	37%	0.3	0.4	8%	0.3	0.3	2%
STWSSI=8	17.1	16.9	-1%	16.4	16.2	-1%	9.8	9.8	0%
STWSSI=9	7.5	7.5	0%	27.2	27.1	0%	38.6	38.6	0%

Source: DAF17, using the corrected method of processing SSI-LF records compared to the method in place in DAF16 and earlier.

Note: Percent change is calculated as a share of the pre-correction distribution. The values shown in the table are rounded to the nearest tenth, but the percentage change was calculated prior to rounding.

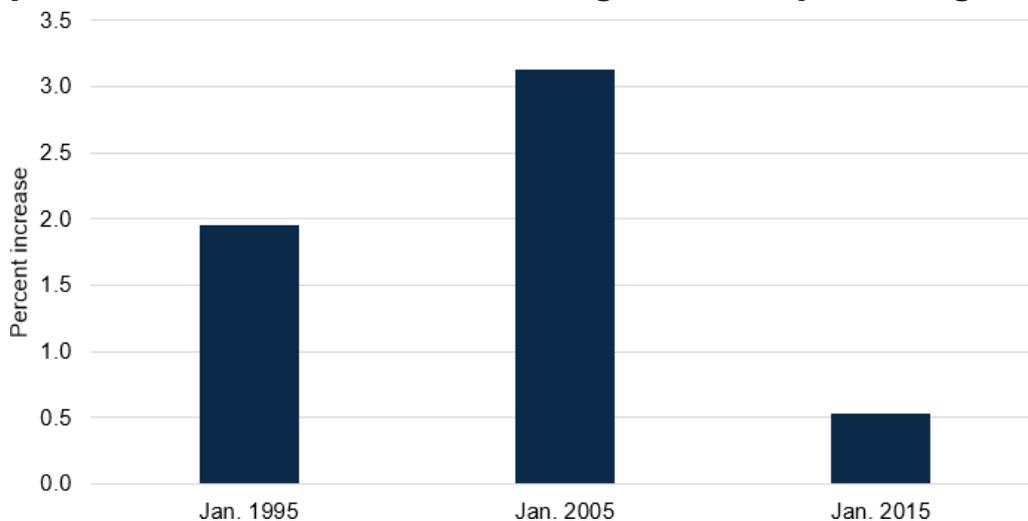
Figure VII.6. Percentage increase in the number of beneficiaries with positive BFWSSI after correcting the SSI-LF processing error



Source: DAF17, using the corrected method of processing SSI-LF records compared to the method in place in DAF16 and earlier.

Note: Positive values summed across all beneficiaries in the DAF in the month shown, with the percentage change calculated by comparing the DAF17 corrected value to the share with positive values before reprocessing.

Figure VII.7. Percentage change in the average dollar value of BFWSSI among positive observations after correcting the SSI-LF processing error



Source: DAF17, using the corrected method of processing SSI-LF records compared to the method in place in DAF16 and earlier.

E. Characteristics of beneficiaries with affected records

Because of the nature of the reprocessing error, beneficiaries with certain characteristics were more likely to have affected records. In this section, we present statistics on the distribution

of sex, age, years since SSI initial eligibility, and primary disabling condition for beneficiaries whose FAMT and EICM were affected by the error. Because other affected variables are closely related to FAMT and EICM, we would expect similar patterns for those variables, but did not include comparisons here for ease of presentation. Again, we present statistics from January of 1995, 2005, and 2015.

Table VII.3 shows the characteristics of beneficiaries whose FAMT records were affected by the processing error compared to those whose records were not affected. The “affected” category consists of beneficiaries for whom the processing error caused a positive value to be set to zero. The “not affected” category consists of beneficiaries with a *positive value* of FAMT both before and after the correction. The largest difference between the two groups is age. On average, beneficiaries whose records were affected are younger than those whose records were not. In particular, a much larger share of beneficiaries whose records were affected are under age 18 (as of January 1995, 2005, or 2015). As noted previously, a new record is often established if a child SSI recipient becomes an adult SSI recipient. Beneficiaries who are under 18 are more likely than older beneficiaries to have been affected by the error because, all else equal, they are more likely to have had a new record established prior to the construction of DAF17. Given the difference in age, it follows that the affected group has fewer years of SSI eligibility on average (at the time they are observed in 1995¹⁴, 2005, or 2015). Other notable differences are that the affected group has a larger share of men (by about 7 percentage points) and a larger share of

¹⁴ The share with missing values of SSIELIG_FIRST in January 1995 is higher than in January 2005 and 2015, and is higher for the group whose records were not affected than for the group whose records were affected. An investigation across other years (not shown) showed that the pattern changed in 2004. In 2004, there were changes in the source files used for SSI and in the PSTA codes that resulted in selection into the DAF. Additionally, in late 2003, there were changes in SSA’s use of some PSTA codes; together these changes mean that the variable we used to define years since SSI eligibility was not as frequently populated for DAF beneficiaries prior to 2004 as it has been since that time. As a result, the distribution on SSIELIG_FIRST in January 1995 is not directly comparable to the later years.

people with diagnoses in group 1 (autistic disorders; developmental disorders; or childhood and adolescent disorders not elsewhere classified), especially in 2005 and 2015.

Table VII.4 shows the characteristics of beneficiaries whose EICM records were affected by the processing error compared to those whose records were not affected. As in Table VII.3, the “affected” category consists of beneficiaries for whom the processing error caused a positive value to be set to zero and the “not affected” category consists of beneficiaries with a positive value both before and after the correction. Because many SSI beneficiaries do not have earned income (and because EICM is an input into FAMT), this table consists of a subset of the beneficiaries described in Table VII.3. Beneficiaries whose EICM records were affected are younger on average but the difference is smaller than for FAMT and it is primarily in the 18 to 29 age group instead of the under-18 group. This is expected given that earned income is rare for SSI beneficiaries under the age of 18. Beneficiaries whose records were affected have slightly fewer years of SSI eligibility on average and there is little difference in the gender distribution in the two groups. For both groups of beneficiaries, the modal diagnosis group is intellectual disability but the share of beneficiaries with an intellectual disability is larger for beneficiaries whose records were affected by the error than for those whose records were not.

Table VII.3. Comparison of beneficiary characteristics for those whose FAMT records were or were not affected by the SSI-LF processing error

	January 1995		January 2005		January 2015	
	Not affected	Affected	Not affected	Affected	Not affected	Affected
Observations	3,201,186	1,012,129	4,647,120	643,166	6,322,528	185,418
Percent female	52.8	46.0	53.3	45.9	49.2	43.4
Age						
Mean	37.5	27.4	39.0	26.9	38.7	26.4
18-29 (%)	16.6	41.7	17.6	45.6	20.5	49.9
30-39 (%)	15.8	14.0	13.4	14.8	14.6	13.6
40-49 (%)	20.0	15.8	14.0	10.2	11.7	9.6
50-59 (%)	19.0	14.8	22.3	13.0	14.8	8.2
60-FRA (%)	20.6	10.3	22.7	11.2	26.2	11.9
Years since SSI award						
Missing (%)	6.1	0.0	0.0	0.0	0.0	0.0
Mean	7.0	6.0	10.0	7.6	11.7	9.1
<1 (%)	10.6	15.2	9.9	13.1	7.7	8.1
1-4 (%)	37.2	50.0	25.5	35.4	24.7	33.7
5-9 (%)	21.3	14.0	21.1	18.9	21.4	24.8
10-19 (%)	19.0	11.7	30.3	26.1	24.8	16.6
20+ (%)	5.9	9.1	13.2	6.6	21.5	16.8
Disabling condition (%)						
Missing	31.6	26.6	13.4	10.2	13.0	8.0
1. Autistic disorders; developmental disorders; or childhood and adolescent disorders not elsewhere classified	1.8	5.2	6.1	16.1	12.1	26.4
2. Intellectual disability	17.3	24.6	19.2	20.8	15.6	14.8
3. Mood disorders; organic mental disorders; schizophrenic and other psychotic disorders; or other mental disorders	19.3	18.5	26.9	23.7	26.2	20.0

Table VII.3. Comparison of beneficiary characteristics for those whose FAMT records were or were not affected by the SSI-LF processing error

	January 1995		January 2005		January 2015	
	Not affected	Affected	Not affected	Affected	Not affected	Affected
4. Endocrine, nutritional, and metabolic diseases; circulatory system; digestive system; genitourinary system; nervous system and sense organs; or respiratory system	15.7	11.0	17.6	13.6	15.7	14.0
5. Musculoskeletal system and connective tissue	5.3	2.6	8.0	4.6	9.3	5.6
6. Infectious and parasitic diseases or injuries	3.7	2.6	4.2	3.2	3.9	3.1
7. Congenital anomalies; blood and blood-forming organs; skin and subcutaneous tissue; or other	1.9	3.8	2.9	6.2	3.5	7.4
8. Unknown value	3.5	5.1	1.7	1.5	0.8	0.7

Source: DAF17, using the corrected method of processing SSI-LF records compared to the method in place in DAF16 and earlier.

Table VII.4. Comparison of beneficiary characteristics for those whose EICM records were or were not affected by the SSI-LF processing error

	January 1995		January 2005		January 2015	
	Not affected	Affected	Not affected	Affected	Not affected	Affected
Observations	175,859	55,017	284,067	41,260	333,376	12,690
Percent female	44.8	45.9	49.7	51.4	49.1	48.7
Age						
Mean	35.3	34.3	38.4	34.9	37.9	34.3
18-29 (%)	2.2	3.9	2.0	3.8	2.1	3.3
30-39 (%)	36.0	36.8	29.0	39.6	35.4	44.9
0-49 (%)	31.4	29.7	24.3	22.1	21.5	24.2
50-59 (%)	17.2	19.3	24.2	20.6	16.0	12.4
60-FRA (%)	10.5	8.4	16.0	10.8	18.8	11.5
Years since SSI award						
Missing (%)	19.0	0.2	0.5	0.5	0.4	0.8
Mean	9.4	8.6	12.2	9.7	13.6	13.1
<1 (%)	6.4	7.2	9.6	9.3	8.1	5.7
1-4 (%)	22.2	35.6	17.6	21.0	18.4	14.3
5-9 (%)	18.5	21.7	17.0	21.1	17.8	20.2
10-19 (%)	26.0	22.9	33.8	39.2	26.3	29.3
20+ (%)	8.0	12.5	21.5	8.9	29.0	29.8
Disabling condition (%)						
Missing	23.4	21.2	7.8	8.4	7.9	7.0
1. Autistic disorders; developmental disorders; or childhood and adolescent disorders not elsewhere classified	0.5	0.6	2.1	3.0	6.7	8.4
2. Intellectual disability	30.6	36.5	32.6	33.8	28.1	34.1
3. Mood disorders; organic mental disorders; schizophrenic and other psychotic disorders; or other mental disorders	20.2	22.2	27.6	31.2	27.5	27.0

Table VII.4. Comparison of beneficiary characteristics for those whose EICM records were or were not affected by the SSI-LF processing error

	January 1995		January 2005		January 2015	
	Not affected	Affected	Not affected	Affected	Not affected	Affected
4. Endocrine, nutritional, and metabolic diseases; circulatory system; digestive system; genitourinary system; nervous system and sense organs; or respiratory system	13.9	9.5	16.0	12.3	14.5	11.8
5. Musculoskeletal system and connective tissue	3.6	2.1	6.4	4.5	8.0	4.8
6. Infectious and parasitic diseases or injuries	3.2	2.9	4.1	4.1	3.9	3.7
7. Congenital anomalies; blood and blood-forming organs; skin and subcutaneous tissue; or other	1.2	0.9	1.8	1.5	2.5	2.8
8. Unknown value	3.5	4.1	1.8	1.3	0.8	0.5

Source: DAF17, using the corrected method of processing SSI-LF records compared to the method in place in DAF16 and earlier.

F. Discussion

Our analysis of the impact of the processing error suggests that across all of the affected variables, the magnitude of the effect was smaller in the years closer to the current DAF year than in years farther in the past. Our checks were limited to comparing DAF17 under the old (incorrect) and new (corrected) processing algorithm, though we would expect to find similar patterns had we performed similar comparisons in earlier versions of the DAF.

Because earlier versions of the DAF with the incorrectly processed SSI-LF data have been used in research and analysis, we think it is important to get a sense of the implications of the change in a broader way than what we have presented here. We continue to work on that analysis and will provide additional details as they become available. We expect to produce: (1) statistics similar to those in this memo for beneficiaries more likely to have had their records affected by the error or of strong policy interest, including young adults ages 18-29, (2) cross-sectional and cohort statistics for groups defined by age, and (3) statistics that look at cohorts of new awardees over time.

VI. ASSESSING DATA NEEDS FOR DAF18

A. Overview of the process

The first step in construction of each version of the DAF is to assess what changes, if any, will be made to the database to improve its overall utility. To do this, Mathematica reviews ad hoc requests for new data items made since construction of the previous DAF, helps SSA conduct a survey of DAF users, and prepares recommendations for SSA regarding which changes should be adopted. The assessment process usually results in a small number of changes each year. Additions to the DAF are highly dependent on the availability of the underlying data and the ease of preparing it for inclusion in the DAF.

B. Process for assessing data needs

The process to assess new DAF needs in each construction cycle involves soliciting input from a number of sources; organizing, tracking, and assessing the input received; and developing an action plan for implementing approved changes.

1. Review ad hoc requests

During each construction cycle for the DAF, researchers from SSA or other organizations request new or revised variables for inclusion in the DAF. Because it is generally difficult to add new variables to the DAF once construction is underway, ad hoc suggestions are logged for later review during the first step of construction of the subsequent DAF database.

2. Survey DAF users

At the beginning of each DAF construction cycle, SSA surveys DAF users to gather suggestions for possible changes to the database. SSA maintains an email distribution list that is used to communicate with users of the DAF, which is continuously updated to include new users. SSA composes an email in consultation with Mathematica for distribution to users asking for their suggestions for the upcoming DAF construction cycle.

3. Timeline

In determining which changes may be feasible and the process that will be used to incorporate new variables into the database, it is important to consider which DAF data source contains the relevant data. Different data sources are processed at different periods during the DAF construction cycle, which begins in January of the calendar year following the DAF year (for example, construction of DAF18 began in February 2019). The ability to incorporate suggestions based on each source file depends on when they are processed for DAF construction relative to when the suggestion was received. Because the source files are used in a predetermined sequence and require different steps to process, Mathematica must determine which requests will be accommodated at varying points during the year. Below is information about the timing for using select SSA data sources during DAF construction.

- **DBAD** – Because assembling the twelve DBAD files used to add new SSDI beneficiaries is the first step in DAF construction, any requests for new data from the DBAD files must be considered as early as possible in the production cycle.
- **CER** – The twelve CERs used to add new SSI beneficiaries to the DAF are assembled at the same time as the DBAD files. Any requests for new data from these files must also be considered as early as possible in the production cycle.
- **MBR** – If Mathematica intends to request additional variables for records selected from the MBR, SSA programmers must modify the custom program to include those variables, and Mathematica must request these modifications from SSA. This is best done early, ideally when the finder request is first submitted, although there is some leeway and it is possible to submit the requests for the custom modifications a little later.
- **SSR** – Mathematica creates a finder file of SSNs to pull data from the SSI-LF. When this is done, all variables available in the SSI-LF are pulled for each record requested by Mathematica, though not all are loaded into the DAF. This means that if additional variables are required, Mathematica can access them directly, without any changes to the code by SSA staff, as must be done for the MBR process.
- **NUMIDENT** – A finder file containing all the SSNs from the finder files created for the MBR and SSR extracts is submitted to SSA for extraction of NUMIDENT data. All standard fields are returned to Mathematica for these SSNs so timing for new data request can wait until the extracts have been returned to Mathematica.
- **Initial medical determinations, appeals, and CDRs (831 & 832/833) files** – These files can be accessed directly by Mathematica programmers and no finder files are required. Processing, including program changes to extract new data items, can begin as soon as the

finder files for the MBR and SSR extracts have been finalized. These files are generally not processed until approximately three months after the DBAD and CER and therefore requests for new variables from them can be finalized later than requests for new variables from the MBR and the SSR.

- **DCF** – These files can be accessed directly by Mathematica programmers and no finder files are required. Processing, including program changes to extract new data items, can begin as soon as the finder files for the MBR and SSR extracts have been finalized. However, because these data files are usually processed for DAF inclusion at approximately the same time as the 831 & 832/833 files, requests for new data from them can be finalized later than requests for new variables from the MBR and the SSR.
- **EN Payment Files** – This file is provided to Mathematica upon request about halfway through construction of the DAF. Currently all data available from this source is carried on the DAF, so any new data request would require a new data source. The investigation into any new source would need to be done very early in the construction process, preferably well before construction begins.
- **VRRMS** – These files are provided to Mathematica upon request about halfway through construction of the DAF. As these files are processed independently from all other DAF processes, requests for changes to the DAF-linkable VRRMS data can be handled substantially later than those for other sources, until about the time the data are requested.
- **RSA** – These files are provided upon request in the fall of each DAF construction year. Mathematica receives all variables available from this file and therefore requests for changes can be handled much later in the DAF construction process. Similar to the EN Payment files, because all variables from this file are already provided, any new source data would need to be considered early in the construction process.
- **MEF** – A finder file containing all the SSNs from the finder files created for the MBR and SSR extracts is submitted to SSA for the creation of the DAF-linkable MEF data file. Changes to or additions of variables need to be considered early in the process and require additional coordination between Mathematica and SSA because the data is not accessible, even after its extraction, to Mathematica staff.

4. Prepare recommendations for SSA

Having collected informal and formal input about ways to improve the DAF user experience, Mathematica compiles a list of all proposed changes and presents recommendations to SSA for review. Each suggested change is considered from a number of points of view, as described below.

First to be considered is SSA's perspective. For example, the RSA data is owned by the Department of Education, not SSA, and therefore any requests to add RSA data to the DAF must take into account the necessity of obtaining permission from the Department of Education.

Next to be considered is the research purpose of the DAF. The suggested additions are examined to determine whether they are appropriate and fit within the scope of the disability research likely to be conducted using the DAF. As the DAF is not intended to replicate all SSA administrative data files, any user requests that would not further the utility of the database for disability research purposes would not be recommended.

The impact on the overall size of the DAF is considered next. For example, a set of monthly variables over multiple years can add significant size to an already large database, making it more unwieldy to work with and potentially creating issues related to storage space.

Finally, Mathematica evaluates the programming resources that would be needed to accomplish each change and to ensure that each set of data processing changes and the accompanying documentation changes could be accomplished with available resources.

APPENDIX A

**MONTHLY GRAPHS OF MEASURES OF CHANGE DUE TO THE SSI
REPROCESSING ERROR IDENTIFIED AND CORRECTED IN DAF17**

This page has been left blank for double-sided copying.

This appendix contains graphs showing various measures of the magnitude of the change in the affected variables due to the error in each month contained in DAF17, from January 1994 through December 2017. There are three sets of graphs: 1) Figures A1-A15 for the continuous variables directly affected as a result of the error (EICMyymm, UINCyymm, FAMTyymm, SAMTyymm, and DUESyymm), 2) Figures A16-A24 for the binary variables that were affected (CONCyymm, PROAyymm, PROByymm), and 3) Figures A25-29 for variables related to STW and BFW (STWSSIyymm, STWCMyymm, BFWSSI_DRAFTyymm, BFWCM_DRAFTyymm).

Figure A.1. EICMyymm: Number of beneficiaries with value > 0

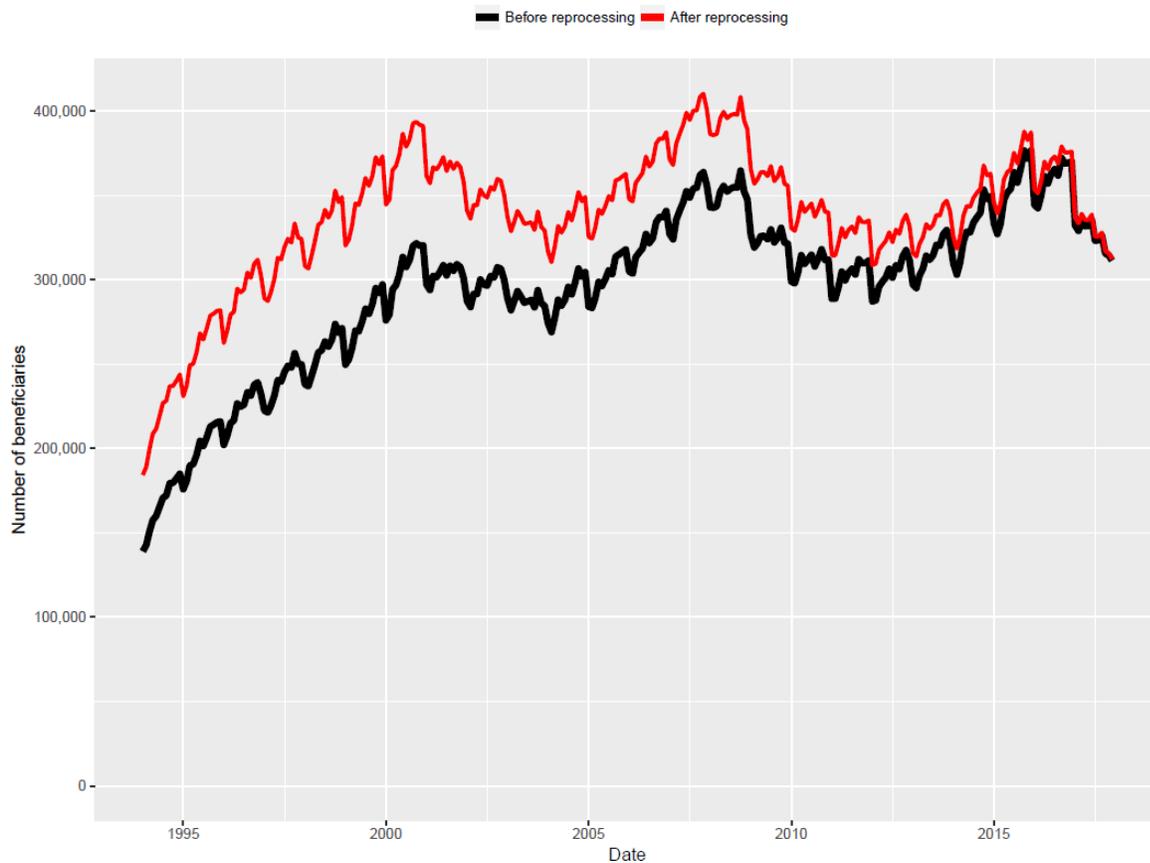


Figure A.2. EICMyymm: Percent of SSI beneficiaries with value > 0

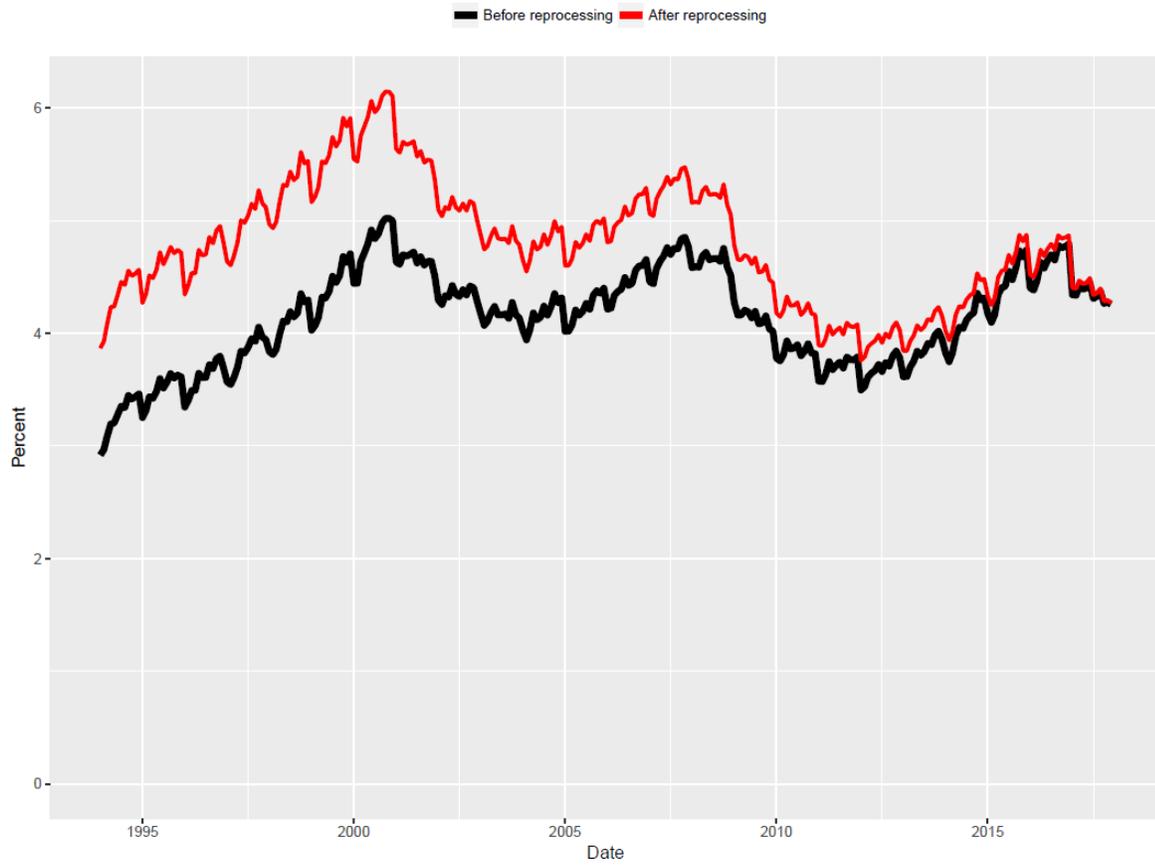


Figure A.3. EICMyymm: Average value among beneficiaries with value > 0

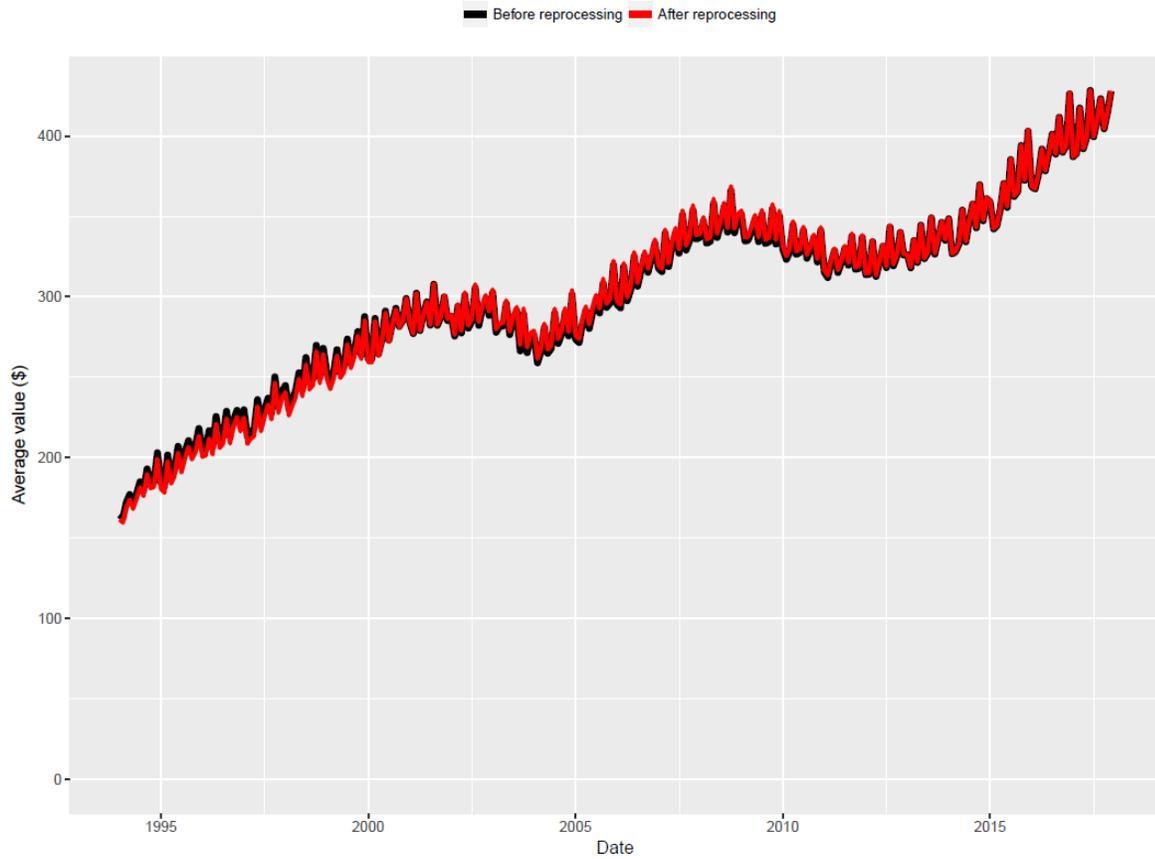


Figure A.4. UINCyymm: Number of beneficiaries with value > 0

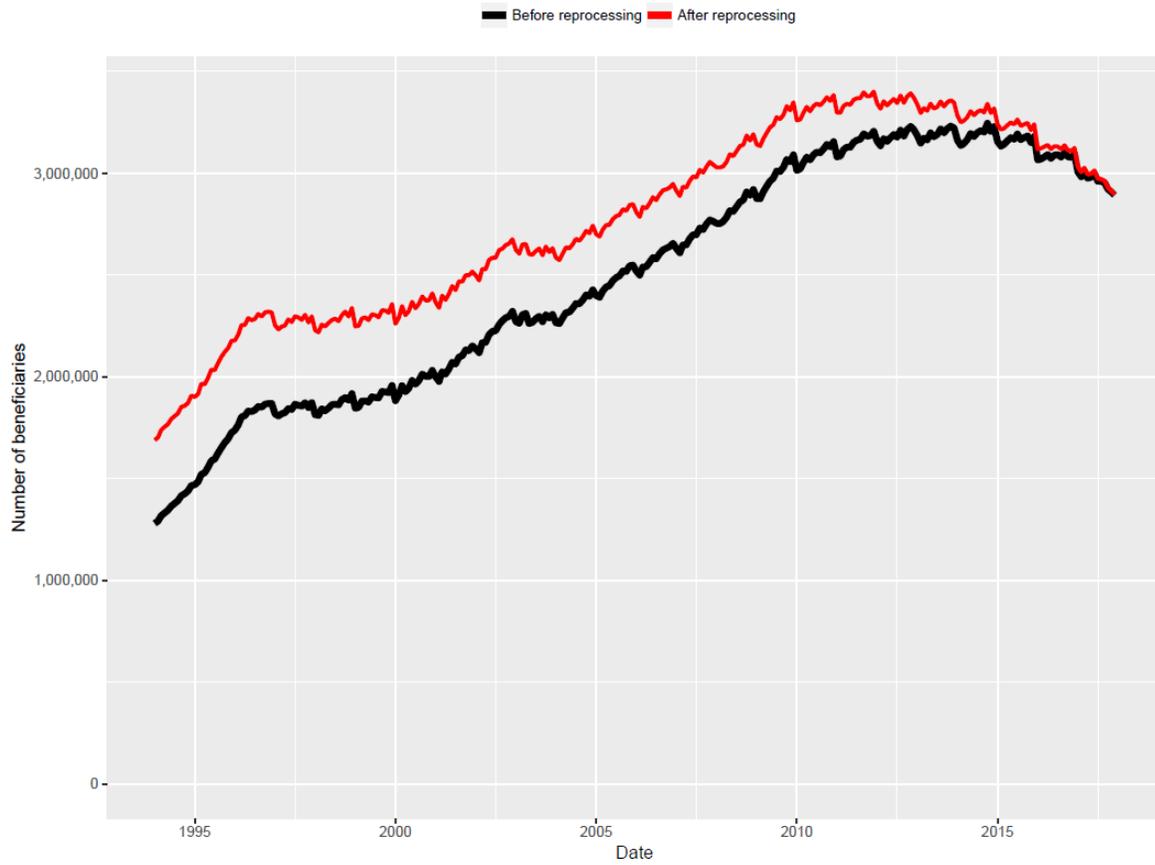


Figure A.5. UINCyymm: Percent of SSI beneficiaries with value > 0

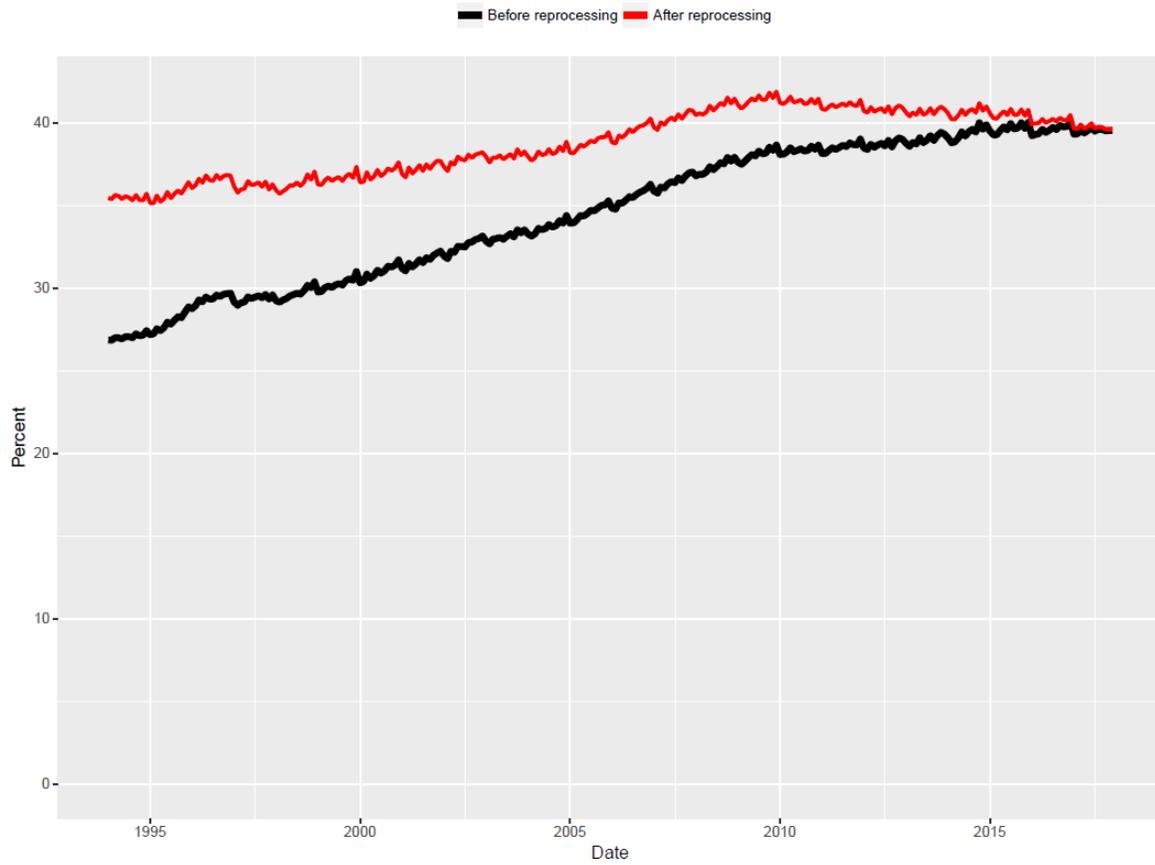


Figure A.6. UINCyymm: Average value among beneficiaries with value > 0

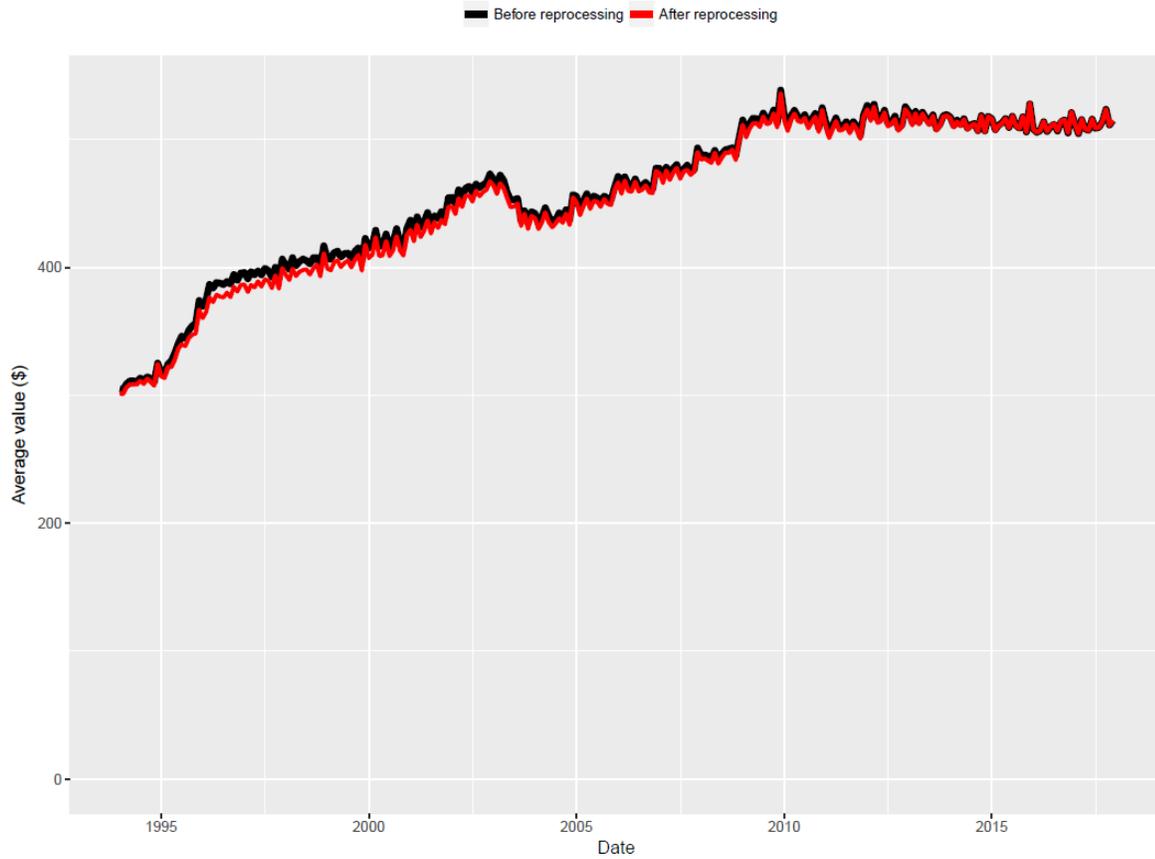


Figure A.7. FAMTyymm: Number of beneficiaries with value > 0

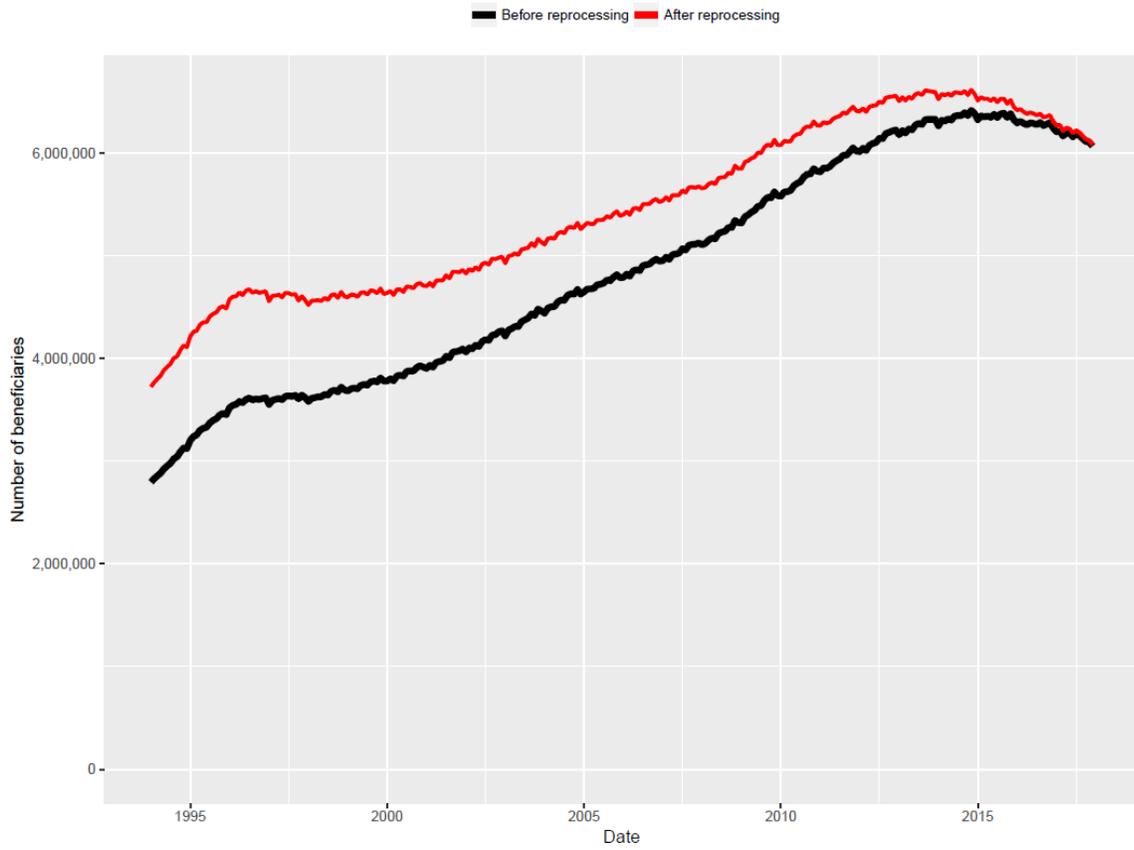


Figure A.8. FAMTyymm: Percent of SSI beneficiaries with value > 0

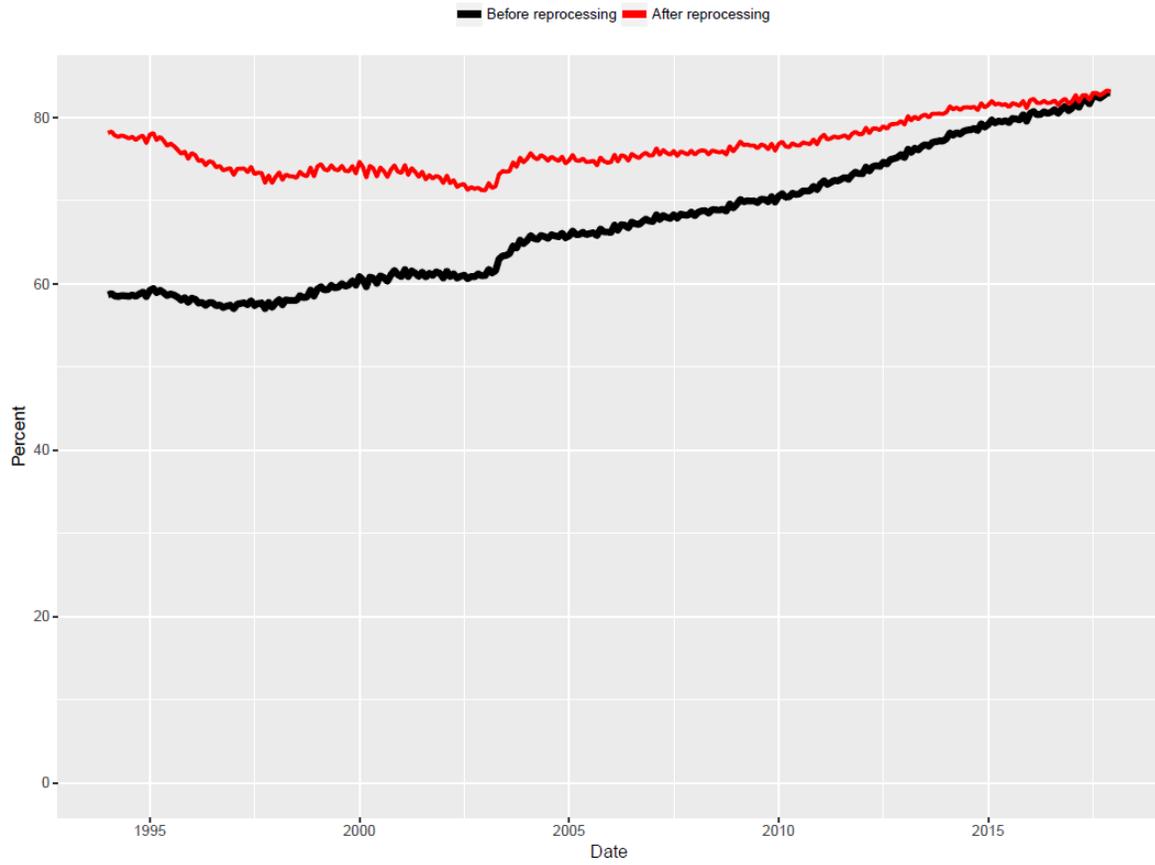


Figure A.9. FAMTyymm: Average value among beneficiaries with value > 0

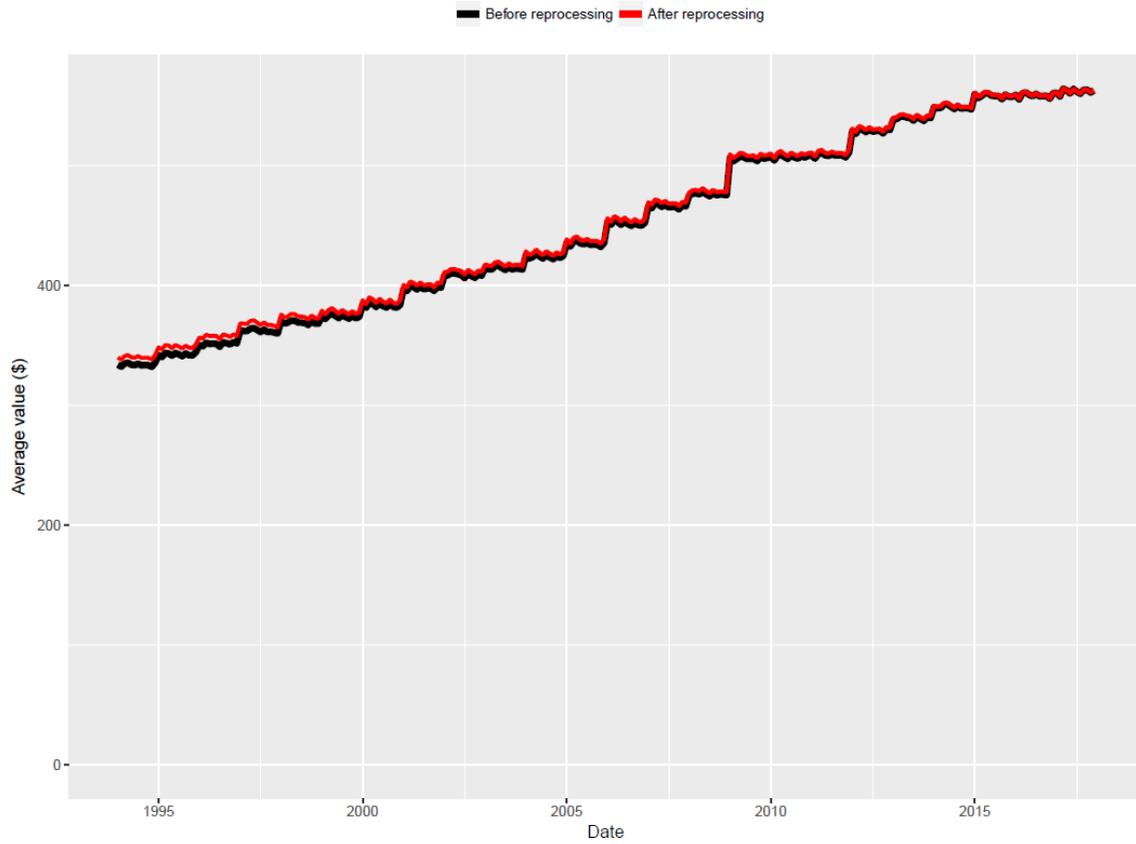


Figure A.10. SAMTyymm: Number of beneficiaries with value > 0

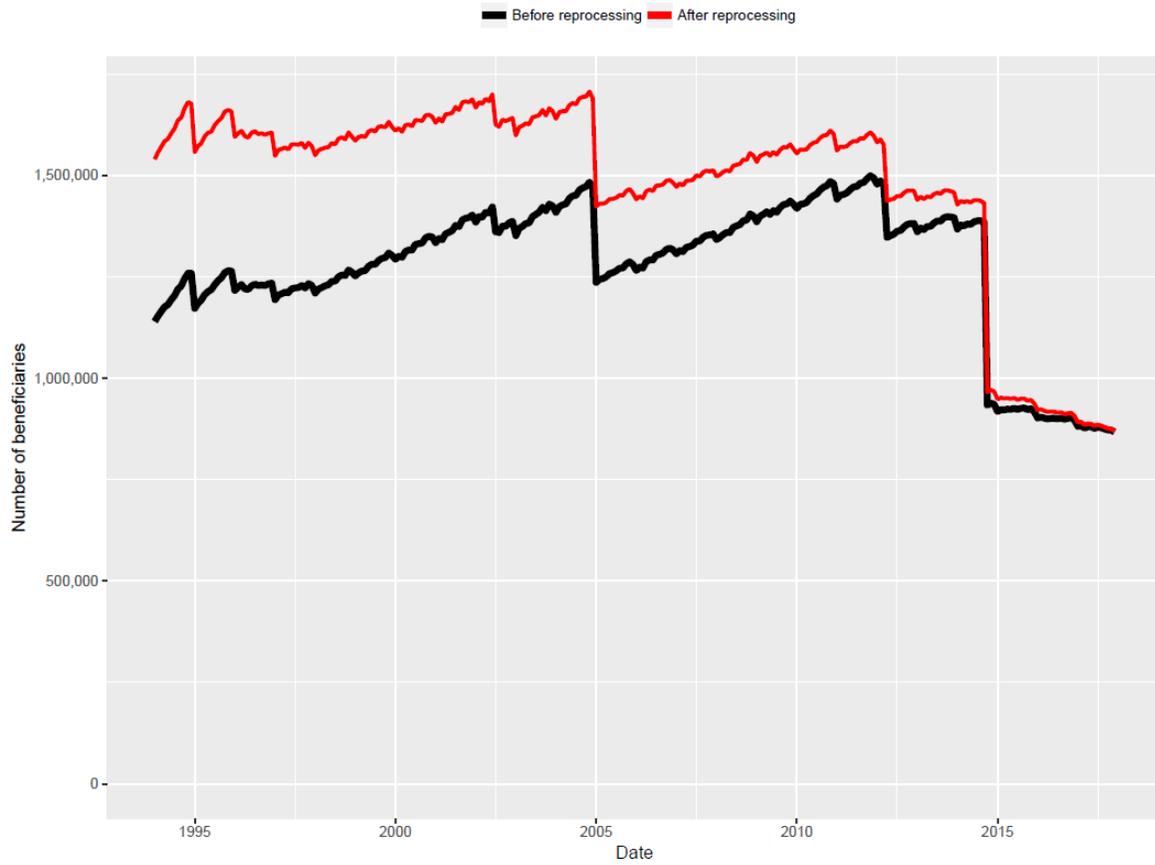


Figure A.11. SAMTyymm: Percent of SSI beneficiaries with value > 0

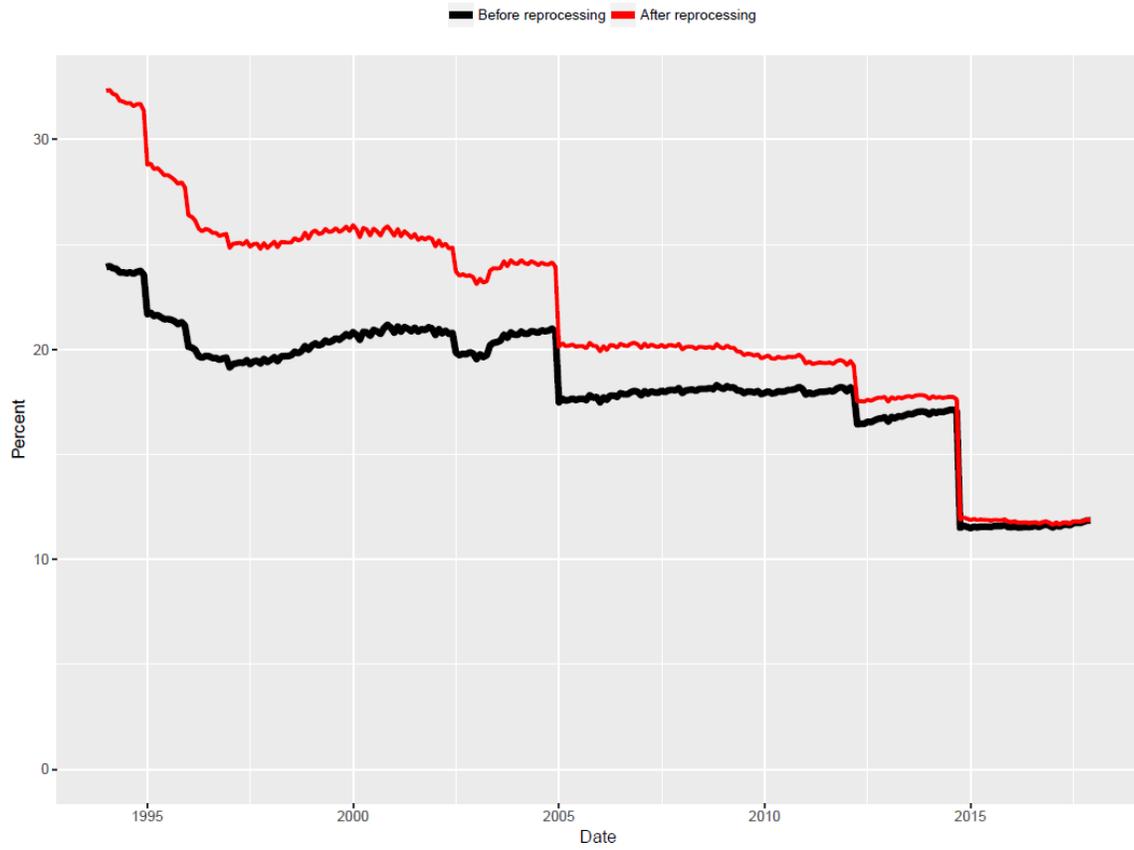


Figure A.12. SAMTyymm: Average value among beneficiaries with value > 0

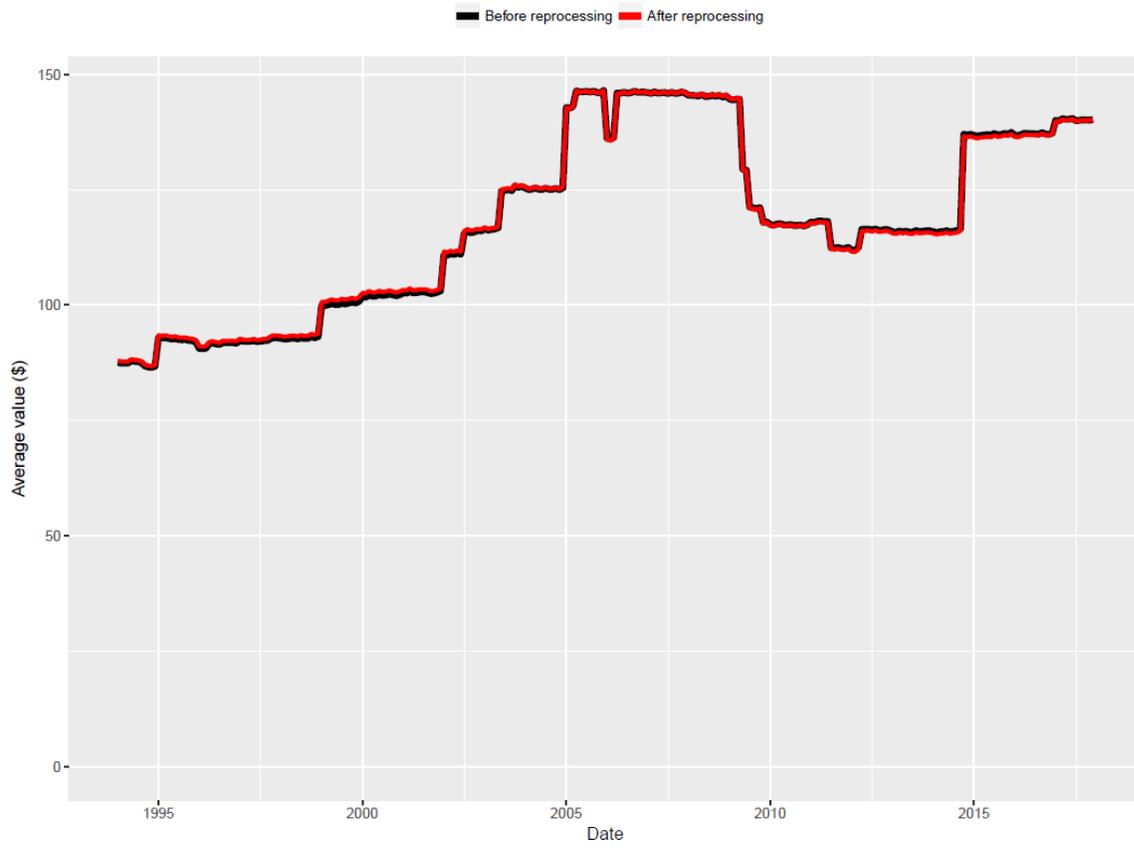


Figure A.13. DUESymm: Number of beneficiaries with value > 0

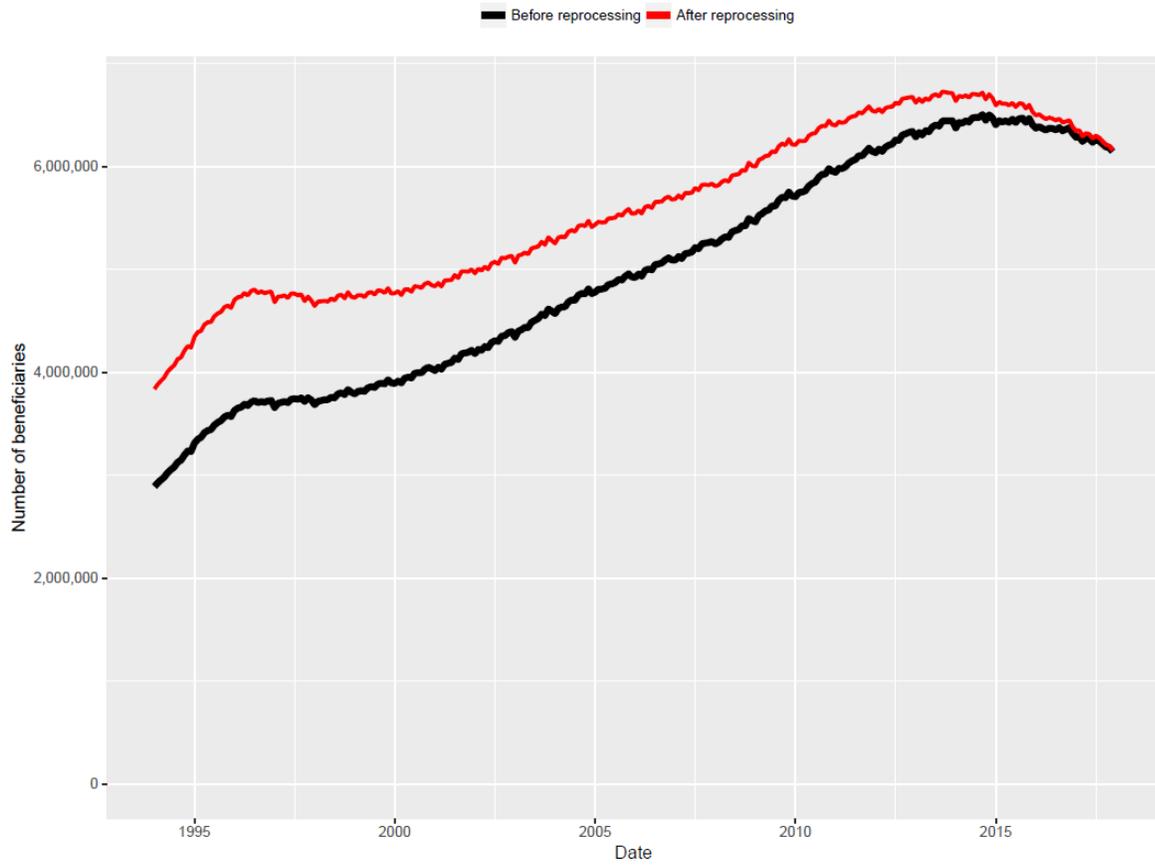


Figure A.14. DUESymm: Percent of SSI beneficiaries with value > 0

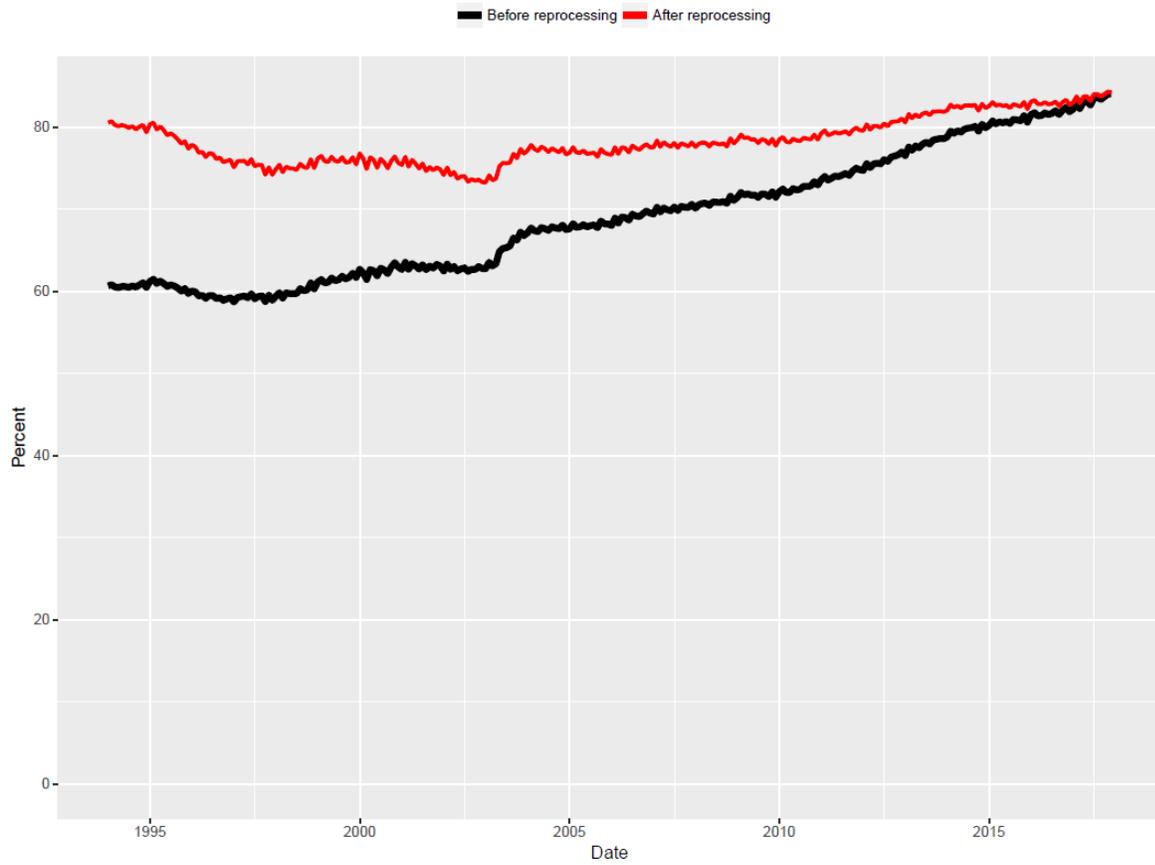


Figure A.15. DUESymm: Average value among beneficiaries with value > 0

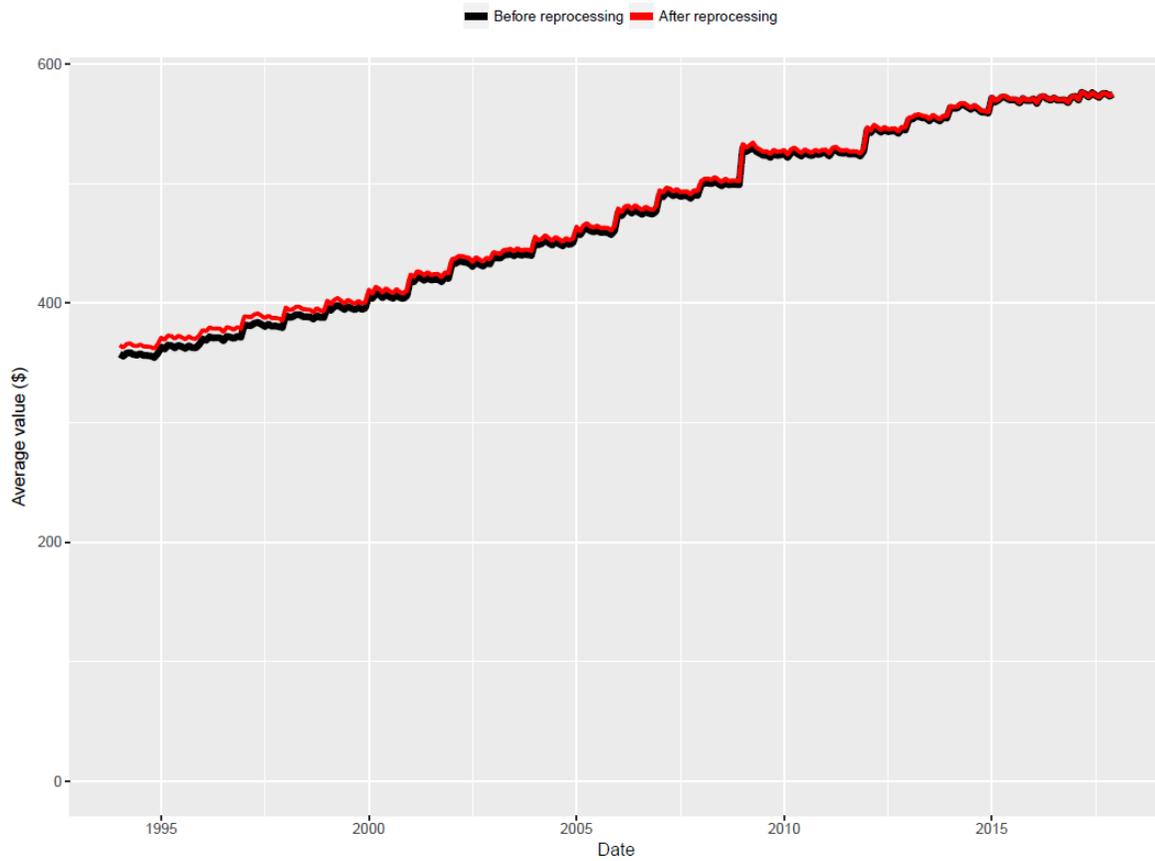


Figure A.16. CONCyymm: Number of beneficiaries with value = 1

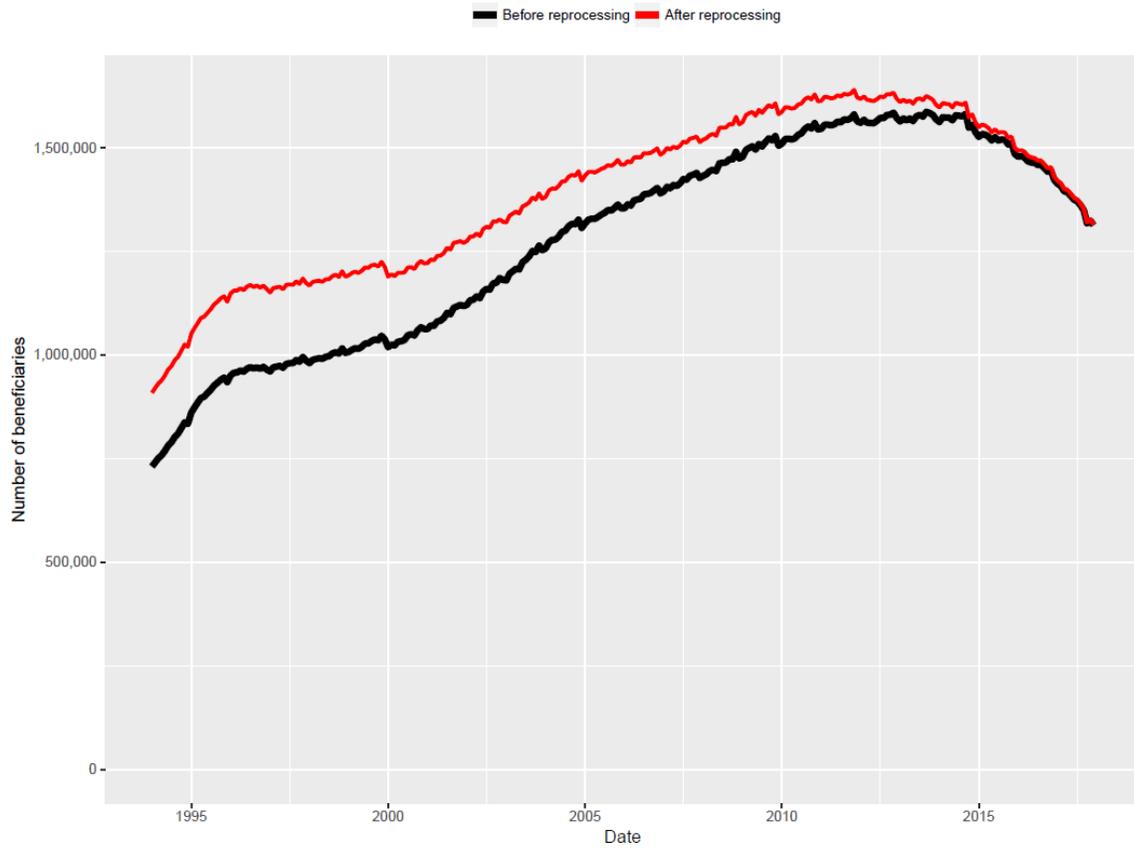


Figure A.17. CONCyymm: Percent of SSI beneficiaries with value = 1

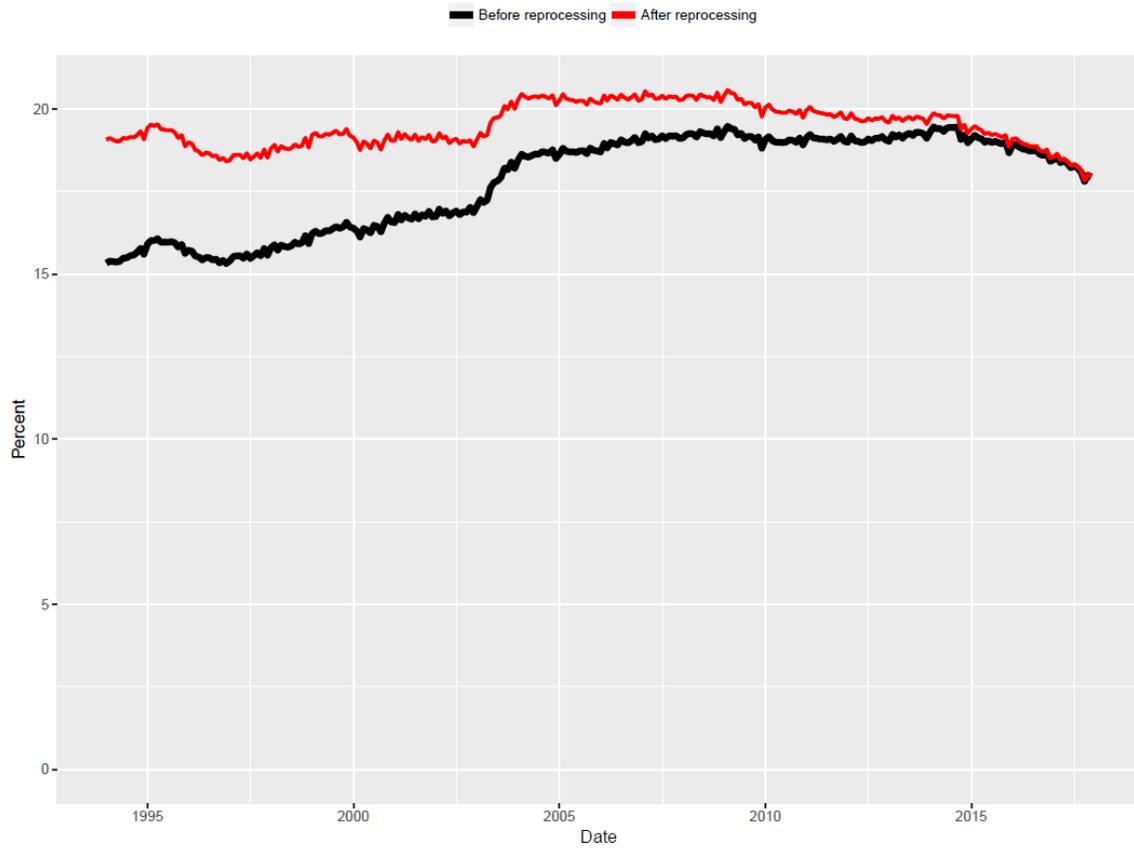


Figure A.18. CONCyymm: Among beneficiaries with populated values, percent with value = 1

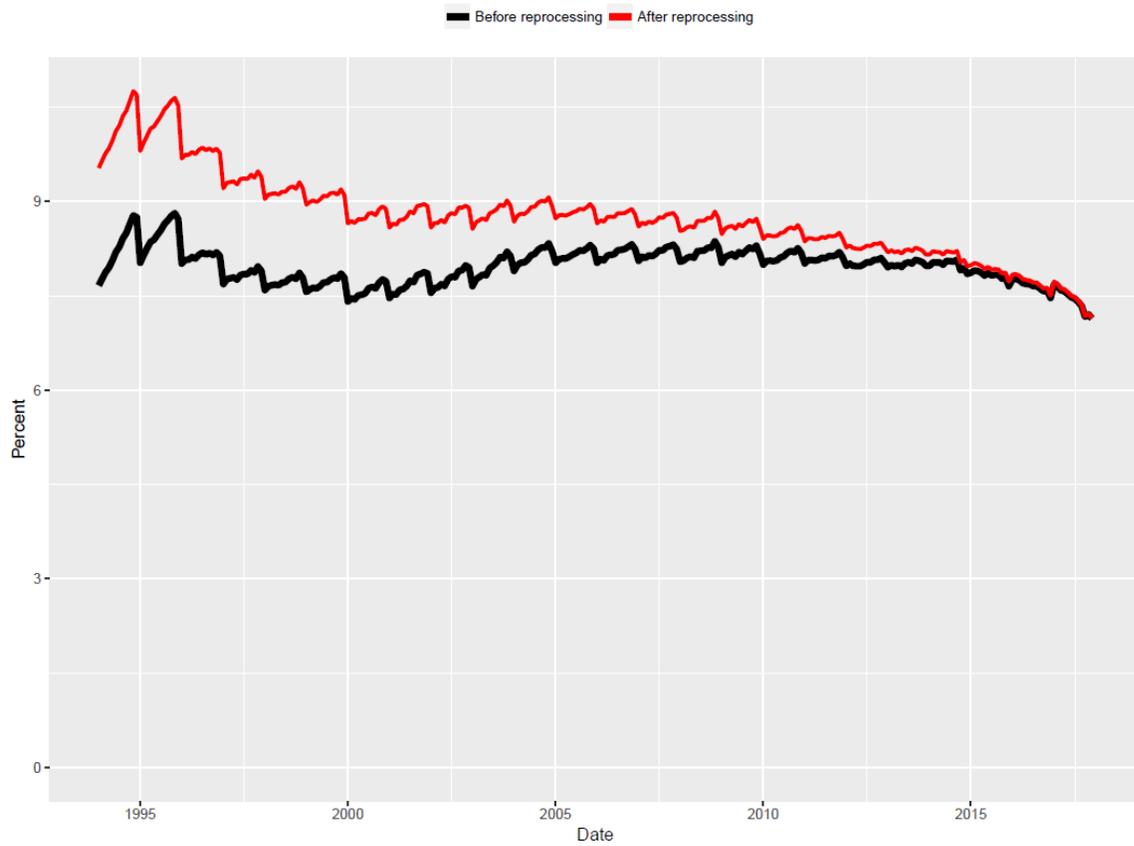


Figure A.19. PROAyymm: Number of beneficiaries with value = 1

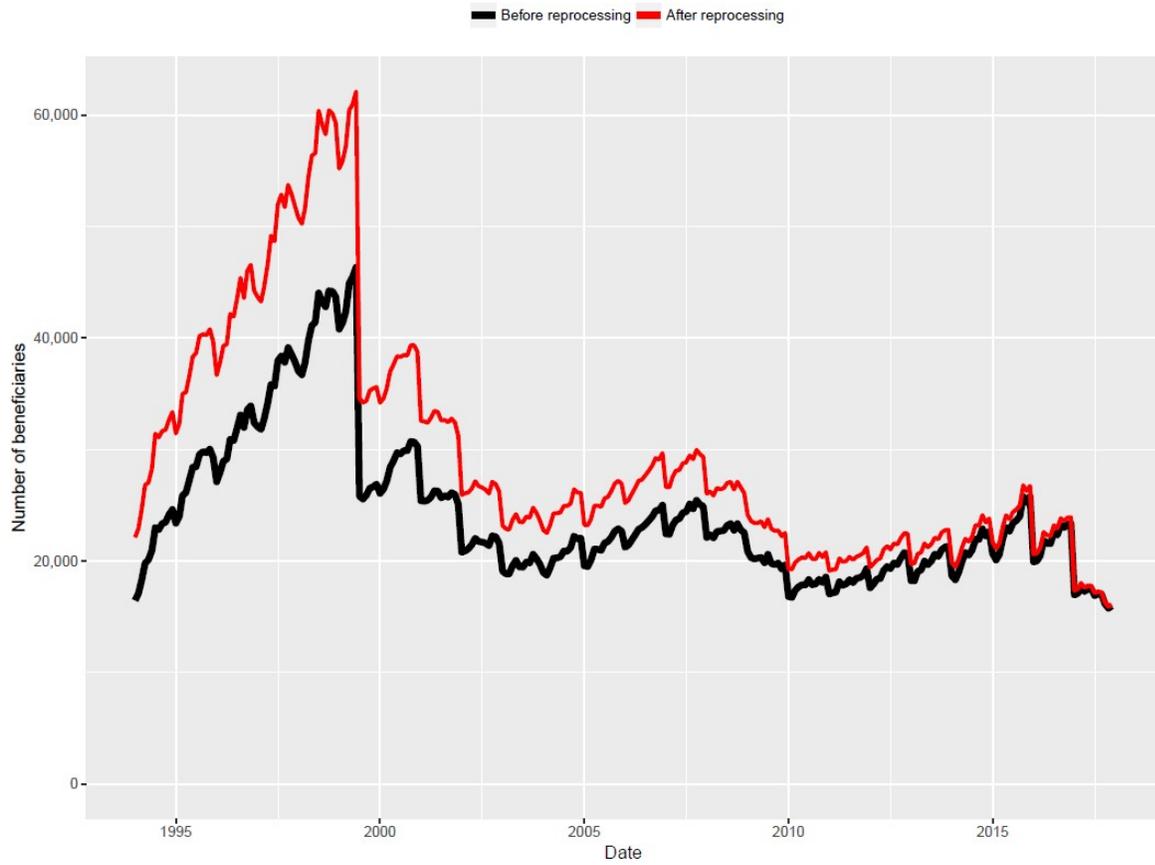


Figure A.20. PROAyymm: Percent of SSI beneficiaries with value = 1

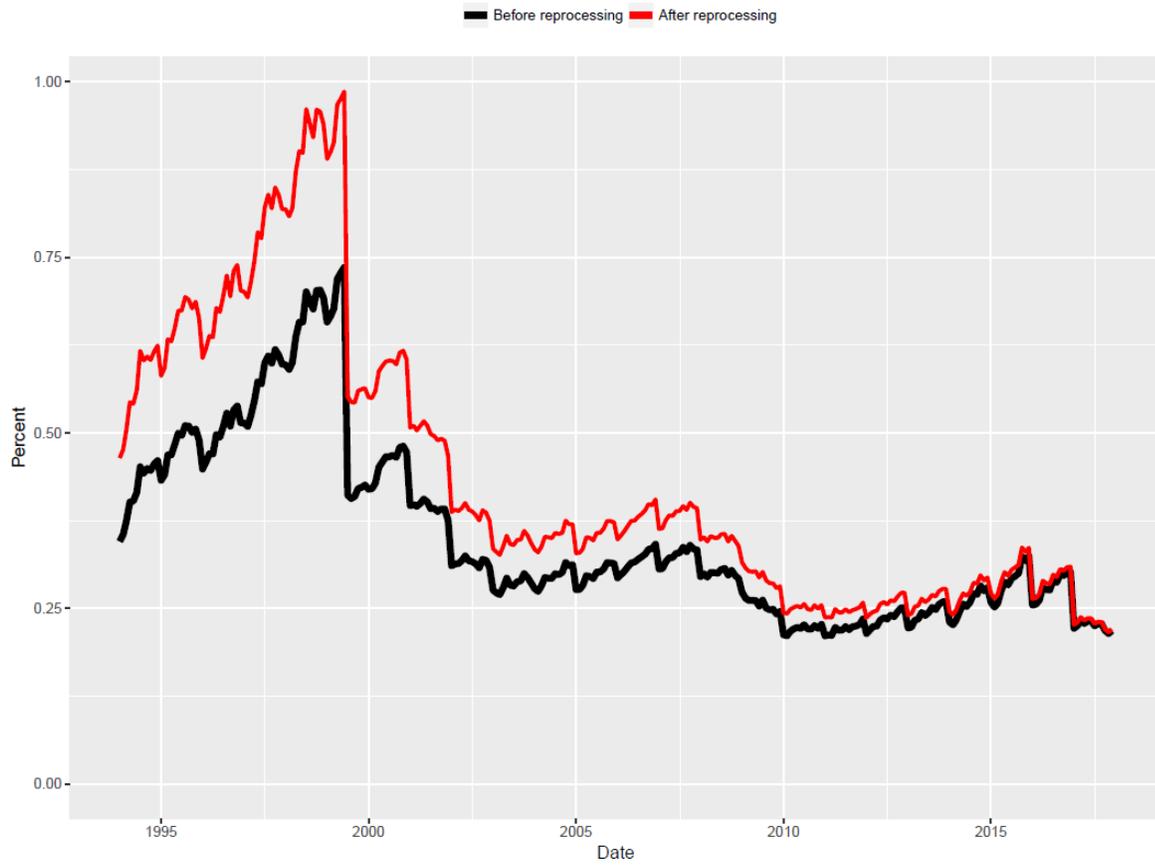


Figure A.21. PROAyymm: Among beneficiaries with populated values, percent with value = 1

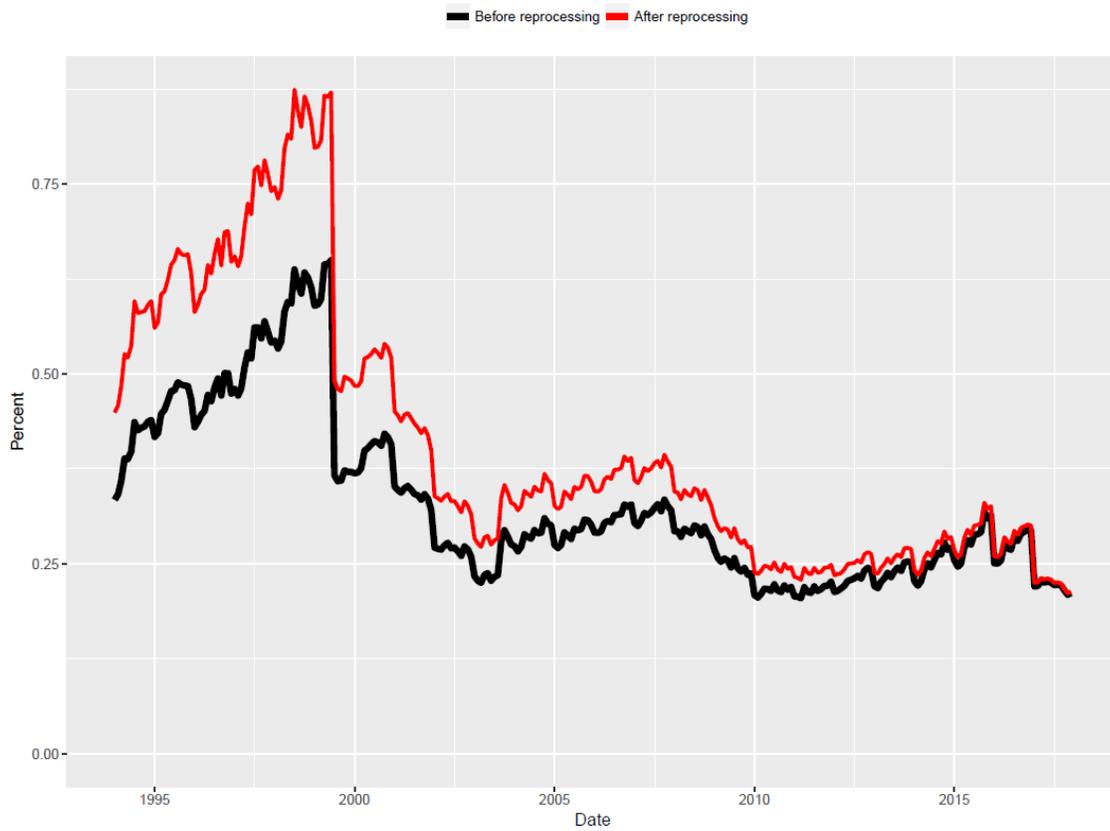


Figure A.22. PROByymm: Number of beneficiaries with value = 1

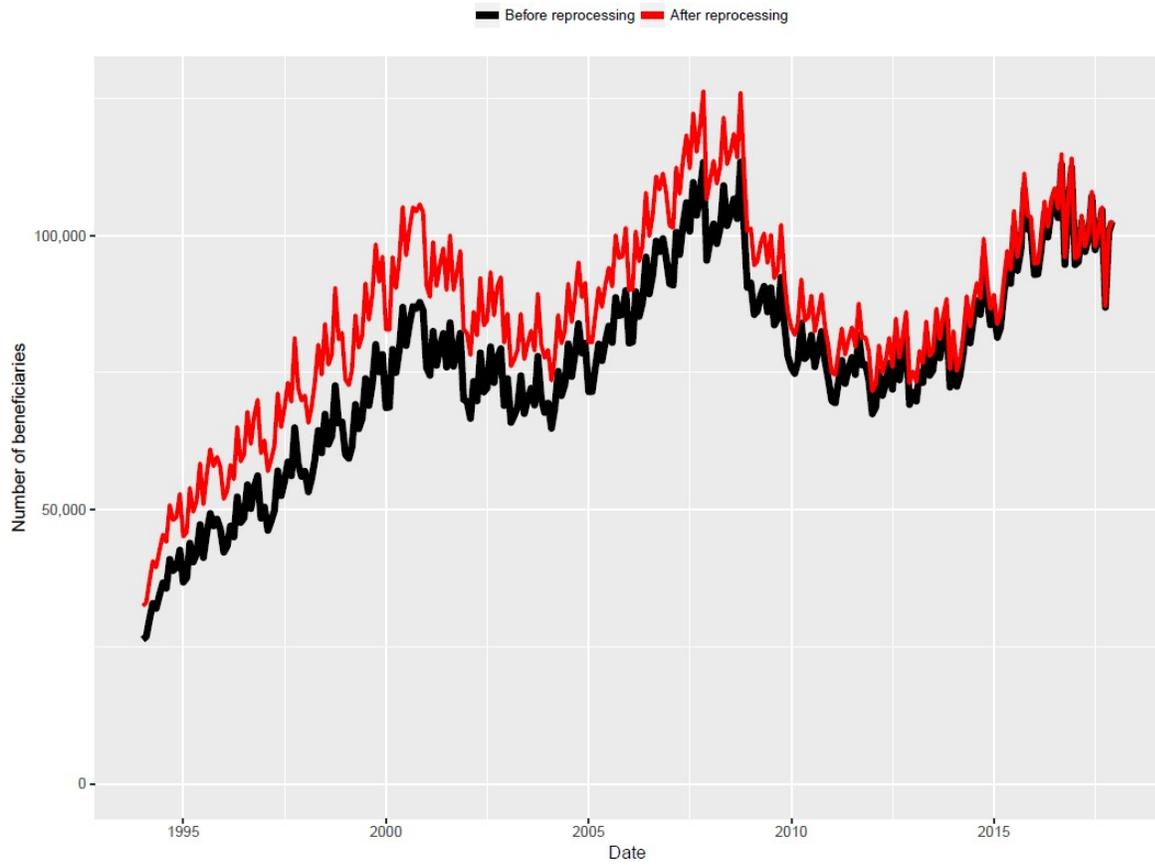


Figure A.23. PROByymm: Percent of SSI beneficiaries with value = 1

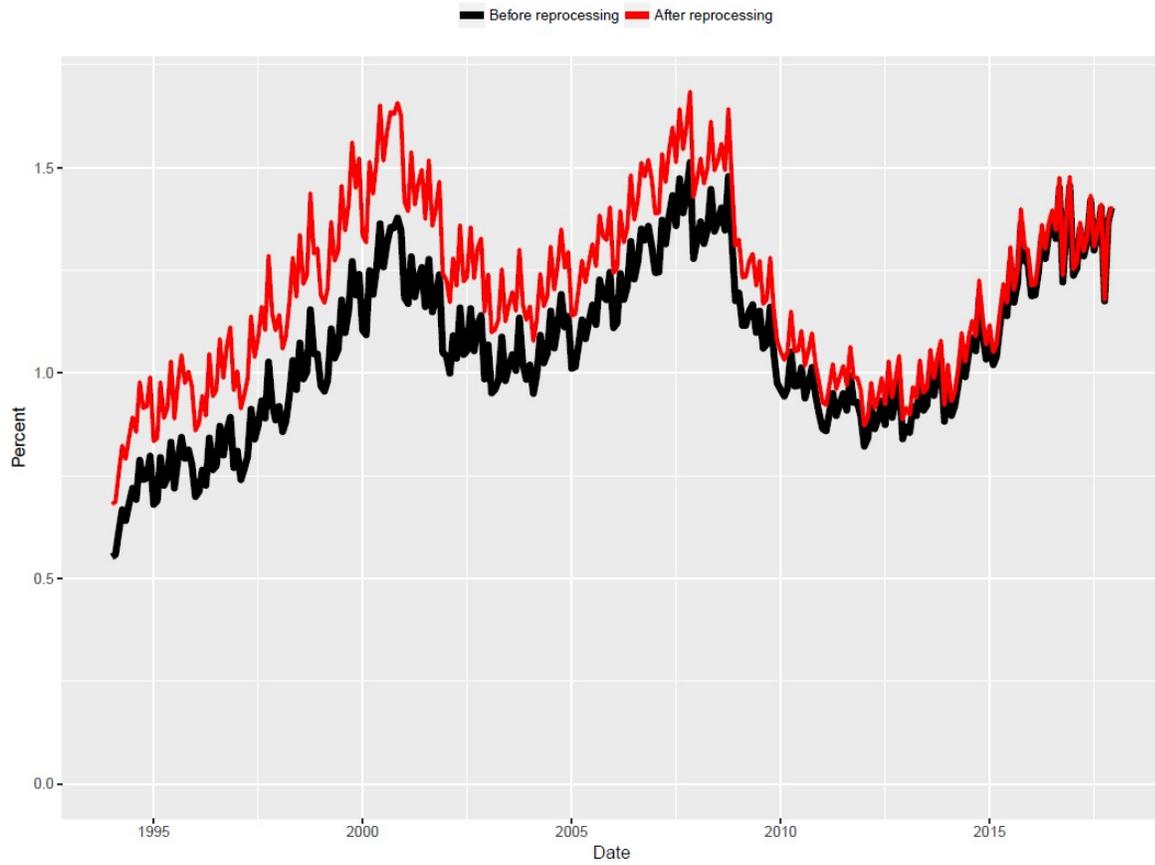


Figure A.24. PROByymm: Among beneficiaries with populated values, percent with value= 1

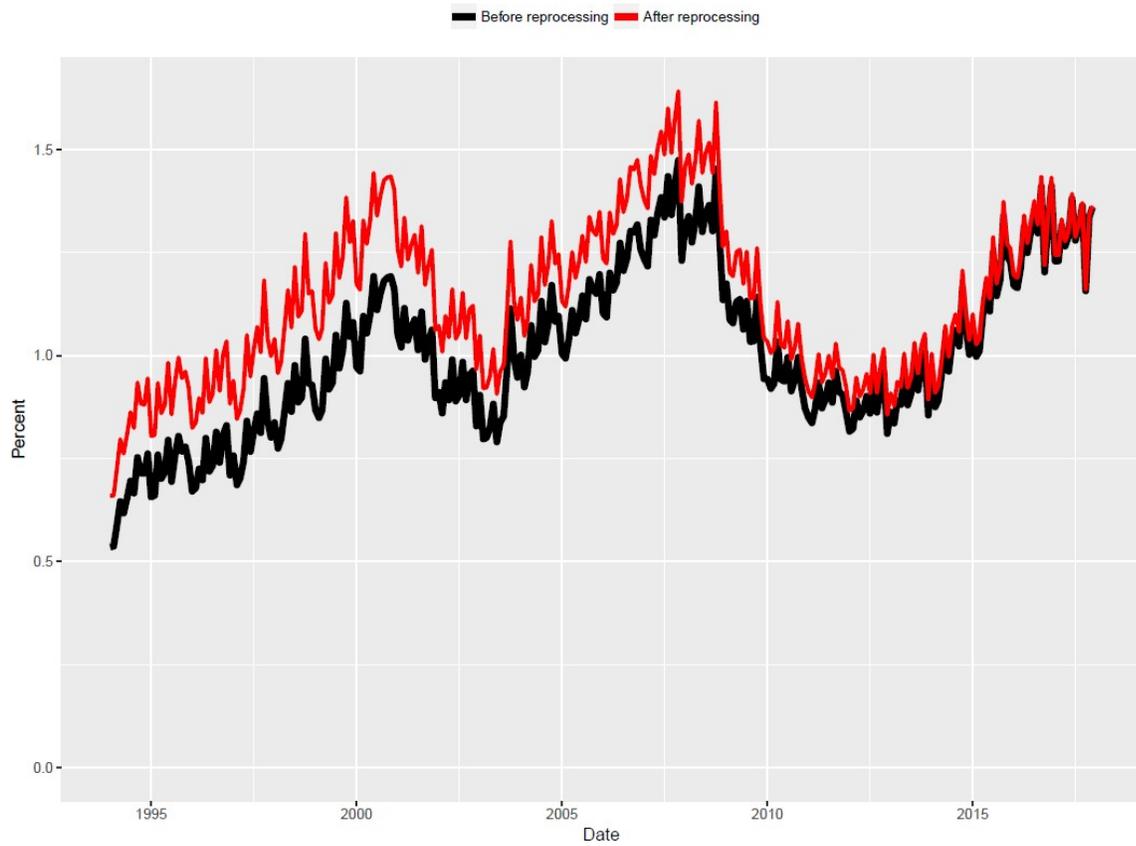


Figure A.25. STWSSlyymm: Number of beneficiaries where STWSSI = 0

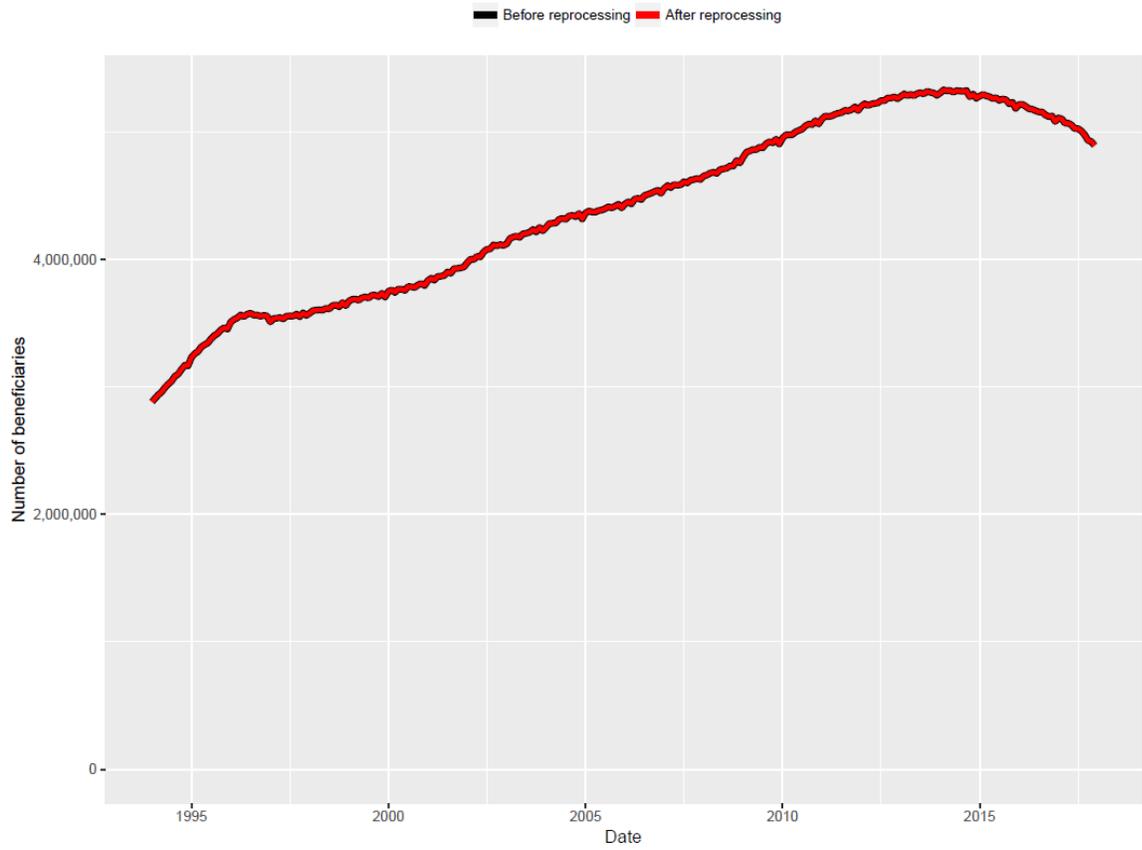


Figure A.26. STWSSlyymm: Number of beneficiaries where STWSSI = 1, 2, or 3

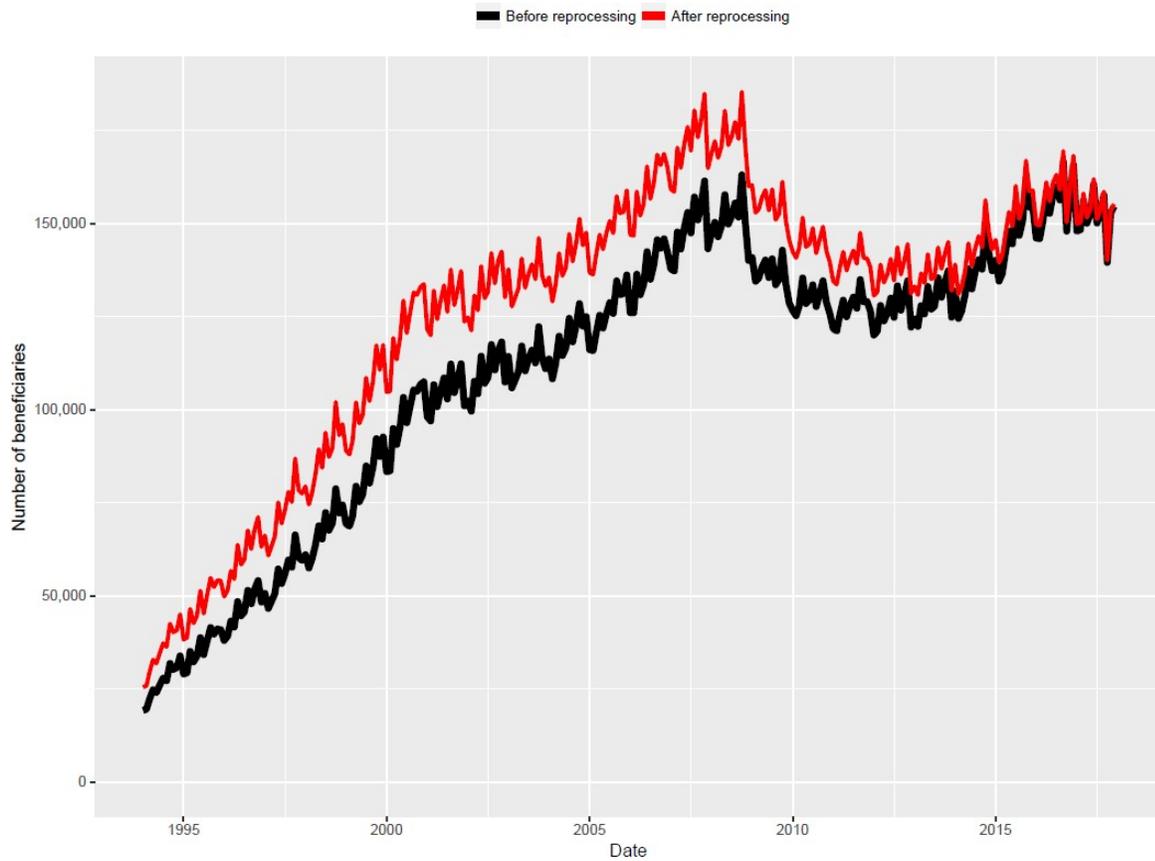


Figure A.27. STWSSlyymm: Number of beneficiaries where STWSSI = 4

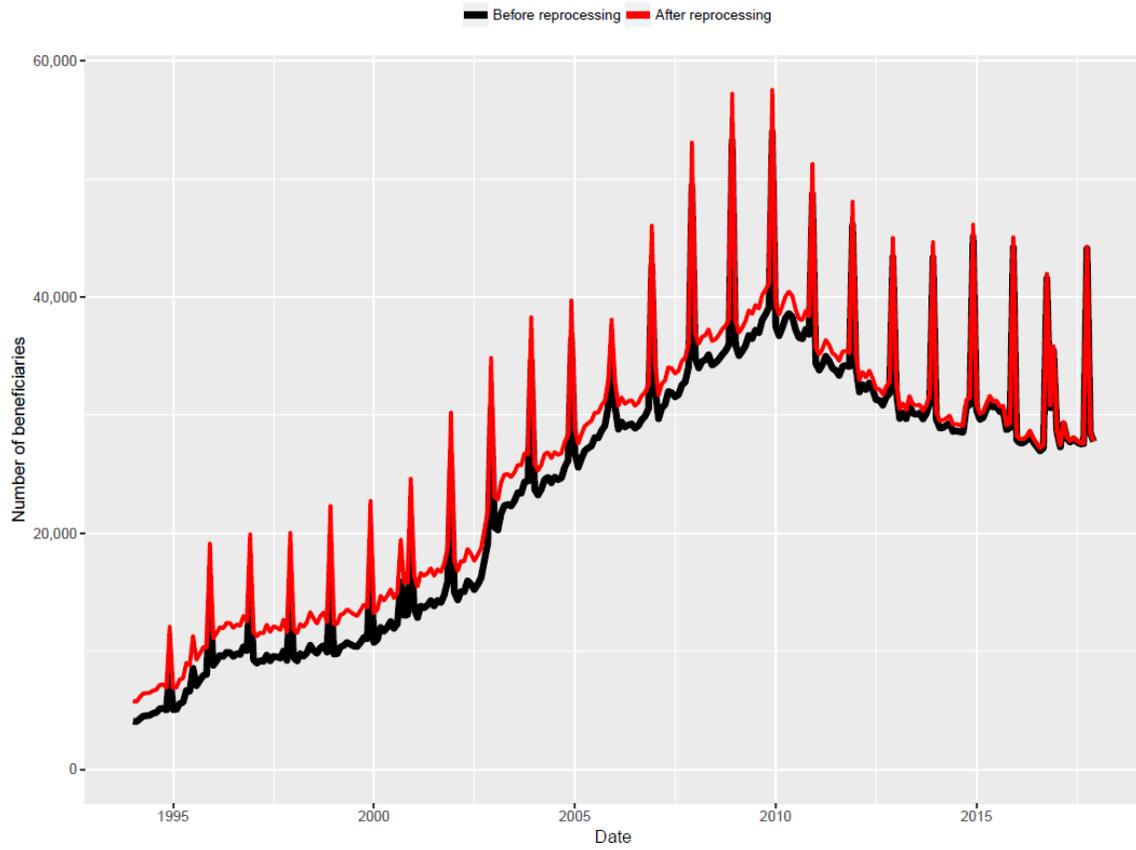


Figure A.28. STWSSlyymm: Number of beneficiaries where STWSSI = 8

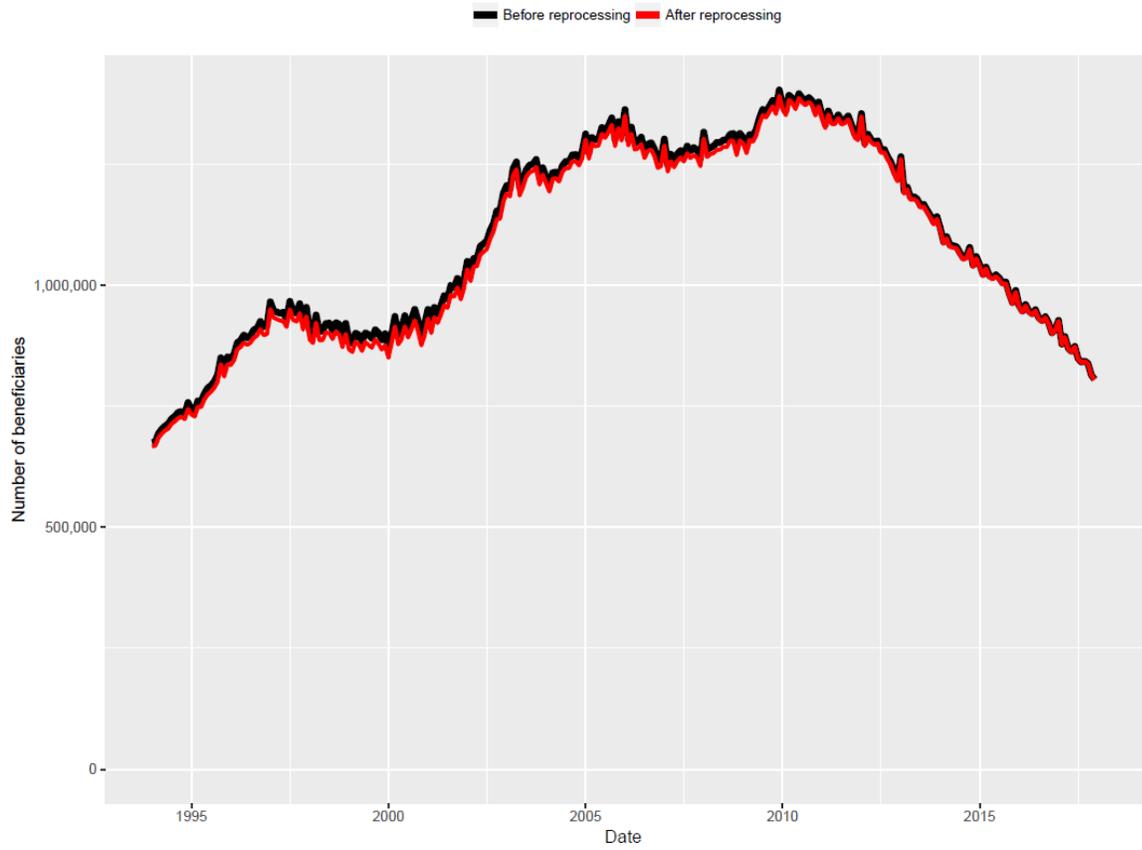


Figure A.29. STWSSlyymm: Number of beneficiaries where STWSSI = 9

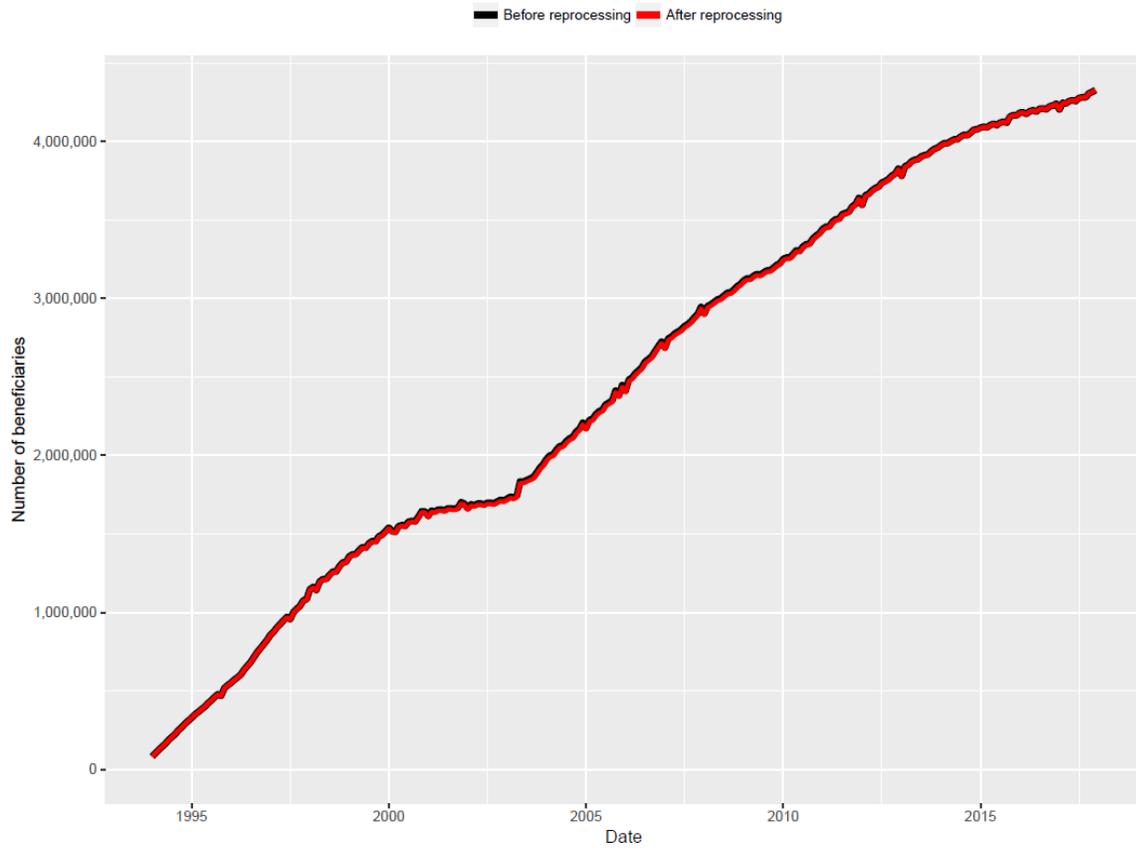


Figure A.30. STWCMymm: Number of beneficiaries where STWCM = 0

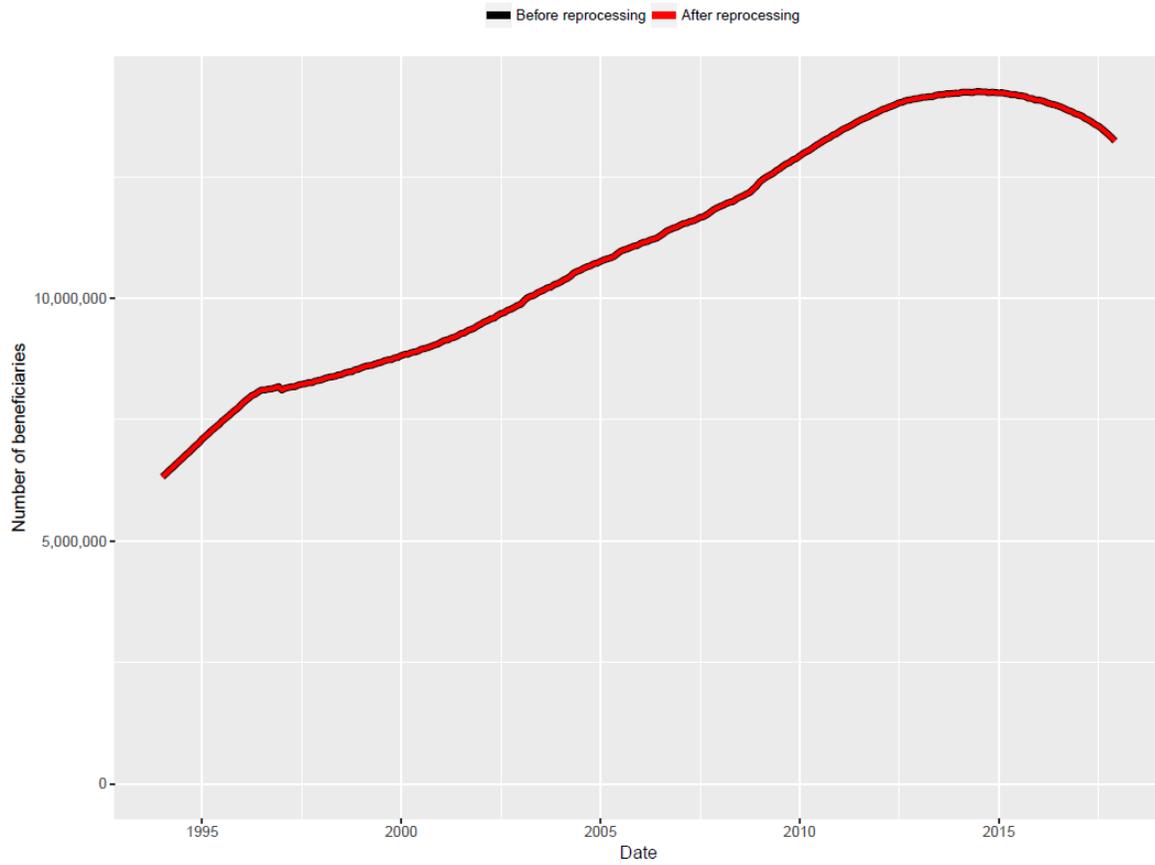


Figure A.31. STWCMymm: Number of beneficiaries where STWCM = 1, 2, or 3

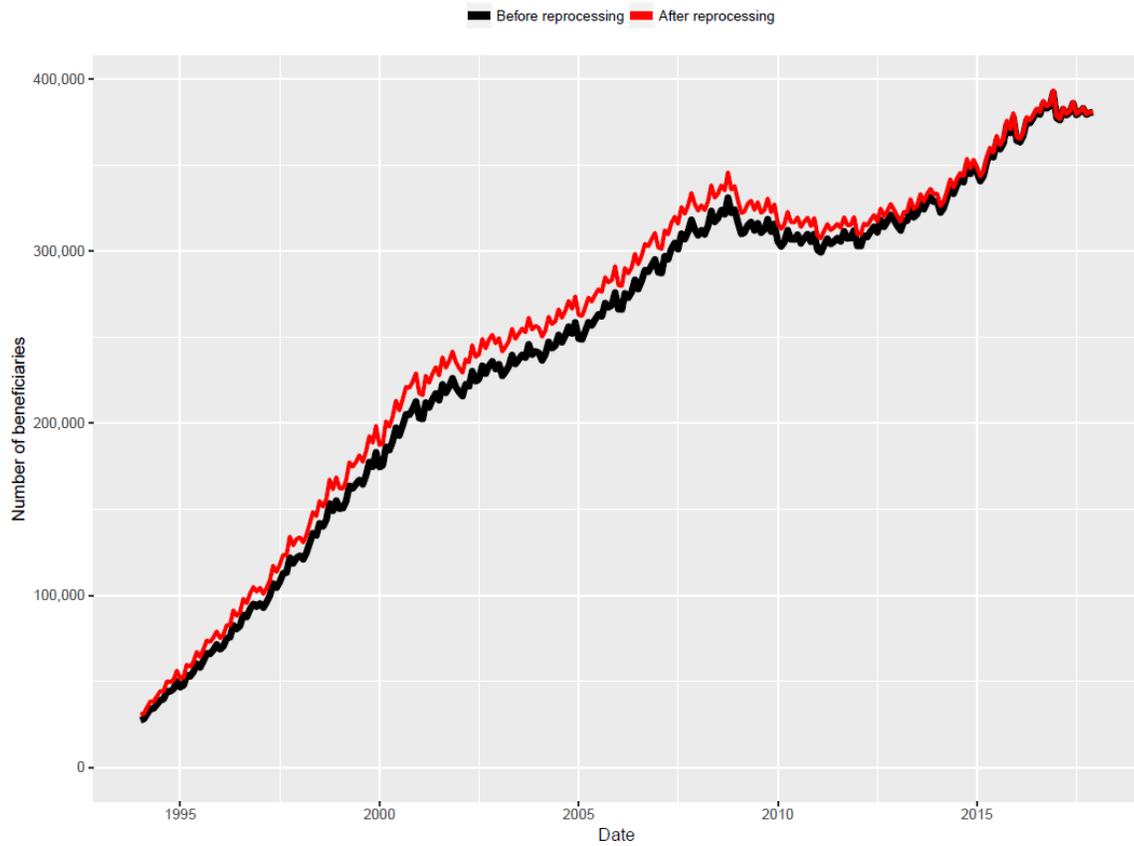


Figure A.32. STWCMymm: Number of beneficiaries where STWCM = 8

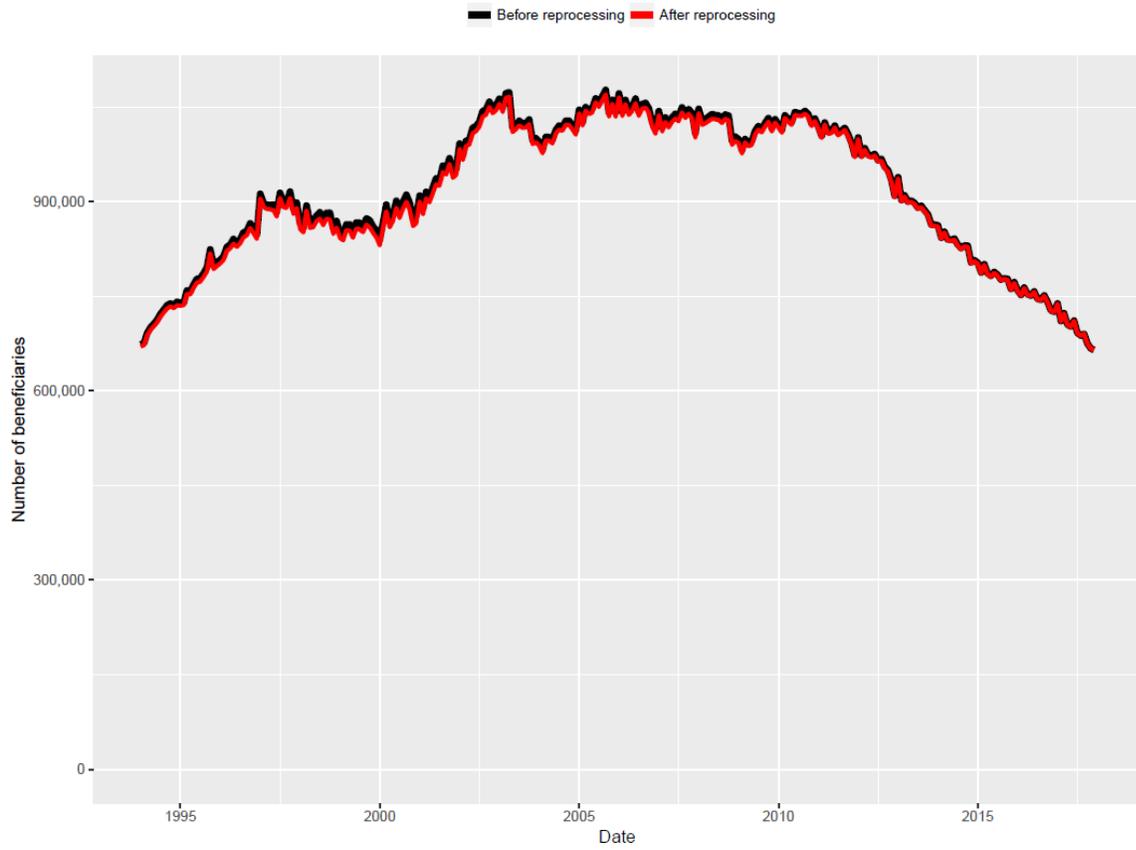


Figure A.33. STWCMymm: Number of beneficiaries where STWCM = 9

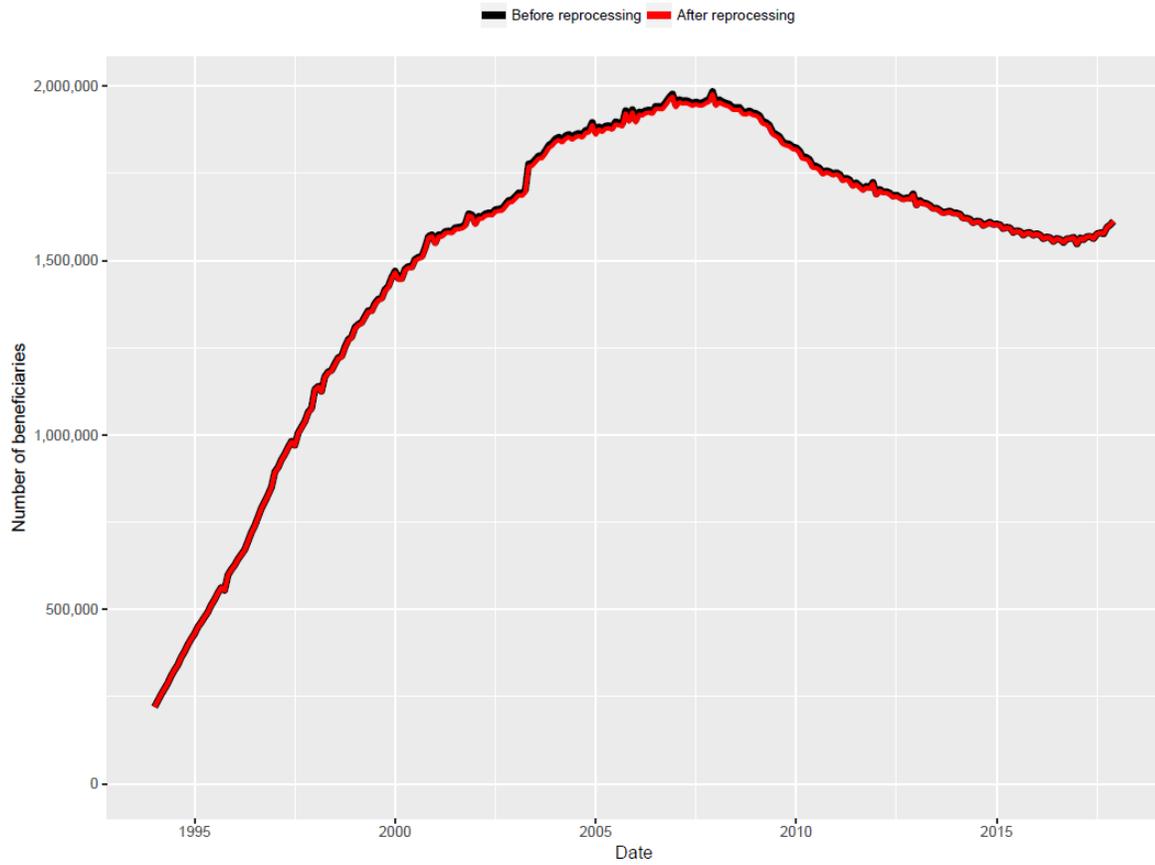


Figure A.34. BFWSSI_DRAFTyymm: Number of beneficiaries with values > 0

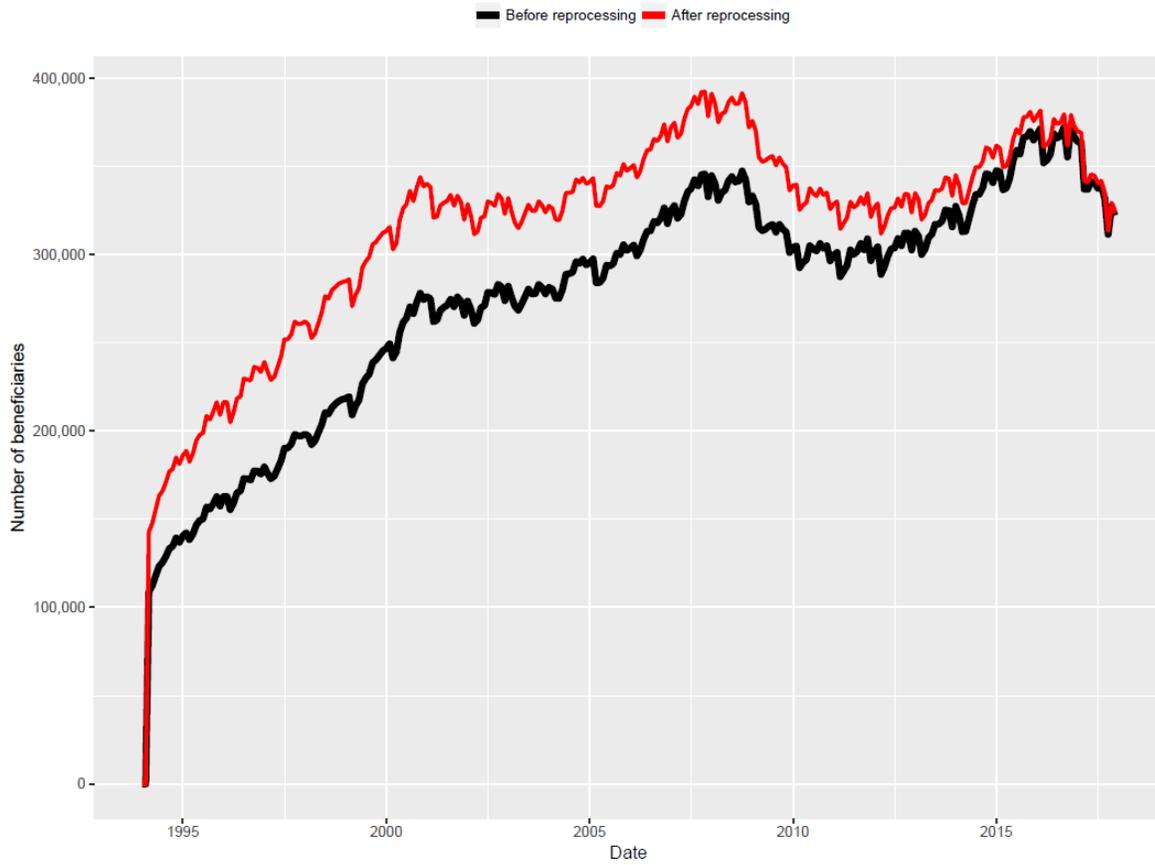


Figure A.35. BFWSSI_DRAFTyymm: Among beneficiaries with STW = 0, 1, 2, or 3, share (%) of beneficiaries with values > 0

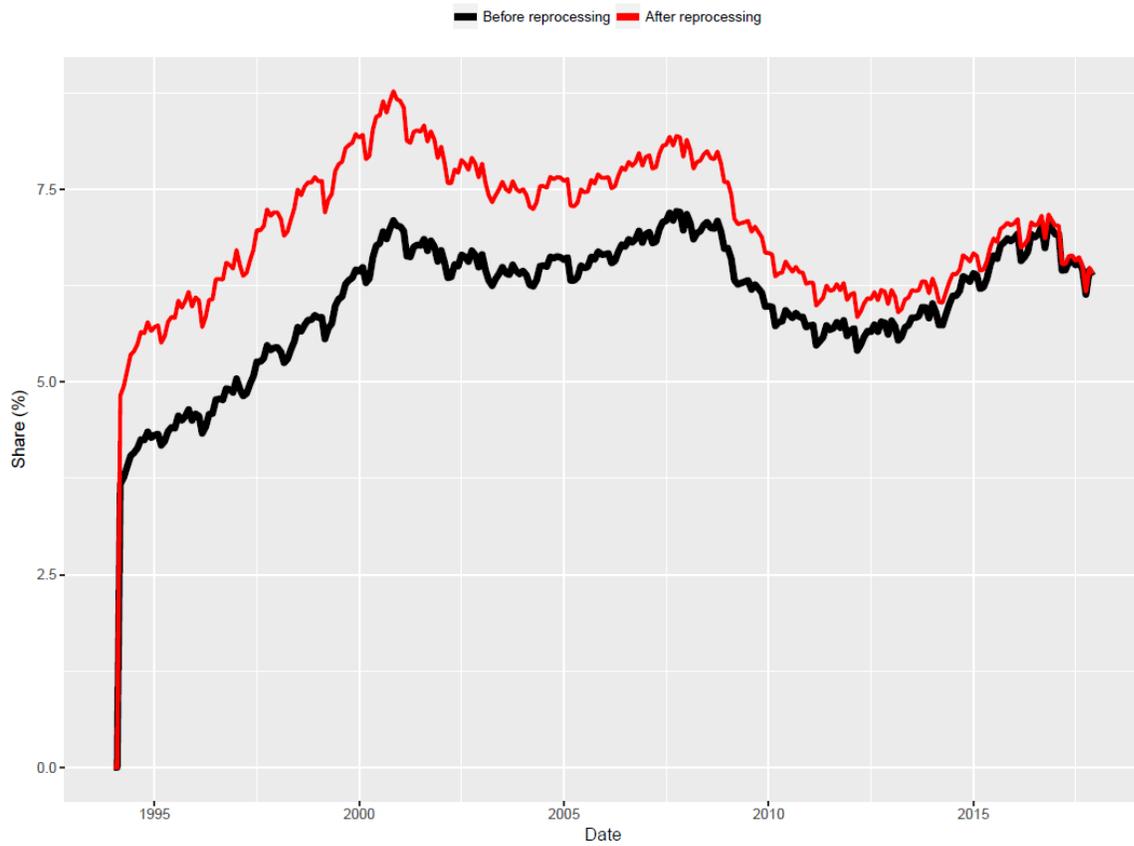


Figure A.36. BFWSSI_DRAFTyymm: Average of values > 0

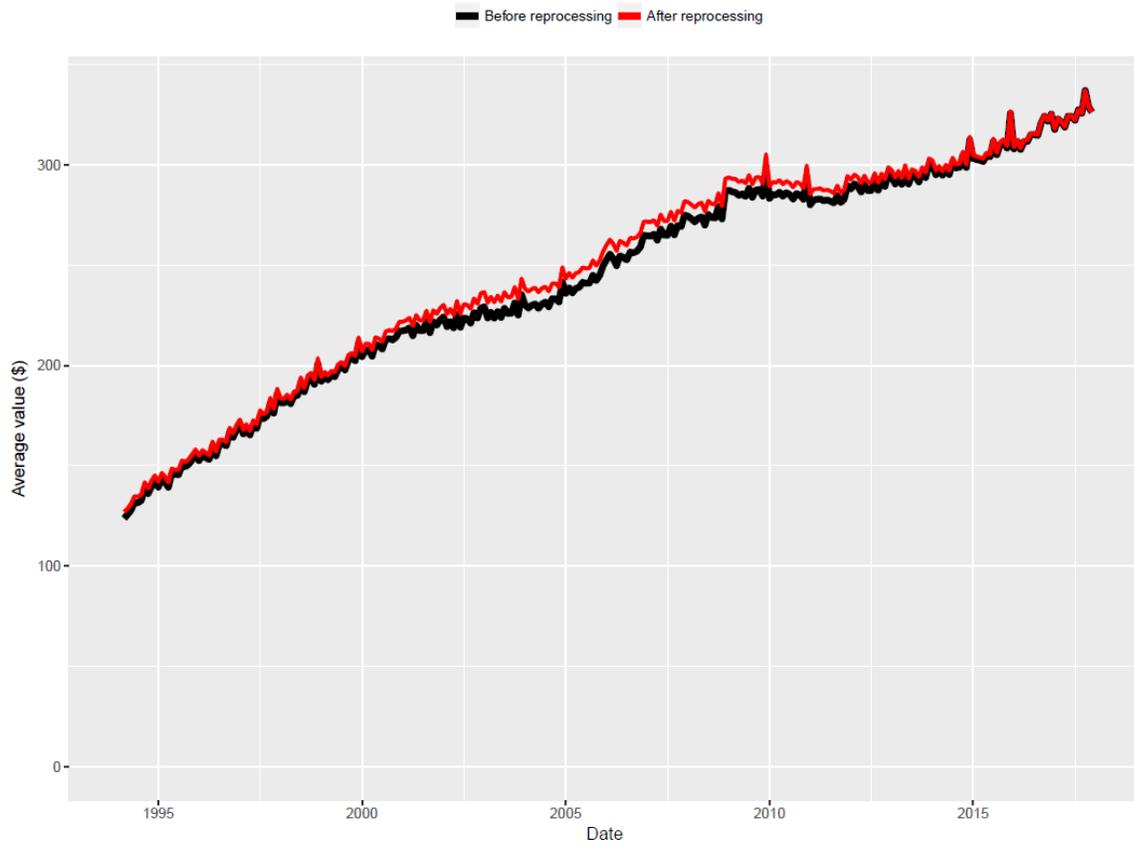


Figure A.37. BFWCM_DRAFTyymm: Number of beneficiaries with values > 0

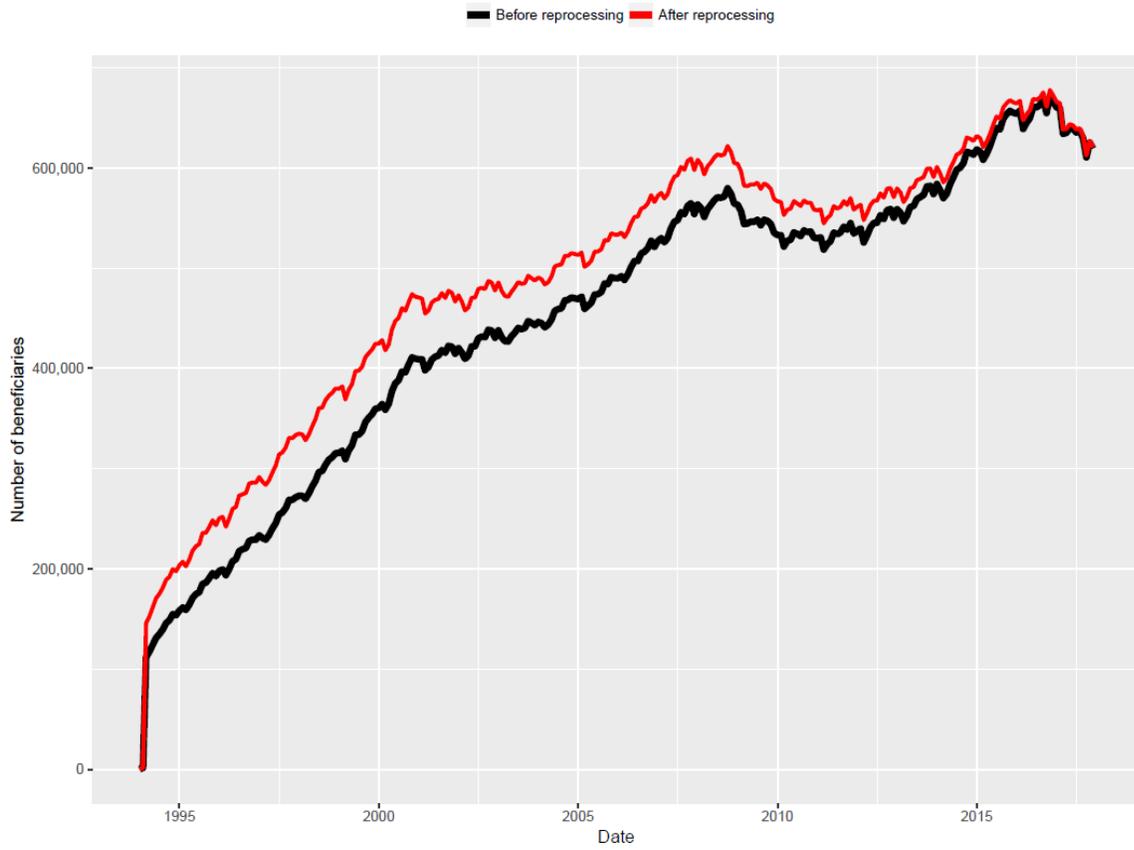


Figure A.38. BFWCM_DRAFTyymm: Among beneficiaries with STW = 0, 1, 2, or 3, share (%) of beneficiaries with values > 0

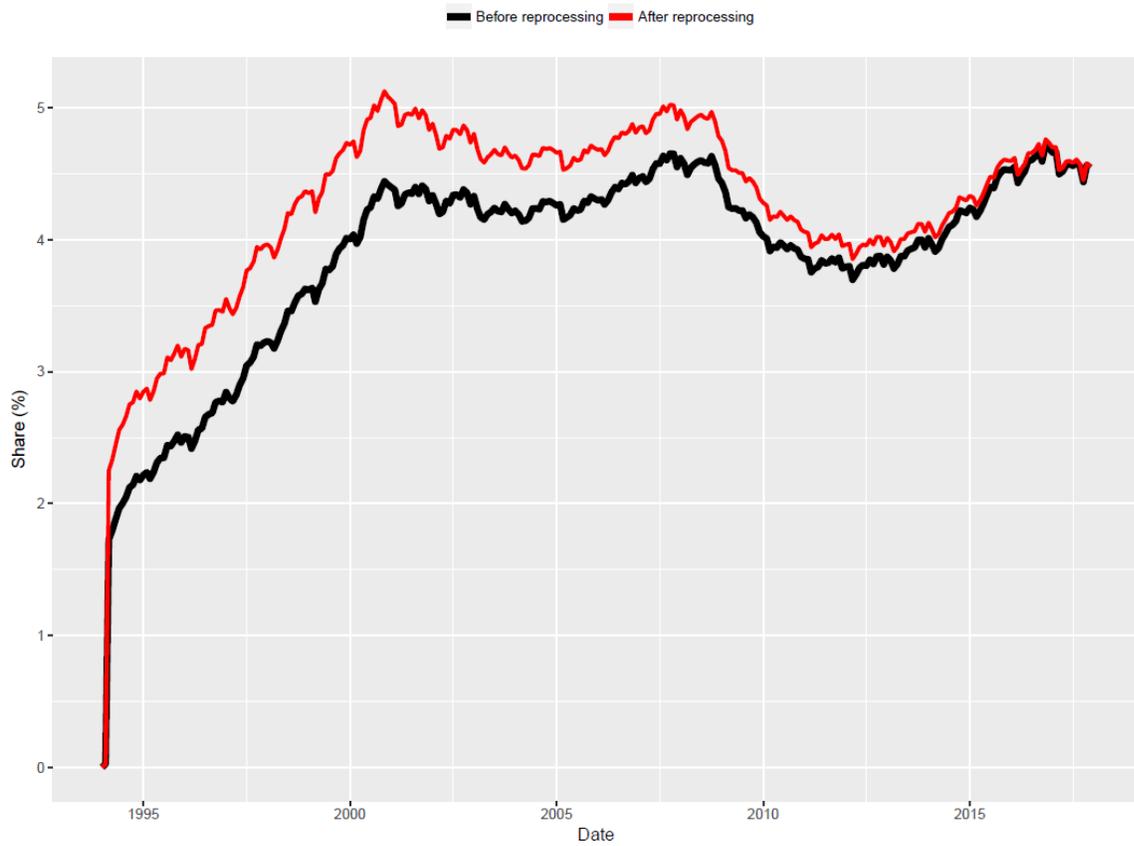
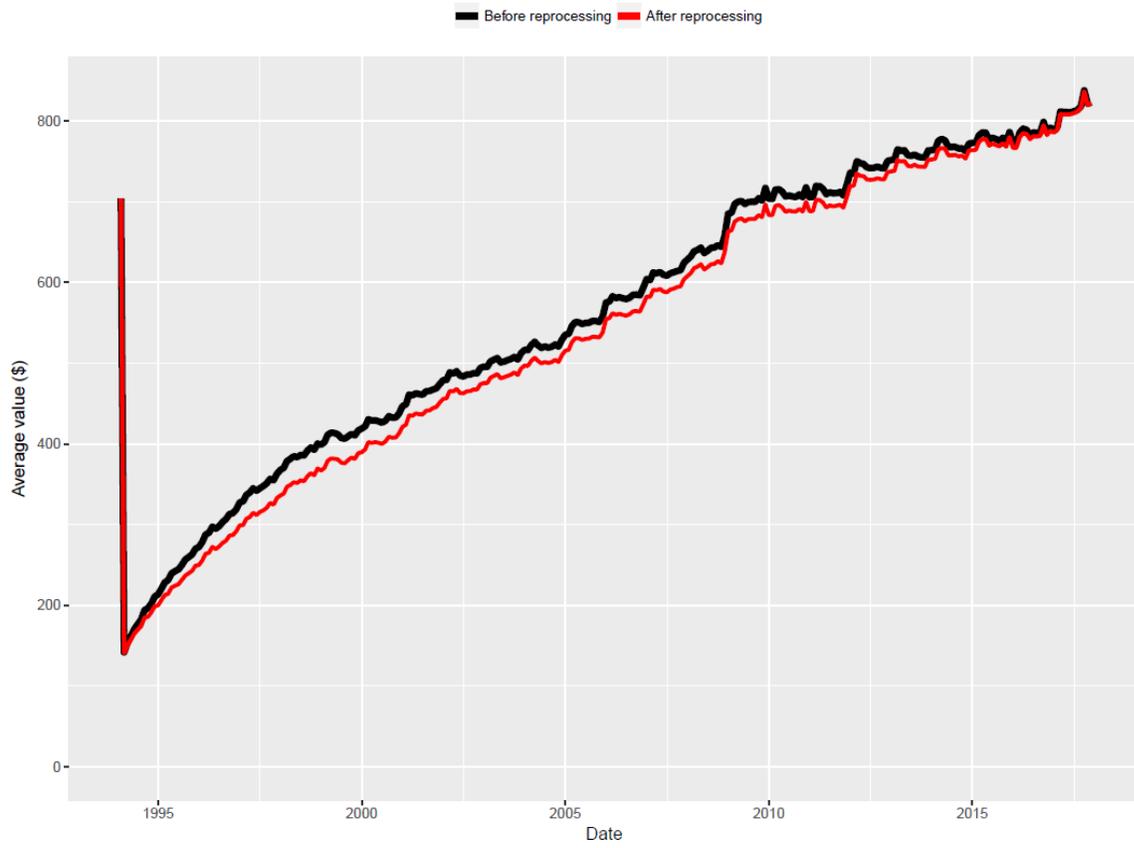


Figure A.39. BFWCM_DRAFTyymm: Average of values > 0



Mathematica

Princeton, NJ • Ann Arbor, MI • Cambridge, MA
Chicago, IL • Oakland, CA • Seattle, WA
Tucson, AZ • Woodlawn, MD • Washington, DC

EDI Global, a Mathematica Company

Bukoba, Tanzania • High Wycombe, United Kingdom



Mathematica
Progress Together

[mathematica.org](https://www.mathematica.org)