What would happen if the deadline for the 2020 Census data collection operation changed? Estimates of apportionment of the House of Representatives and distribution of federal Medicaid funding under different deadlines.

Jonathan Auerbach and Steve Pierson
Technical Report*
Office of Science Policy
American Statistical Association

September 17, 2020

Abstract

We consider whether apportionment of the House of Representatives and the distribution of federal Medicaid funding would change were the 2020 Census data collection operation to end on September 30th, October 31st, or other deadlines. We extrapolate the percent of households enumerated by the Census Bureau from reports dated between August 19th, 2020 and September 12th, 2020 to determine the percent that will be enumerated each day if field operations continue to count households at their currently decreasing rate. We then use omission, erroneous enumeration, and imputation rates from 2010 Census operations to construct three scenarios of 2020 Census quality. For each scenario, we estimate the changes in apportionment and Medicaid funding that would occur under each deadline. Our estimates suggest that under a September 30th deadline, California, Ohio, or Idaho could gain seats in the House of Representatives, while Florida or Montana could lose seats (that they might not have under an October 31st deadline). They also suggest Texas, Florida, Arizona, Georgia, and North Carolina could collectively forfeit as much as five hundred million dollars in federal Medicaid funding each year under a September 30th deadline (that they would receive under an October 31st deadline). (Medicaid funding represents only a small portion of the estimated $1.5 trillion distributed using census data.) Though the exact winners and losers change depending on the assumptions underlying each scenario, our estimates suggest significant consequences to an early cessation of data collection operations.

Introduction

The U.S. Census Bureau is currently in the process of counting the United States population. This count, referred to as the enumeration, occurs once every ten years and will determine how the U.S. population is represented in the House of Representatives (as well as the Electoral College, state, and local governments), an estimated $1.5 trillion in government funds are distributed (Reamer (2018)), and countless personal and business decisions are made.

The Census Bureau plans to cease its data collection operations on September 30th according to its website (at the time of this report). The deadline leaves half a month (at the time of this report) for the Bureau to contact households that have not yet responded; any remaining households are then filled-in by the Census Bureau, using a statistical process called imputation. Although imputation can be remarkably accurate,

*This report was written by researchers at the American Statistical Association to evaluate policy issues of statistical importance. This report is intended for research purposes only and is not an official position, statement, or policy of the American Statistical Association on these issues. The authors can be contacted by email at jonathan@amstat.org.

historically, the census count has been most reliable when households self-respond or the Bureau is able to follow-up with households directly.

This report investigates the consequences of extending the data-collection deadline to October 31st or later so that the Bureau has additional time to follow up with non-responding households and engage in other post-processing operations. We focus on two consequences: whether a later deadline would change the apportionment of the House of Representatives and whether it would change the distribution of federal Medicaid funding. Other analyses have examined the relationship between the 2020 Census count and apportionment or funding. For example, see Seeskin and Spencer (2018), Elliott et al. (2019), and Frey (2020), on which our analysis is based. However, to our knowledge, none have investigated the consequences of moving the Census deadline specifically.

The Census is a complicated process, operating under unprecedented conditions, positive and negative–for example, enumerators have access to new data sources and improved technology albeit during a global pandemic and recession. The accuracy of the Census may only become clear once the count is finalized and the results are verified. For the purpose of investigating the value of a deadline extension, however, we consider three scenarios of Census accuracy: a neutral scenario in which the aforementioned conditions lead to a 2020 Census of similar quality to the 2010 Census; an optimistic scenario, in which the 2020 Census is higher quality than the 2010 Census; and a pessimistic scenario, in which the 2020 Census is lower quality than the 2010 Census. Our approach is similar to Elliott et al. (2019), which also uses 2010 Census operations to predict the outcome of the 2020 Census.

In all three scenarios, we find apportionment and funding would better reflect the U.S. population if the deadline were extended–in some cases significantly better. We stress that our analysis only considers three of many possible scenarios. While other scenarios could produce a different set of winners and losers, our findings suggest that apportionment and funding would still benefit from a deadline extension.

We present the details of our work in four sections: (1) we forecast the percent of households that will be enumerated by current Census Bureau operations for each state under different deadlines; (2) we present the three scenarios, each of which assume the count will be subject to a certain amount of omissions, erroneous enumerations, and imputations (depending on whether a respondent self-responds, is enumerated by current operations before the deadline, or enumerated after the deadline or by additional operations); (3) we calculate the apportionment of the House of Representatives and distribution of FY15 federal Medicaid funding under each scenario; and (4) we discuss some limitations of our approach. An Appendix contains all figures, tables, and code.

Section 1. Forecasting the percent enumerated with a logistic curve

We begin by forecasting the percent of households that will be enumerated on any given date were Census operations to continue at their current pace. We obtain the total response rates by state from 2020census.gov, which provides daily reports of the percent enumerated to date, starting with August 19th. The data are displayed in Figure 1, ordered by the percent of households enumerated before September 12th. As of this date, Alabama, Louisiana, and Georgia have the lowest number of households enumerated. Idaho, West Virginia, and Hawaii have the highest.

Enumeration is clearly increasing as field operations continue. One might be tempted to assume enumeration will continue to grow at a constant rate—that is, the number of households enumerated will grow linearly, by the same amount every day. Figure 2 shows the percent that would be enumerated under the constant growth assumption, adding a best-fit line to the data displayed in Figure 1. (The growth rate is computed using the least squares algorithm.) A consequence of this assumption is that nearly all states will have 100 percent of their population enumerated by the September 30th deadline.

However, linear growth is implausible; it is unlikely that 100 percent of residents can be counted by September 30th. In fact, many states, such as Idaho, Alaska, and Maine, are clearly increasing at a decreasing rate. A popular alternative to linear growth is logistic growth, which allows for enumeration to grow at a decreasing rate. (The growth rate is assumed proportional to the percent of uncounted households, see Weisstein (2003)

for details.) Figure 3 adds a best-fit logistic curve to the data displayed in Figure 1. (The growth rate is computed using the nonlinear least squares algorithm, and the max value parameter—also called the capacity or asymptote—is set to 100 percent.) A consequence of this assumption is that, only half of all states will have at least 96 percent of households enumerated by September 30th. If the deadline is extended to October 31st, more than 99 percent of U.S. households will be counted, and all states will have more than 95 percent counted. However, a deadline of December 31st is necessary for at least 99 percent to be counted in each state. The projected number enumerated for each state by select deadlines is displayed in both Figure 4 and Table 1 at the end of this analysis.

For the remainder of the analysis, we assume the percent of persons enumerated by the Census is equal to the percent of households enumerated. This assumption is made by Elliott et al. (2019) and partially supported by the 2010 Census (as reported in the 2010 Census Coverage Measurement persons and household reports). For example, Figure 5 displays the $\log_{10}$-number of correctly enumerated persons for each state against the $\log_{10}$-number of correctly enumerated households. The best-fit line has a slope of 1, suggesting the number of persons and households enumerated are proportional across states. Consequently, were this best-fit line used to estimate the number of persons enumerated from the number of households enumerated—the percent enumerated would remain the same.

Section 2. Three scenarios of omissions, erroneous enumerations, and imputations

The number counted by a census never equals the true population (in the case of the 2020 Census, where every resident was living on April 1st, 2020); errors are bound to occur. For example, the number counted in the 2010 Census was equal to the true population minus omissions plus erroneous enumerations (for example, duplicates) plus imputations (the Census Bureau's attempt at filling in incorrect or missing persons). See the Census Bureau website for a description of the components of coverage of the 2010 Census and Anderson and Fienberg (1999) and Freedman and Wachter (2007) for a general discussion of U.S. census errors.

The quality of the 2020 Census count—as determined by the number of omissions, erroneous enumerations, and imputations—will be reported to the public following the completion of the Census and post-Census studies (or a similar decomposition). For this analysis, we assume self-responding individuals will be counted with the lowest number of omissions, erroneous enumerations, and imputations; individuals enumerated during current Census operations (see Section 1) will have the second lowest; and individuals enumerated after (or in addition to current operations) will have the highest. Specifically, we base our omission, erroneous enumerations, and imputation rates on the 2010 Census (as reported by the 2010 Census Coverage Measurement reports). For self-responding individuals, we assume 2.5 percent of the count will be erroneously enumerated and 0.3 percent will be imputed (Table 18). For non-self-responding individuals counted by current operations, we assume 4.3 percent erroneously enumerated and 2.6 percent imputed (Table 19). For remaining individuals, we assume 8 percent erroneously enumerated and 17.3 percent imputed (Table 19). Thus, the value of an extended deadline is that more individuals are enumerated through reliable operations and fewer are subject to the 8 percent erroneous enumeration and 17.3 percent imputation rates.

The omission of persons from the 2010 Census was not reported in the 2010 Census Coverage Measurement report by Census operation. However, the number of omissions is reported at the state level, and at this level it is proportional to the number of erroneous enumerations and imputations across states (see Figure 6. Note that erroneous enumeration and imputation rates in the Measurement report are per Census count and omissions are per population). We assume the same relationship holds for the self-responding and non-self-responding persons described in the previous paragraph, and we estimate the number of omissions from the number of erroneous enumerations and imputations via multivariate regression (using the least squares algorithm, all variables are on the log scale).

The omission, erroneous enumeration, and imputation rates described in this paragraph form the basis of our first scenario. We also consider two other scenarios, which are identical to scenario 1, except the proportion omitted, erroneously enumerated, and imputed are halved and doubled in scenarios 2 and 3 respectively.

Section 3. Determining apportionment of the House of Representatives and distribution of federal Medicaid funds.

We compute the number of seats apportioned to each state and the distribution of federal Medicaid funds were the Census Bureau's enumeration operations to end by select deadlines. The computations use an estimate of the 2020 Census count, which we derive by combining the 2020 population projections of Elliott et al. (2019) with the logistic-growth forecast from Section 1 and the omission, error, and imputation rates outlined in Section 2—for each scenario and deadline.

We calculate apportionment using the *method of equal proportions*, detailed on the Census Bureau website. (See Wright and Cobb (2005) for an account of the apportionment process.) Figures 7.1-7.3 display the states we estimate will gain (green) or lose (red) seats if the Census Bureau's enumeration operations cease by select dates and fail to count the entire U.S. population. Together, these scenarios suggest California, Ohio, or Idaho could gain seats with a September 30th deadline (relative to an October 31st deadline), while Florida or Montana could lose seats. In addition, we find an October 31st deadline may not be far enough to guarantee the most representative apportionment. In fact, in the pessimistic scenario, even if enumeration were extended until 2021, California would still have a seat belonging to Montana. It is possible that similar misrepresentations will occur at the state and local level.

We calculate federal funding for state Medicaid expenditures using the *Federal Medical Assistance Percentage formula*, detailed on the Department of Health and Human Services website. We start with the total amount of federal Medicaid funding for each state in Fiscal Year 2015 (see Reamer (2018)). We then calculate the Federal Medical Assistance Percentage (FMAP) formula using personal income data from the U.S. Bureau of Economic Analysis and our estimated 2020 Census counts to determine the change in state funding under the scenarios in Section 3. The estimated gains and losses for each state are displayed in Figures 8.1-8.3. We find 18 states could lose a total of eight hundred million dollars a year under a September 30th deadline (relative to an October 31st deadline), with Texas, Florida, Arizona, Georgia, and North Carolina losing a total of five hundred million dollars. Note that Medicaid funding represents only a small portion of the estimated $1.5 trillion distributed using census data.

Section 4. Discussion

It is often quipped that predictions are difficult—especially those about the future. However, we believe our analysis provides an objective, however approximate, basis for quantifying the consequences of a Census miscount that could arise from ending the Census Bureau's enumeration operations by the September 30th deadline. We hope our findings help inform the discussion over whether the deadline for these operations should be extended.

Our analysis rests on three assumptions: (1) that the logistic-growth forecast of households reflects the percent of residents that will be enumerated by the Census if operations continue at their current pace; (2) that the 2020 population projections closely describe the 2020 population; and (3) that the Census Bureau will not adjust the Census counts beyond what is described in the Census Coverage Measurement reports. We conclude by briefly discussing some of these assumptions.

Proponents of a September 30th deadline may argue that the logistic growth assumption is pessimistic and that enumeration will continue at a faster pace. Perhaps, for example, the Census Bureau may employ additional operations to count those households. The curves in Figure 3 then represent what the enumeration process would have been without these additional operations; additional operations are assumed to enumerate the remaining population with reduced accuracy.

Our forecast assumes the max value parameter (also called the capacity or asymptote) is 100 percent. We then estimate the net undercount under the different scenarios using the 2010 Census Coverage Measurement reports. Alternatively, we could estimate the max value as an additional parameter, either estimating it from the data or from the data of previous censuses.

Apportionment (but not distribution of funding) is somewhat sensitive to the 2020 population. We use the 2020 population projections from the Urban Institute (Elliott et al. (2019)). However, using projections from The Brooking Institution (for example, Frey (2020)) the Census Bureau' 2019 resident population projections, or the U.S. Bureau of Economic Analysis gives different results. In all cases, extending the deadline yields more representative apportionment and funding. Our analysis may be sensitive to rounding and other approximations—for example, the Census Coverage Measurement reports round counts to the nearest thousand and rates to the nearest tenth—but we are unable to determine the magnitude of this sensitivity from the data.

Our analysis only considers the count quality of the 2020 Census: whether the number enumerated by the Census Bureau will be close to the actual number of residents. Extending the deadline could also improve the characteristic quality, whether the sociodemographic information collected by the Census represents the sociodemographic characteristics of U.S. residents. It is possible that extending the 2020 Census deadline would produce a large improvement in characteristic quality. However, the Census Bureau does not currently report the percent of 2020 households enumerated by sociodemographics, and therefore the characteristic quality of the 2020 Census could not be investigated using the methods of this analysis.

References

Anderson, Margo, and Stephen E Fienberg. 1999. *Who Counts?: The Politics of Census-Taking in Contemporary America*. Russell Sage Foundation.

Elliott, Diana, Rob Santos, Steven Martin, and Charmaine Runes. 2019. "Assessing Miscounts in the 2020 Census." *Washington, DC: Urban Institute.* www.urban.org/research/publication/assessing-miscounts-2020-census.

Freedman, David A, and Kenneth W Wachter. 2007. "Methods for Census 2000 and Statistical Adjustments." *Handbook of Social Science Methodology.*

Frey, William H. 2020. *Census Day Is Here. How Is Our Nation Changing? The Brookings Institution.* www.brookings.edu/research/the-2020-census-is-here-what-will-it-tell-us/.

Reamer, Andrew. 2018. "Counting for Dollars 2020: The Role of the Decennial Census in the Geographic Distribution of Federal Funds." *Institute of Public Policy.* https://gwipp.gwu.edu/counting-dollars-2020-role-decennial-census-geographic-distribution-federal-funds#Briefs.

Seeskin, Zach, and Bruce Spencer. 2018. "Balancing 2020 Census Cost and Accuracy: Consequences for Congressional Apportionment and Fund Allocations." Working Paper. Evanston, IL: Northwestern University, Institute for Policy. https://www.ipr.northwestern.edu/documents/working-papers/2018/wp-18-10.pdf.

Weisstein, Eric W. 2003. "Logistic Equation." https://mathworld.wolfram.com/LogisticEquation.html.

Wright, T, and G Cobb. 2005. "Counting and Apportionment: Foundations of America's Democracy." *Statistics: A Guide to the Unknown*, 35–67.

Appendix

Figure 1: Percent of households counted between August 19 and September 12

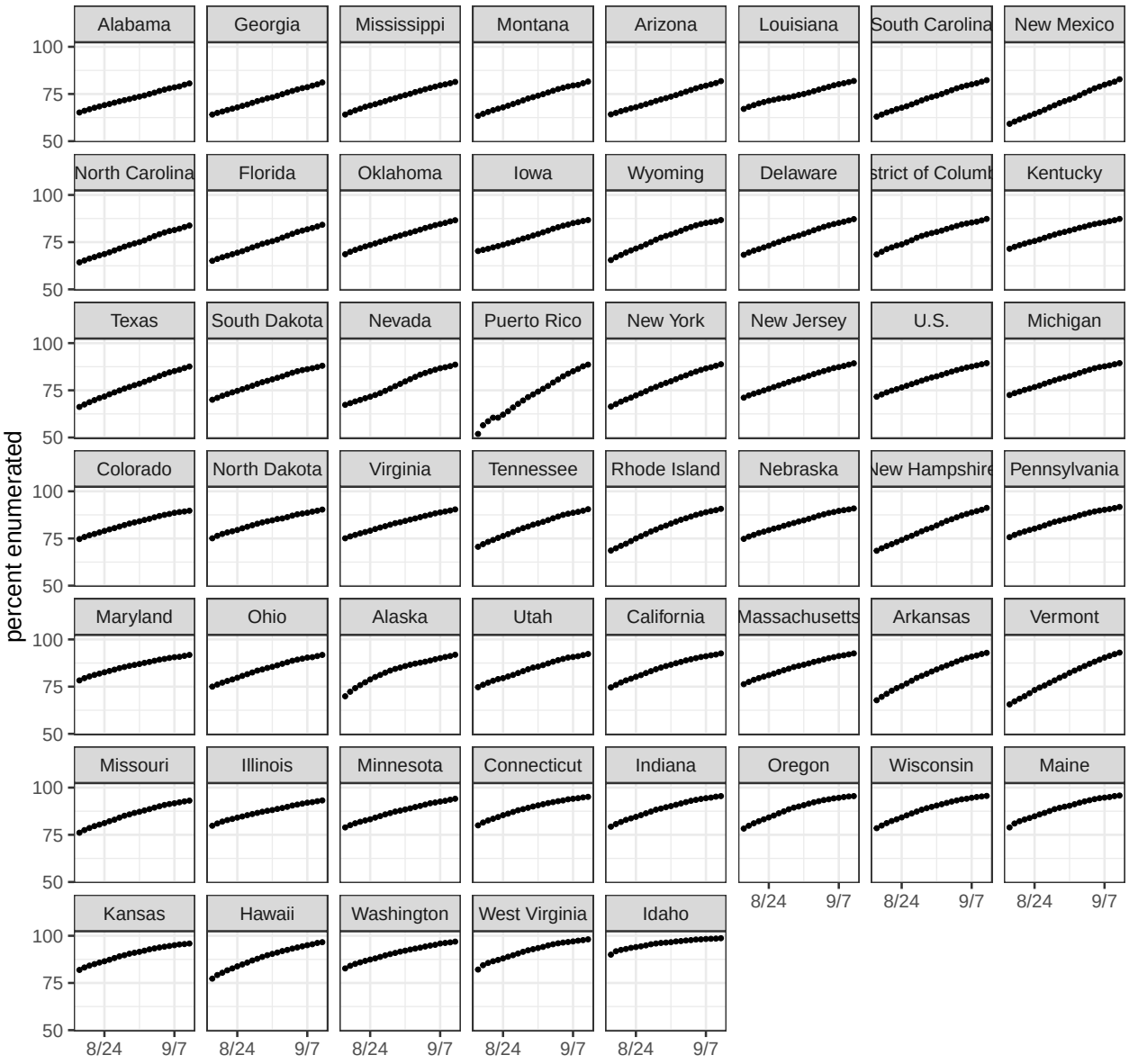

Figure 2: Percent counted if enumeration increases at constant rate

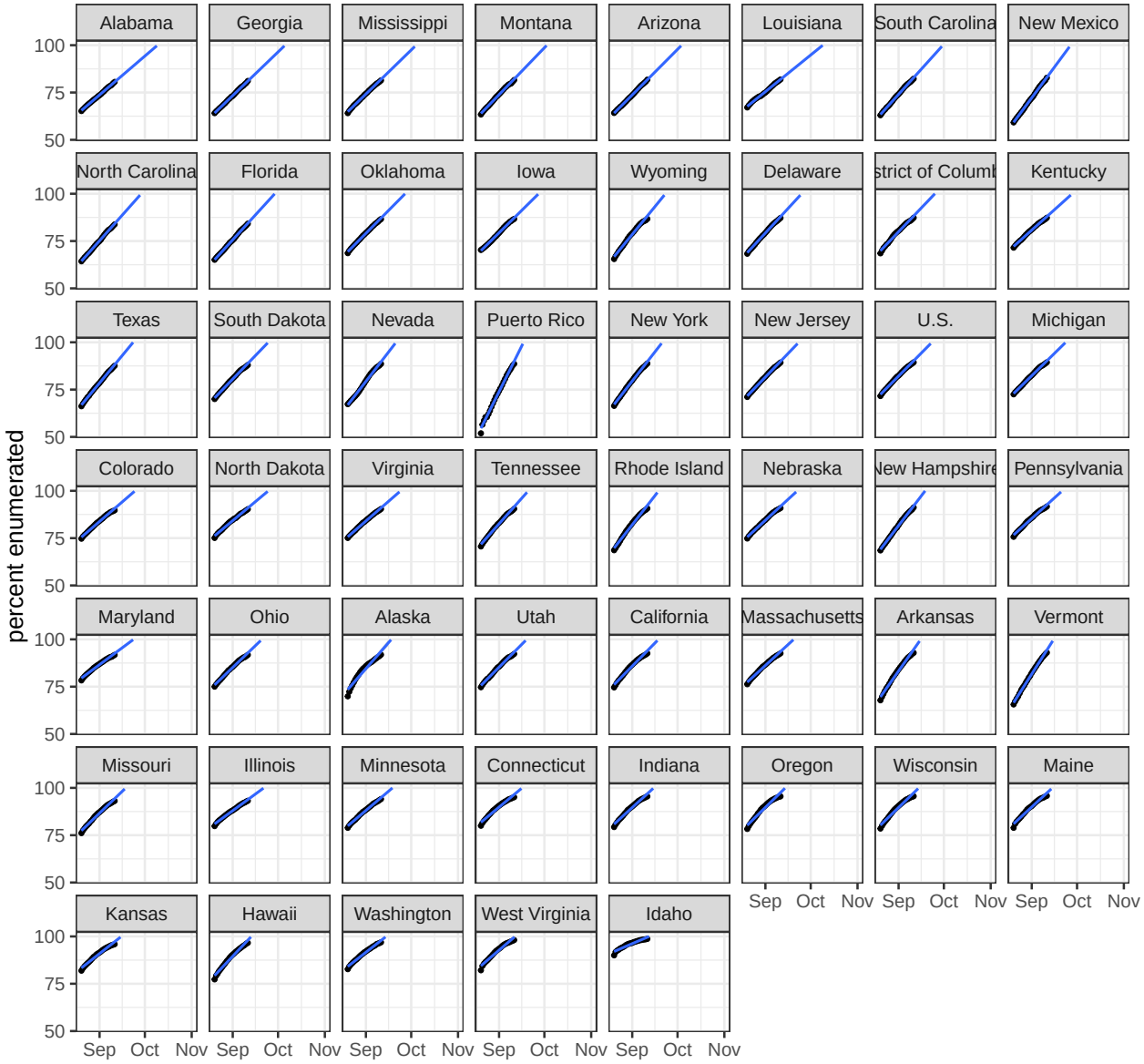

Figure 3: Percent counted if enumeration increases at current decreasing rate

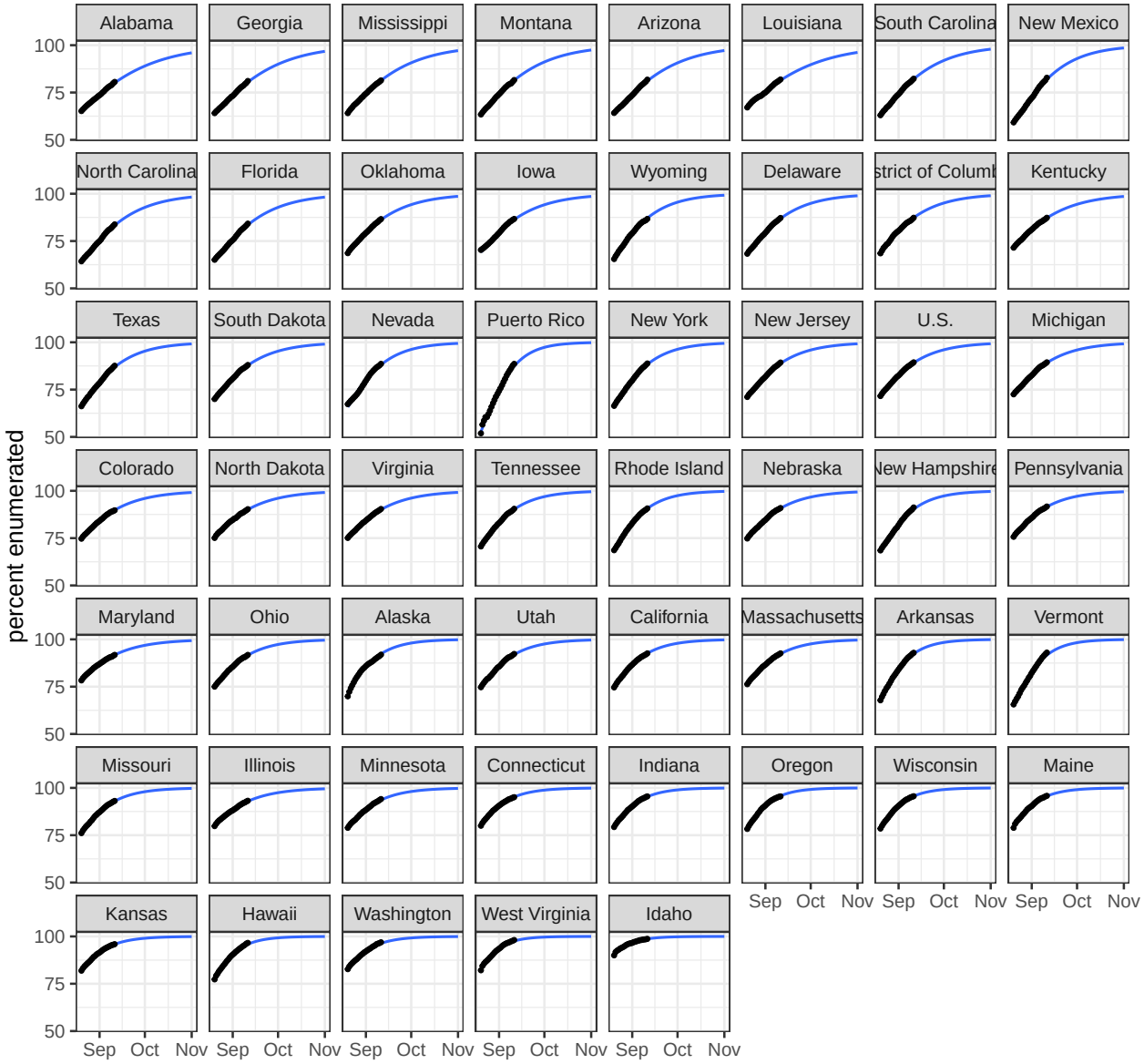

Figure 4: Estimated percent of households not enumerated by state if present trends continue (under current operations)

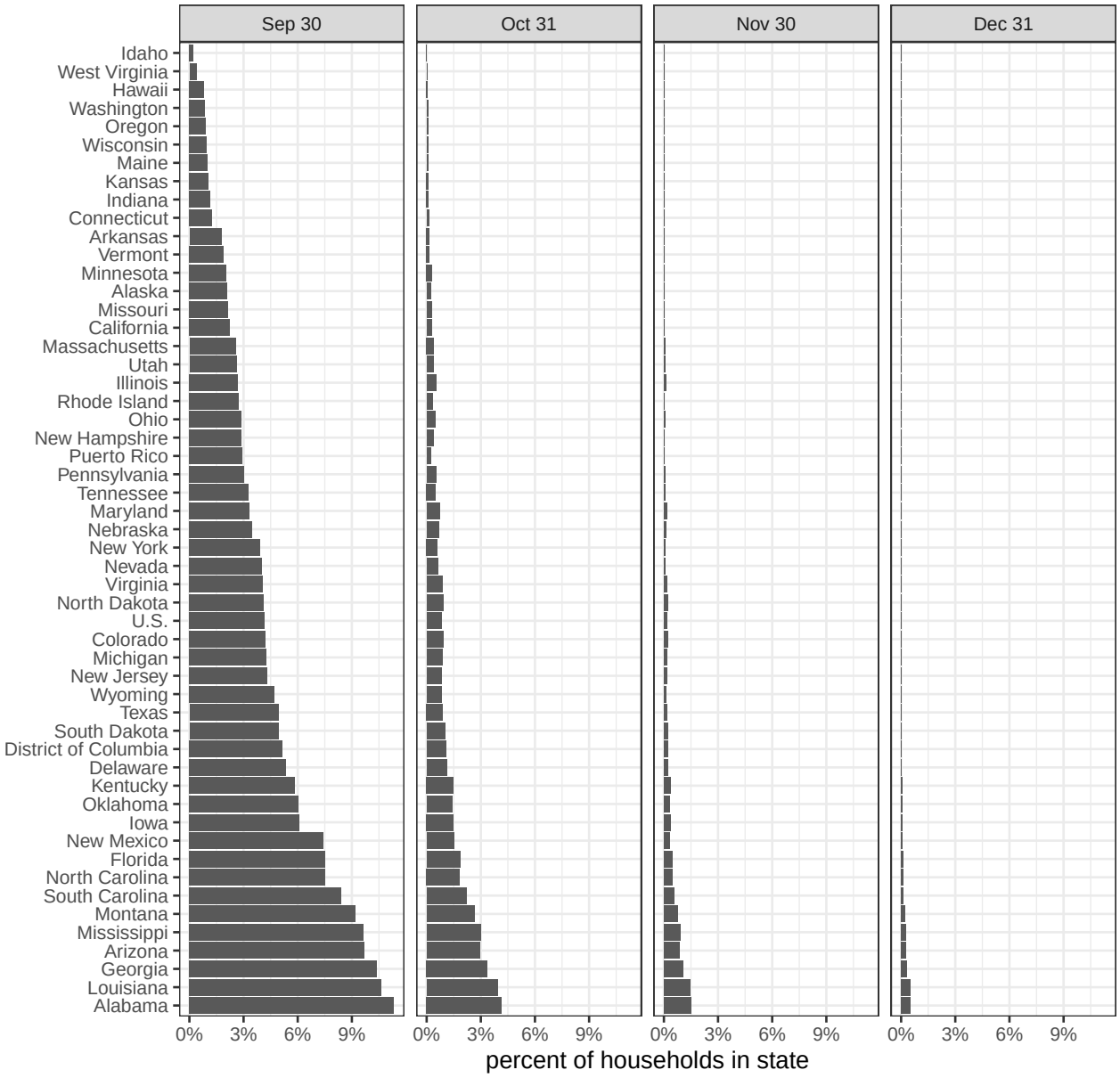

Figure 5: Number of persons correctly enumerated in each state is proportional to the number of households correctly enumerated (2010 Census)

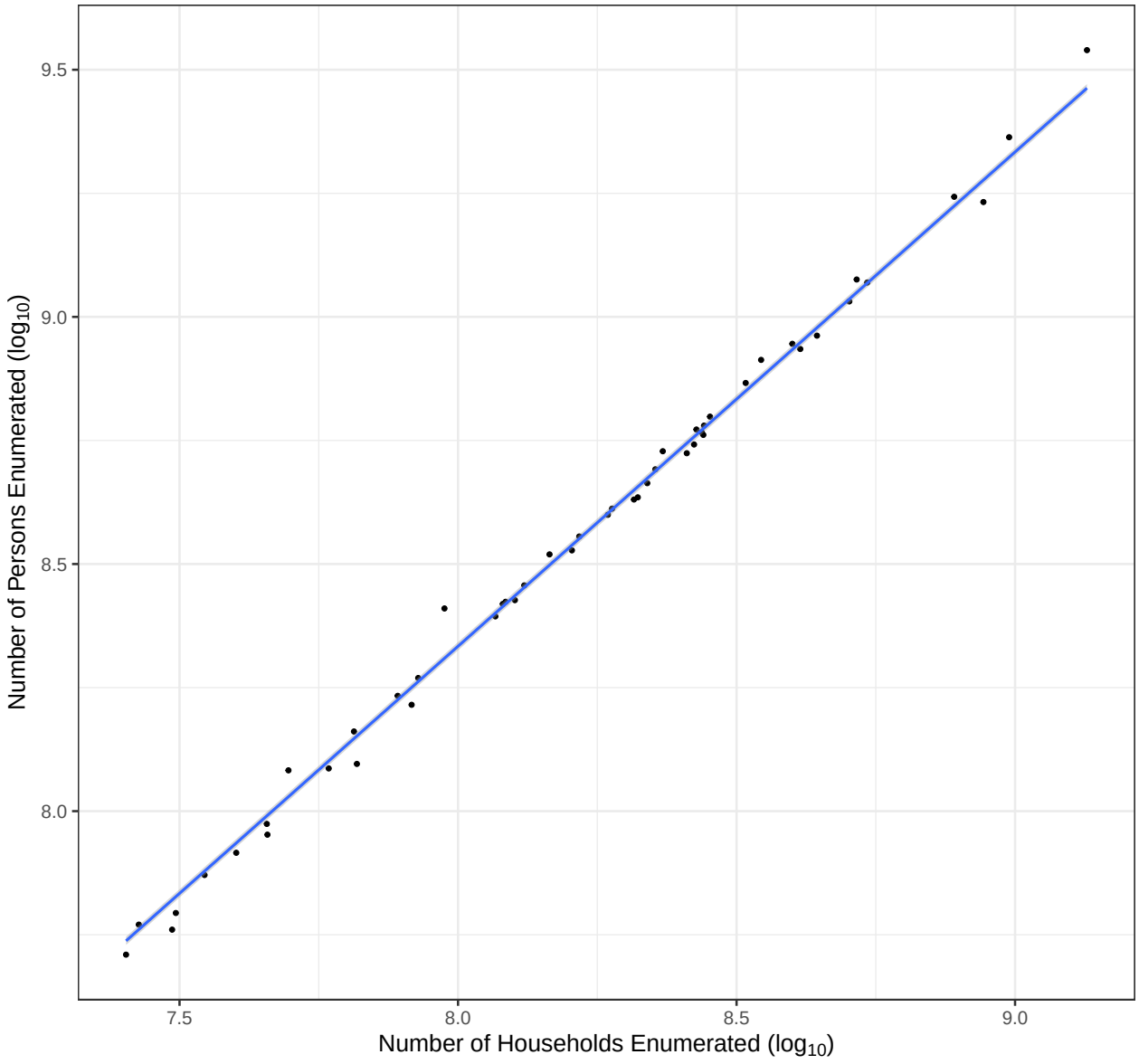

Figure 6: Number of person omissions in each state is proportional to the number of erroneous enumerations plus imputations (2010 Census)

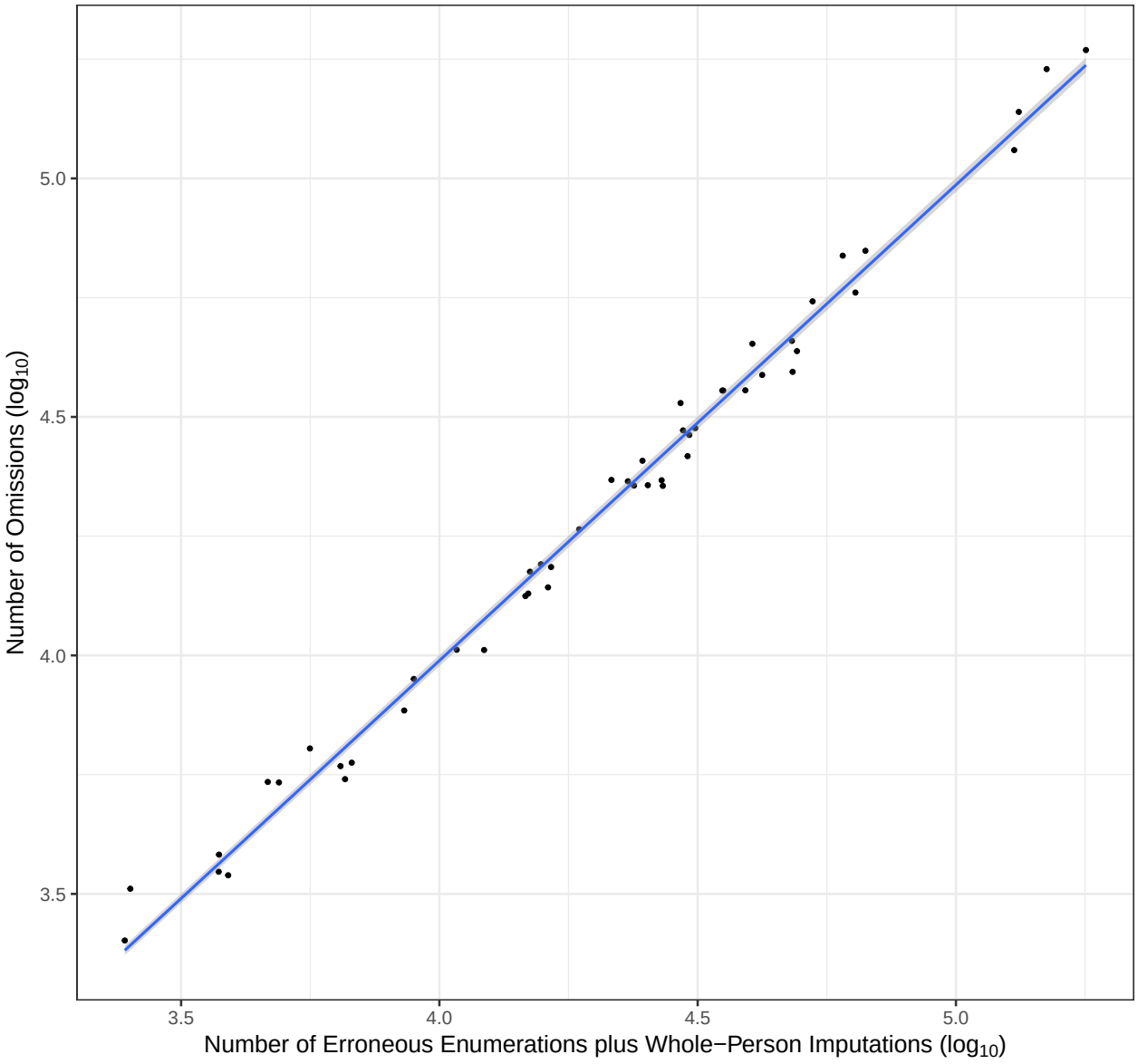

Figure 7.1: Change in apportionment if Census operations continue until select deadlines under scenario 1 (similar quality to 2010)

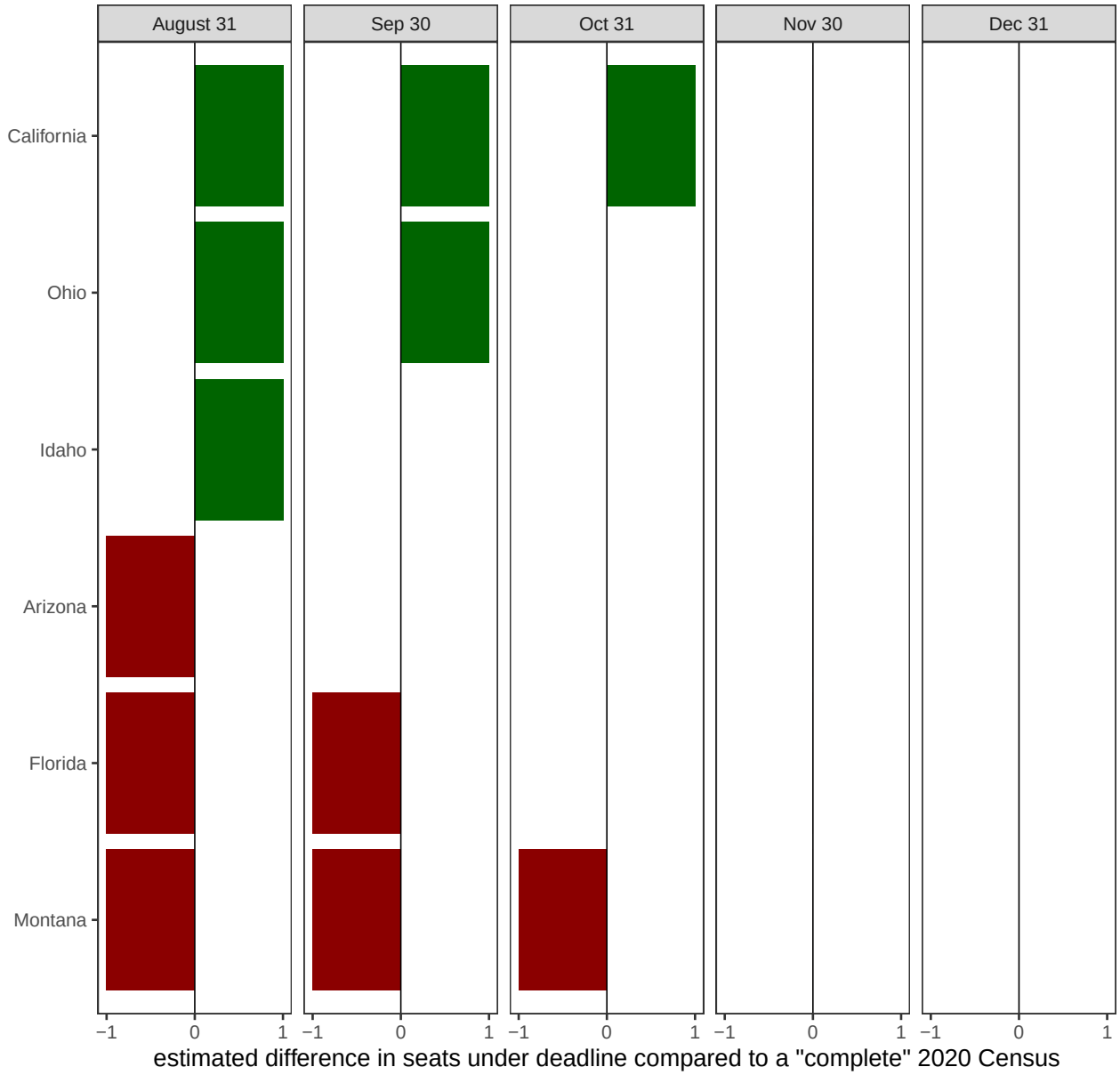

Figure 7.2: Change in apportionment if Census operations continue until select deadlines under scenario 2 (better quality than 2010)

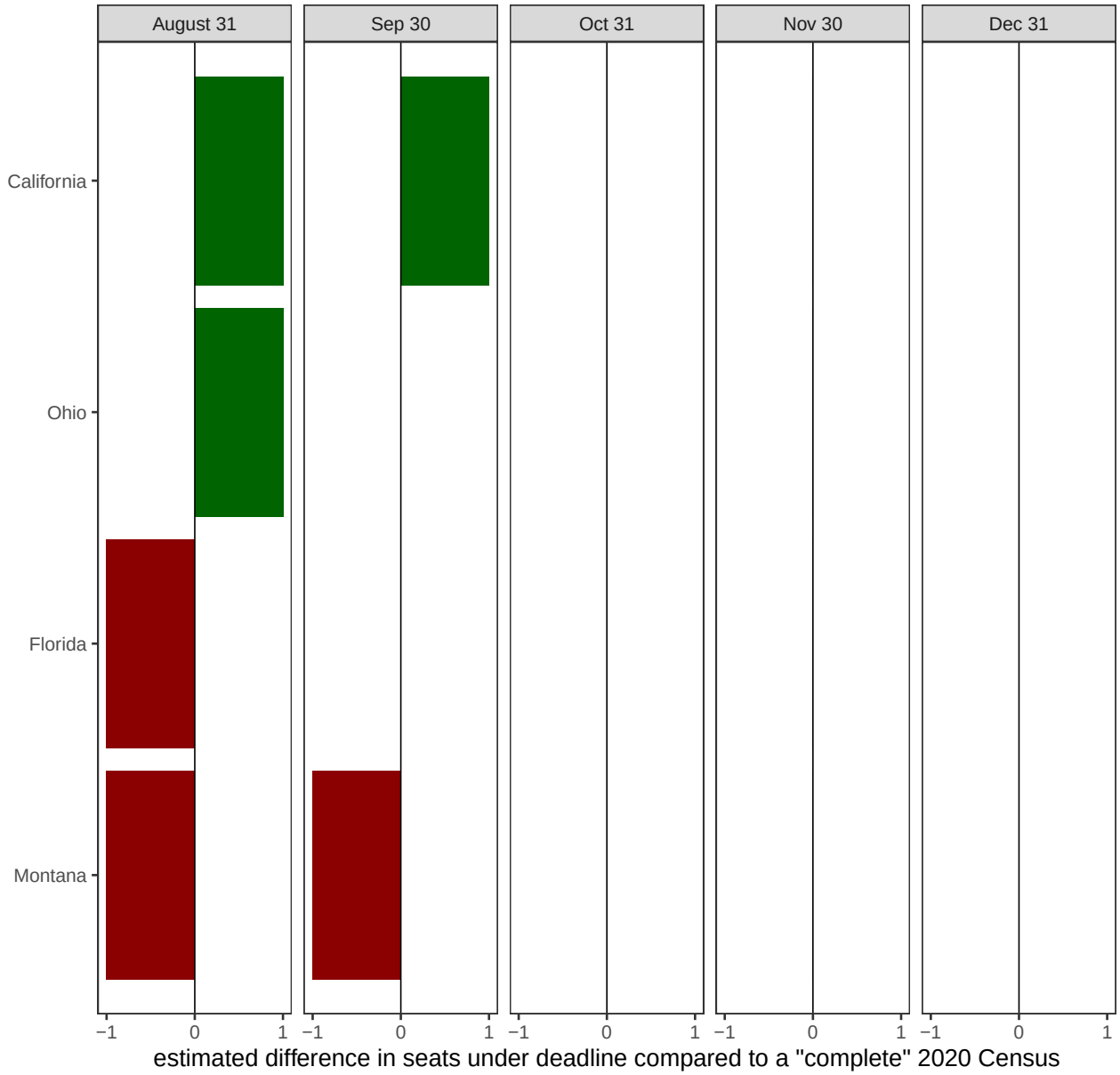

Figure 7.3: Change in apportionment if Census operations continue until select deadlines under scenario 3 (worse quality than 2010)

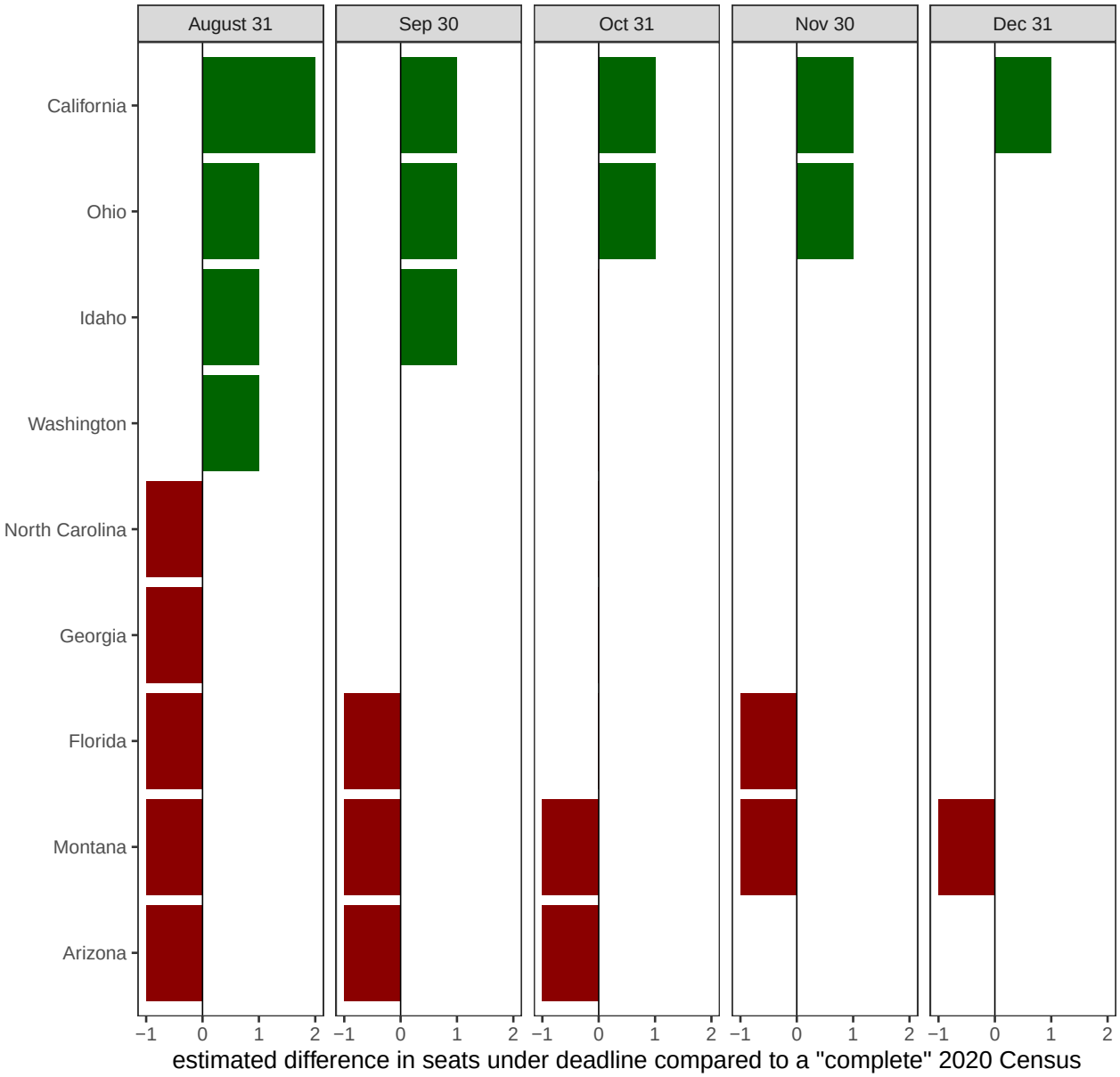

Figure 8.1: Loss of federal Medicaid funds if Census operations continue to select deadlines under scenario 1 (similar quality to 2010)

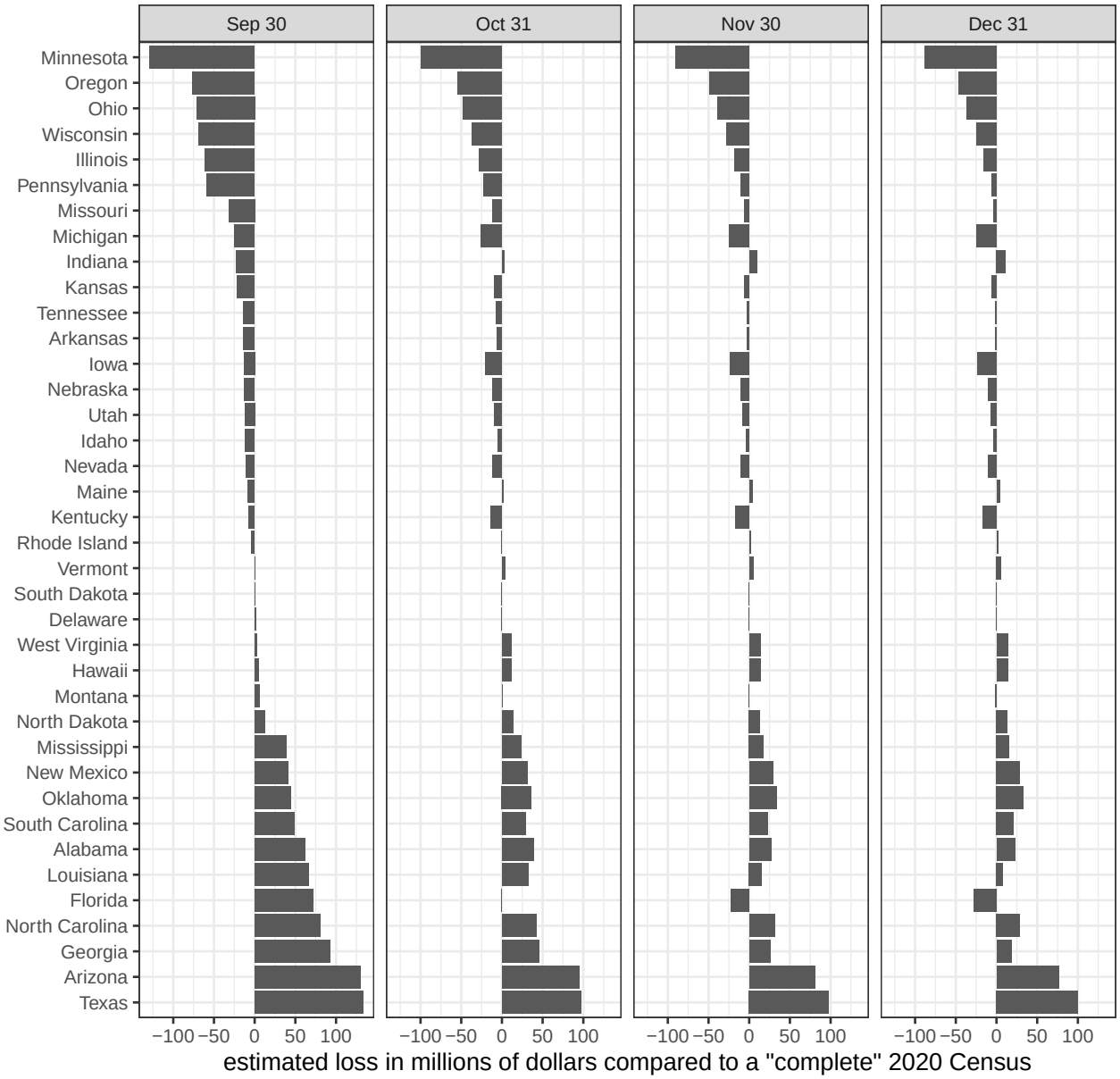

Figure 8.2: Loss of federal Medicaid funds if Census operations continue to select deadlines under scenario 2 (better quality than 2010)

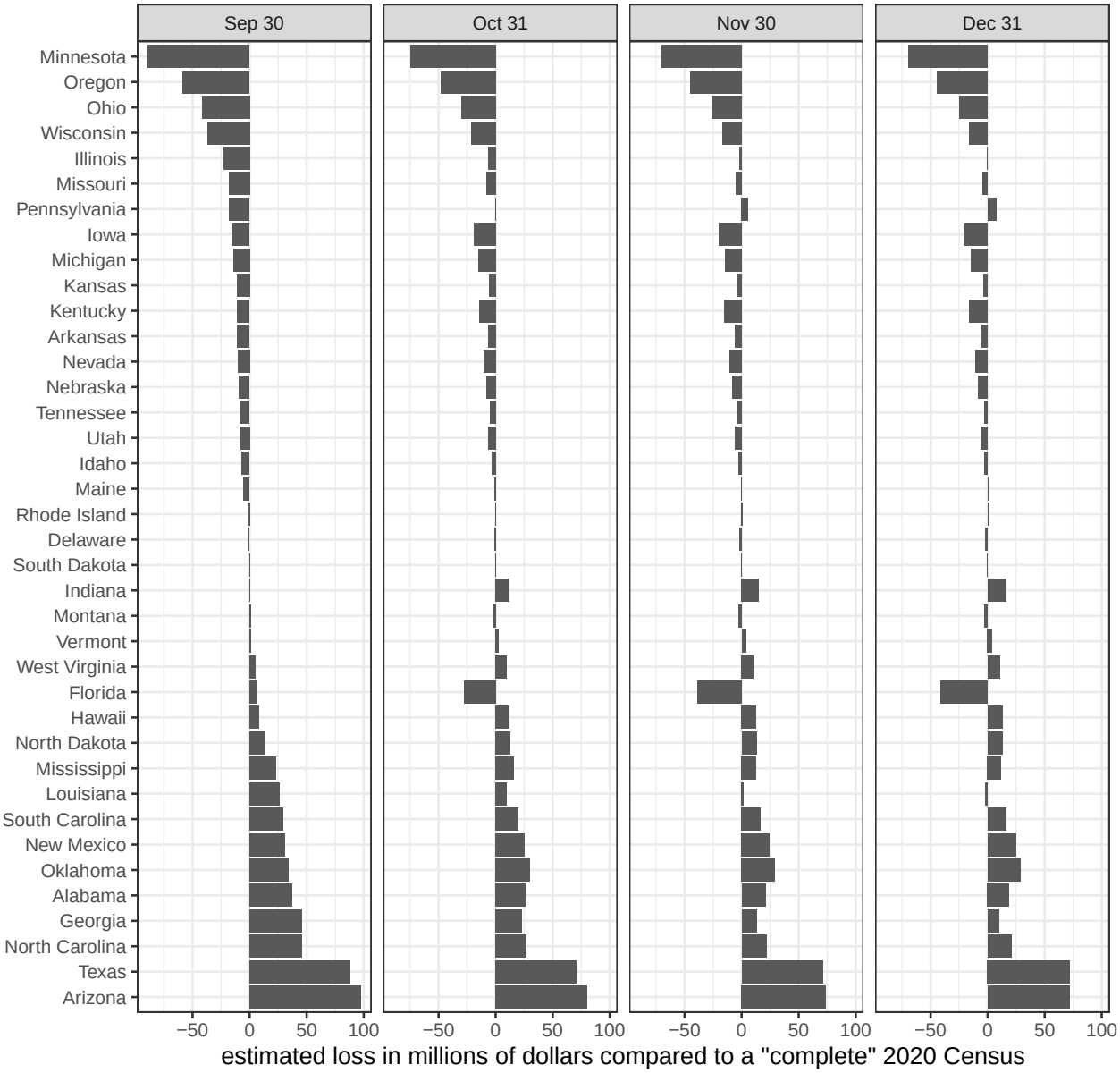

Figure 8.3: Loss of federal Medicaid funds if Census operations continue to select deadlines under scenario 3 (worse quality than 2010)

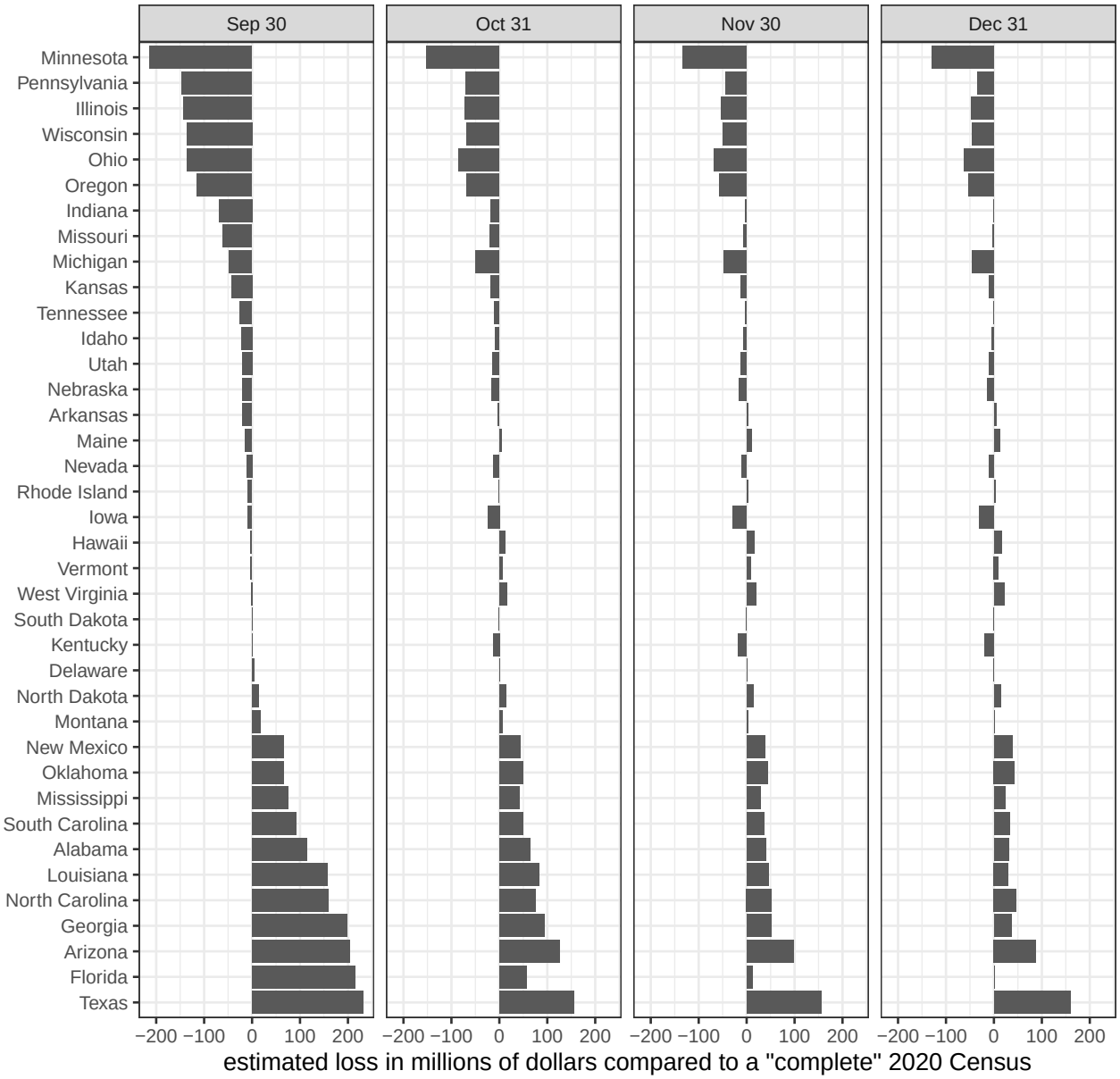

Table 1: The predicted percent of households enumerated by state under select deadlines if current enumeration rate continue

State	estimated percent enumerated by Sep 30	Oct 31	Nov 30	Dec 31
U.S.	95.82	99.19	99.84	99.97
Alabama	88.66	95.85	98.50	99.49
Alaska	97.89	99.76	99.97	100.00
Arizona	90.32	97.06	99.12	99.75
Arkansas	98.23	99.85	99.99	100.00
California	97.77	99.71	99.96	99.99
Colorado	95.79	99.07	99.79	99.96
Connecticut	98.73	99.86	99.98	100.00
Delaware	94.65	98.87	99.76	99.95
District of Columbia	94.85	98.92	99.77	99.95
Florida	92.49	98.13	99.53	99.89
Georgia	89.60	96.63	98.92	99.67
Hawaii	99.19	99.95	100.00	100.00
Idaho	99.78	99.99	100.00	100.00
Illinois	97.33	99.48	99.90	99.98
Indiana	98.86	99.89	99.99	100.00
Iowa	93.91	98.51	99.63	99.91
Kansas	98.96	99.89	99.99	100.00
Kentucky	94.17	98.51	99.62	99.91
Louisiana	89.39	96.07	98.56	99.50
Maine	99.00	99.91	99.99	100.00
Maryland	96.69	99.29	99.84	99.97
Massachusetts	97.43	99.60	99.93	99.99
Michigan	95.72	99.14	99.82	99.97
Minnesota	97.96	99.70	99.95	99.99
Mississippi	90.35	97.02	99.09	99.74
Missouri	97.85	99.71	99.96	99.99
Montana	90.78	97.35	99.24	99.80
Nebraska	96.51	99.34	99.87	99.98
Nevada	95.97	99.38	99.90	99.99
New Hampshire	97.10	99.63	99.95	99.99
New Jersey	95.68	99.16	99.83	99.97
New Mexico	92.56	98.48	99.69	99.94
New York	96.07	99.40	99.91	99.99
North Carolina	92.47	98.16	99.55	99.90
North Dakota	95.88	99.08	99.79	99.95
Ohio	97.13	99.53	99.92	99.99
Oklahoma	93.97	98.57	99.66	99.92
Oregon	99.10	99.93	99.99	100.00
Pennsylvania	96.95	99.46	99.90	99.98
Rhode Island	97.27	99.67	99.96	99.99
South Carolina	91.59	97.81	99.43	99.86
South Dakota	95.07	98.97	99.78	99.96
Tennessee	96.73	99.50	99.92	99.99
Texas	95.07	99.10	99.83	99.97
Utah	97.39	99.61	99.94	99.99

(continued)

State	estimated percent enumerated by Sep 30	Oct 31	Nov 30	Dec 31
Vermont	98.11	99.85	99.99	100.00
Virginia	95.94	99.13	99.81	99.96
Washington	99.14	99.92	99.99	100.00
West Virginia	99.61	99.98	100.00	100.00
Wisconsin	99.04	99.92	99.99	100.00
Wyoming	95.28	99.17	99.85	99.98
Puerto Rico	97.09	99.77	99.98	100.00

```
########################################
# Apportionment and Funding Projections #
########################################

# Date: 9/15/2020

# This R script explores apportionment of the House of Representatives and federal
## Medicaid funding if enumeration were to continue at its current pace until
## September 30th, October 31st, or other deadlines. The percent enumerated is
## forecast by fitting a logistic curve to the number enumerated to date between
## Aug 19 and Sep 12 via nonlinear least squares. (The maximum value parameter (aka
## capacity/asymptote) is set to 100 percent). The 2020 population is assumed to
## equal the Urban Institute's population estimate. Errors, imputations, and
## omissions are assumed to be similar to the 2010 Census.

# Please contact Jonathan Auerbach with any questions or corrections: jonathan@amstat.org

#read archived response rates
## source: https://2020census.gov/en/response-rates/nrfu.html#dd234650384-co
response_rates <- read_csv(file = "response_rates.csv")

##percent enumerated if response rate grows at current rate (linear growth)
response_rates %>%
  gather(key = "Source", value = "Response", -State) %>%
  mutate(Type = str_sub(Source, 1, 4),
         Date = as.Date(str_c(str_sub(Source, 6), " 2020"), format = "%B %d %Y")) %>%
  group_by(State, Date) %>%
  summarize(Response = sum(Response))  %>%
  ggplot() +
  aes(Date, Response) +
  geom_point() +
  facet_wrap(~ factor(State,
                levels = response_rates$State[
                  order(response_rates$`Self Sep 10` +
                        response_rates$`NRFU Sep 10`)])) +
  labs(title = "Figure 1: Percent of households counted between Aug 19 and Sep 12",
       x = "", y = "percent enumerated") +
  scale_y_continuous(breaks = c(50, 75, 100), limits = c(NA, 100)) +
  scale_x_date(breaks = c(as.Date("2020-8-24"),
                          as.Date("2020-9-07")),
               labels = c("8/24", "9/7"))

##percent enumerated if response rate grows at current rate (linear growth)
response_rates %>%
  gather(key = "Source", value = "Response", -State) %>%
  mutate(Type = str_sub(Source, 1, 4),
         Date = as.Date(str_c(str_sub(Source, 6), " 2020"), format = "%B %d %Y")) %>%
  group_by(State, Date) %>%
  summarize(Response = sum(Response))  %>%
  ggplot() +
  aes(Date, Response) +
  geom_point() +
  geom_smooth(method = "lm", fullrange = TRUE) +
```

```
  facet_wrap(~ factor(State,
                levels = response_rates$State[
                  order(response_rates$`Self Sep 10` +
                        response_rates$`NRFU Sep 10`)])) +
  labs(title = "Figure 2: Percent counted if enumeration increases at constant rate",
       x = "", y = "percent enumerated") +
  scale_y_continuous(breaks = c(50, 75, 100), limits = c(NA, 100)) +
  scale_x_date(breaks = c(as.Date("2020-8-31"),
                          as.Date("2020-9-30"),
                          as.Date("2020-10-31")),
               labels = c("Sep", "Oct", "Nov"),
               limits = c(as.Date("2020-08-19"), as.Date("2020-10-31")))


##percent enumerated if responses increase at current decreasing rate (logistic growth)
response_rates %>%
  gather(key = "Source", value = "Response", -State) %>%
  mutate(Type = str_sub(Source, 1, 4),
         Date = as.Date(str_c(str_sub(Source, 6), " 2020"),
                        format = "%B %d %Y")) %>%
  group_by(State, Date) %>%
  summarize(Response = sum(Response))  %>%
  ggplot() +
  aes(Date, Response) +
  geom_smooth(method="nls",
              formula=y ~ 100 / (1 + exp(-alpha - beta * x)),
              method.args = list(start=c(alpha = 1, beta = 1e-10)),
              se = FALSE,
              fullrange = TRUE)  +
  geom_point() +
  facet_wrap(~ factor(State,
                levels = response_rates$State[
                  order(response_rates$`Self Sep 10` +
                        response_rates$`NRFU Sep 10`)])) +
  labs(title = "Figure 3: Percent counted if enumeration increases at current
       decreasing rate",
       x = "", y = "percent enumerated") +
  scale_y_continuous(breaks = c(50, 75, 100), limits = c(NA, 100)) +
  scale_x_date(breaks = c(as.Date("2020-8-31"),
                          as.Date("2020-9-30"),
                          as.Date("2020-10-31")),
               labels = c("Sep", "Oct", "Nov"),
               limits = c(as.Date("2020-08-19"), as.Date("2020-10-31")))

population <- read_csv(file = "census_population.csv")
census_2010_person_errors <-
  read_csv(file = "census_2010_coverage_measurement_persons_table_14.csv")
census_2010_household_errors <-
  read_csv(file = "census_2010_coverage_measurement_households_table_6.csv")

census_2010_person_errors <-
  census_2010_person_errors %>%
  mutate(`Estimated Population (Thousands)` =
```

```
          `Census Count (Thousands)` * (1 - `Percent Undercount Est. (%)`))

census_2010_errors <-
census_2010_household_errors %>%
  select(State,
         `Number of Households (Thousands)` = `Census Count (Thousands)`,
         `Correct Household Enumerations (%)` = `Correct Enumerations (%)`) %>%
  left_join(
    census_2010_person_errors %>%
  select(State,
         `Number of Persons (Thousands)` = `Census Count (Thousands)`,
         `Correct Person Enumerations (%)` = `Correct Enumerations (%)`)) %>%
  filter(State != "U.S.")

fit_persons_counted_per_household <-
    lm(I(log10(
        1e3 * `Number of Persons (Thousands)` *
          `Correct Person Enumerations (%)`)) ~
      offset(I(log10(
        1e3 * `Number of Households (Thousands)` *
          `Correct Household Enumerations (%)`))),
      data = census_2010_errors)

fig5 <-
ggplot(census_2010_errors %>%
         filter(! (State %in% c("U.S.", "District of Columbia")))) +
  aes(log10(1e3 * `Number of Households (Thousands)` *
              `Correct Household Enumerations (%)`),
      log10(1e3 * `Number of Persons (Thousands)` *
              `Correct Person Enumerations (%)`)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ offset(x)) +
  labs(x = expression("Number of Households Enumerated ("*log[10]*")"),
       y = expression("Number of Persons Enumerated ("*log[10]*")"),
       title = "Figure 5: Number of persons correctly enumerated in each state is
       proportional to the number of households correctly enumerated (2010 Census)")

fig6 <-
census_2010_person_errors %>%
  filter(! (State %in% c("U.S.", "District of Columbia"))) %>%
  ggplot() +
  aes(log10(`Census Count (Thousands)` *
              (`Erroneous Enumerations Est. (%)` +
               `Whole-Person Census Imputations (%)`)),
      log10(`Census Count (Thousands)` * `Omissions Est. (%)`)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x + 0) +
  labs(x = expression("Number of Erroneous Enumerations
                       plus Whole-Person Imputations ("*log[10]*")"),
       y = expression("Number of Omissions ("*log[10]*")"),
       title = "Figure 6: Number of person omissions in each state is proportional
       to the number of erroneous enumerations plus imputations (2010 Census)")
```

```
pred_omitted <- function(imputs, errors, pop) {
  x1 <- with(census_2010_person_errors,
             log10(1e3 * `Census Count (Thousands)` *
                         `Erroneous Enumerations Est. (%)`))
  x2 <- with(census_2010_person_errors,
             log10(1e3 * `Census Count (Thousands)` *
                         `Whole-Person Census Imputations (%)`))
  y <- with(census_2010_person_errors,
             log10(1e3 * `Census Count (Thousands)` *
                         `Omissions Est. (%)`))
  fit <- lm(y ~ x1 + x2)
  predict(fit, newdata = data.frame(x1 = log10(pop * errors),
                                    x2 = log10(pop * imputs)))
}

  pred_percent_enumerated <- function(pct_self_respond,
                                      pct_early_nrfu,
                                      pct_late_nrfu,
                                      self_response_overcount,
                                      early_nrfu_overcount,
                                      late_nrfu_overcount)
    pct_self_respond * self_response_overcount +
    pct_early_nrfu * early_nrfu_overcount +
    pct_late_nrfu * late_nrfu_overcount

data_for_prediction <-
  response_rates %>%
  gather(key = "Source", value = "Response", -State) %>%
  mutate(Type = str_sub(Source, 1, 4),
         Date = as.Date(str_c(str_sub(Source, 6), " 2020"), format = "%B %d %Y")) %>%
  group_by(State, Date) %>%
  summarize(Response = sum(Response)) %>%
  mutate(date_numeric = as.numeric(Date) - min(as.numeric(Date)) + 1)

population_pred <- function(model_errors = FALSE,
                            error_self_response = 1,
                            error_early_nrfu = 1,
                            error_late_nrfu = 1) {

if(model_errors == FALSE) {
  population_pred <- tibble(State = response_rates$State,
                            Percent_by_Oct = 0,
                            Percent_by_Nov = 0,
                            Percent_by_Dec = 0,
                            Percent_by_Jan = 0)

for(state in unique(population_pred$State)) {
  fit_enumeration_completion <- nls(
    Response ~ 100 / (1 + exp(-alpha - beta * date_numeric)),
    start = c(alpha = 1, beta = 1e-10),
    data = data_for_prediction,
    subset = State == state)
```

```
  population_pred$Percent_by_Oct[population_pred$State == state] <-
    predict(fit_enumeration_completion, data.frame(date_numeric = 42))

  population_pred$Percent_by_Nov[population_pred$State == state] <-
    predict(fit_enumeration_completion, data.frame(date_numeric = 73))

  population_pred$Percent_by_Dec[population_pred$State == state] <-
    predict(fit_enumeration_completion, data.frame(date_numeric = 103))

  population_pred$Percent_by_Jan[population_pred$State == state] <-
    predict(fit_enumeration_completion, data.frame(date_numeric = 134))
    }
  }


if(model_errors == TRUE) {
population_pred <- tibble(State = response_rates$State,
                          Percent_by_Sep = 0,
                          Percent_by_Oct = 0,
                          Percent_by_Nov = 0,
                          Percent_by_Dec = 0,
                          Percent_by_Jan = 0)

for(state in population$State) {

  pop <- population$`Urban Population 2020`[population$State == state]

  self_response <- 1 -
    (pred_omitted(imputs = .2 * error_self_response,
                  errors = (2.1 + .3) * error_self_response,
                  pop = pop) / pop
     + (.2 + 2.1 + .3) * error_self_response) / 100

  early_nrfu <- 1 -
    (pred_omitted(imputs = 2.6 * error_early_nrfu,
                  errors = (3.7 + .6) * error_early_nrfu,
                  pop = pop) / pop
     + (2.6 + 3.7 + .6) * error_early_nrfu)  / 100

  late_nrfu <- 1 -
    (pred_omitted(imputs = 17.3 * error_late_nrfu,
                  errors = (6.8 + 1.2) * error_late_nrfu,
                  pop = pop) / pop
     + (17.3 + 6.8 + 1.2) * error_late_nrfu)  / 100

  fit_enumeration_completion <-
    nls(Response ~ 100 / (1 + exp(-alpha - beta * date_numeric)),
        start = c(alpha = 1, beta = 1e-10),
        data = data_for_prediction,
        subset = State == state)

  population_pred$Percent_by_Sep[population_pred$State == state] <-
    pred_percent_enumerated(
```

```
    pct_self_respond =
      response_rates$`Self Sep 10`[response_rates$State == state],
    pct_early_nrfu =
      (predict(fit_enumeration_completion, data.frame(date_numeric = 13)) -
         response_rates$`Self Sep 10`[response_rates$State == state]),
    pct_late_nrfu =
      100 - predict(fit_enumeration_completion, data.frame(date_numeric = 13)),
    self_response_overcount = self_response,
    early_nrfu_overcount = early_nrfu,
    late_nrfu_overcount = late_nrfu)

 population_pred$Percent_by_Oct[population_pred$State == state] <-
  pred_percent_enumerated(
    pct_self_respond =
      response_rates$`Self Sep 10`[response_rates$State == state],
    pct_early_nrfu =
      (predict(fit_enumeration_completion, data.frame(date_numeric = 42)) -
         response_rates$`Self Sep 10`[response_rates$State == state]),
    pct_late_nrfu =
      100 - predict(fit_enumeration_completion, data.frame(date_numeric = 42)),
    self_response_overcount = self_response,
    early_nrfu_overcount = early_nrfu,
    late_nrfu_overcount = late_nrfu)

  population_pred$Percent_by_Nov[population_pred$State == state] <-
  pred_percent_enumerated(
    pct_self_respond =
      response_rates$`Self Sep 10`[response_rates$State == state],
    pct_early_nrfu =
      (predict(fit_enumeration_completion, data.frame(date_numeric = 73)) -
         response_rates$`Self Sep 10`[response_rates$State == state]),
    pct_late_nrfu =
      100 - predict(fit_enumeration_completion, data.frame(date_numeric = 73)),
    self_response_overcount = self_response,
    early_nrfu_overcount = early_nrfu,
    late_nrfu_overcount = late_nrfu)

 population_pred$Percent_by_Dec[population_pred$State == state] <-
  pred_percent_enumerated(
    pct_self_respond =
      response_rates$`Self Sep 10`[response_rates$State == state],
    pct_early_nrfu =
      (predict(fit_enumeration_completion, data.frame(date_numeric = 103)) -
         response_rates$`Self Sep 10`[response_rates$State == state]),
    pct_late_nrfu =
      100 - predict(fit_enumeration_completion, data.frame(date_numeric = 103)),
    self_response_overcount = self_response,
    early_nrfu_overcount = early_nrfu,
    late_nrfu_overcount = late_nrfu)

  population_pred$Percent_by_Jan[population_pred$State == state] <-
  pred_percent_enumerated(
    pct_self_respond =
```

```
        response_rates$`Self Sep 10`[response_rates$State == state],
      pct_early_nrfu =
        (predict(fit_enumeration_completion, data.frame(date_numeric = 134)) -
           response_rates$`Self Sep 10`[response_rates$State == state]),
      pct_late_nrfu =
        100 - predict(fit_enumeration_completion, data.frame(date_numeric = 134)),
      self_response_overcount = self_response,
      early_nrfu_overcount = early_nrfu,
      late_nrfu_overcount = late_nrfu)
    }
  }
  population_pred
}

population_pred() %>%
  gather(key = "deadline", value = "amount", -State) %>%
  ggplot() +
  aes(factor(State,
             levels = State[deadline == "Percent_by_Oct"][
               order(amount[deadline == "Percent_by_Oct"])]),
             weight = 1 - amount / 100) +
  geom_bar() +
  facet_grid(~ factor(deadline,
                      levels = c("Percent_by_Oct",
                                 "Percent_by_Nov",
                                 "Percent_by_Dec",
                                 "Percent_by_Jan"),
                      labels = c("Sep 30", "Oct 31", "Nov 30", "Dec 31"))) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1)) +
  coord_flip() +
  labs(y = 'percent of households in state',
       x = "",
       title = "Figure 4: Estimated percent of households remaining by state
                if present trends continue (under current operations)")

fig5
fig6

#read population data
population_1 <- left_join(population, population_pred(model_errors = TRUE),
                          by = "State")
population_2 <- left_join(population, population_pred(model_errors = TRUE,
                                                      error_self_response = .5,
                                                      error_early_nrfu = .5,
                                                      error_late_nrfu = .5),
                          by = "State")
population_3 <- left_join(population, population_pred(model_errors = TRUE,
                                                      error_self_response = 2,
                                                      error_early_nrfu = 2,
                                                      error_late_nrfu = 2),
                          by = "State")

#function to compute apportionment
```

```r
census_multiplier <- function(n) 1/sqrt(n * (n - 1))

apportionment <- function(pop) {
  #1. calculate priority values for each state and 2 to 60 seats
  state_seat <- expand.grid(State = seq_along(population$State),
                            Seat  = 2:60)
  state_seat$`Priority Value` <-
    mapply(function(i, j) census_multiplier(j) * pop[i],
           i = state_seat$State,
           j = state_seat$Seat
    )
  #2. each state gets one "free" seat
  assignment <- tibble(State = population$State,
                       `House Seats` = 1)
  #3. rank state seats by priority value and assign the first 385 seats
  for(rank in 1:385) {
    State <- assignment$State[
      state_seat$State[order(state_seat$`Priority Value`,
                             decreasing = TRUE)][rank]
    ]
    assignment$`House Seats`[assignment$State == State] <-
      assignment$`House Seats`[assignment$State == State] + 1
  }
  assignment
}

plot_apportionment <- function(population) {
  tibble(
    State = apportionment(population$`Population 2010`)$`State`,
    `Current Number of Seats` =
      apportionment(population$`Population 2010`)$`House Seats`,
    `Expected Number Full Population` =
      apportionment(population$`Urban Population 2020`)$`House Seats`,
    `Expected Number Currently Enumerated` =
      apportionment(population$`Urban Population 2020` *
                      population$Enumerated / 100)$`House Seats`,
    `Expected Number Self Reported` =
      apportionment(population$`Urban Population 2020` *
                      population$`Self Respond` / 100)$`House Seats`,
    `Expected Number Projection Sep 1` =
      apportionment(population$`Urban Population 2020` *
                      population$`Percent_by_Sep` / 100)$`House Seats`,
    `Expected Number Projection Oct 1` =
      apportionment(population$`Urban Population 2020` *
                      population$`Percent_by_Oct` / 100)$`House Seats`,
    `Expected Number Projection Nov 1` =
      apportionment(population$`Urban Population 2020` *
                      population$`Percent_by_Nov` / 100)$`House Seats`,
    `Expected Number Projection Dec 1` =
      apportionment(population$`Urban Population 2020` *
                      population$`Percent_by_Dec` / 100)$`House Seats`,
    `Expected Number Projection Jan 1` =
      apportionment(population$`Urban Population 2020` *
```

```
                      population$`Percent_by_Jan` / 100)$`House Seats`
  ) %>%
  transmute(State,
            `Sep 1` = `Expected Number Projection Sep 1` -
              `Expected Number Full Population`,
            `Oct 1` = `Expected Number Projection Oct 1` -
              `Expected Number Full Population`,
            `Nov 1` = `Expected Number Projection Nov 1` -
              `Expected Number Full Population`,
            `Dec 1` = `Expected Number Projection Dec 1` -
              `Expected Number Full Population`,
            `Jan 1` = `Expected Number Projection Jan 1` -
              `Expected Number Full Population`) %>%
  filter(`Sep 1` + `Oct 1` + `Nov 1` + `Dec 1` + `Jan 1` != 0) %>%
  gather(key = "population", value = "gain", -State) %>%
  ggplot() +
  aes(factor(State,
             levels =
               State[order(gain[population == "Sep 1"] +
                             .5 * gain[population == "Oct 1"] +
                             .25 * gain[population == "Nov 1"])]),
      weight = gain,
      fill = gain > 0) +
  geom_bar() +
  coord_flip() +
  facet_grid(~ factor(population,
                      levels = c("Sep 1",
                                 "Oct 1",
                                 "Nov 1",
                                 "Dec 1",
                                 "Jan 1"))) +
  geom_hline(yintercept = 0, color = "black")  +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
  legend.position = "none") +
  scale_fill_manual(values = c("TRUE" = "dark green", "FALSE" = "dark red")) +
  scale_y_continuous(breaks = c(-2, -1, 0, 1, 2))
}

plot_apportionment(population_1) +
  labs(y = 'estimated difference in seats under deadline',
       x = "",
       title = "Figure 7.1: Change in apportionment if Census operations continue until
                   select deadlines under scenario 1 (similar quality to 2010)")

plot_apportionment(population_2) +
  labs(y = 'estimated difference in seats under deadline',
       x = "",
       title = "Figure 7.2: Change in apportionment if Census operations continue until
                   select deadlines under scenario 2 (better quality than 2010)")

plot_apportionment(population_3) +
  labs(y = 'estimated difference in seats under deadline',
       x = "",
```

```
       title = "Figure 7.3: Change in apportionment if Census operations continue until
                   select deadlines under scenario 3 (worse quality than 2010)")

#read BEA Personal Income data
#BEA site: https://apps.bea.gov/iTable/index_regional.cfm
bea_personal_income <- read_csv("bea_income.csv")
reamer_fy15_expenditures <- read_csv("reamer_table_3_1.csv")

fmap <- function(pop) {
  personal_income <-
    bea_personal_income %>%
    filter(Description ==
             "Personal income (millions of dollars, seasonally adjusted)") %>%
    transmute(State, `Personal Income` = 1e6 * `2020:Q1`) %>%
    left_join(pop, by = "State")

  if("U.S." %in% personal_income$State[is.na(personal_income$population)])
    personal_income$population[is.na(personal_income$population)] <-
      sum(personal_income$population, na.rm = TRUE)

  personal_income %>%
    filter(State != "United States") %>%
    mutate(`Income per Capita` = `Personal Income`/`population`,
           `U.S. Income per Capita` = personal_income %>%
                                      filter(State == "U.S.") %>%
                                      mutate(`Income per Capita` =
                                             `Personal Income`/`population`) %>%
                                      pull(`Income per Capita`),
           fmap = 1 - .45 * (`Income per Capita`/`U.S. Income per Capita`)^2,
           fmap = ifelse(fmap < .5, .5, ifelse(fmap > .83, .83, fmap)) ) %>%
    transmute(State, fmap) %>%
    filter(! (State %in% c("U.S.", "District of Columbia") )) %>%
    left_join(reamer_fy15_expenditures %>%
                transmute(State,
                          Medicaid = 100 * `Medicaid Traditional` / FMAP +
                                     100 * `Medicaid Medicare Part D Clawback` /
                            EFMAP)) %>%
    mutate(`Federal Portion` = Medicaid * fmap)
    }

fmap_plot <- function(population)
  tibble(fmap(bea_personal_income %>%
              filter(Description == "Population (midperiod, persons) 1/") %>%
              transmute(State, `population` = `2020:Q1`)) %>%
         transmute(State, `Federal Portion`),
       `Oct 1` = fmap(population %>%
                        transmute(State,
                                  population = Percent_by_Oct *
                                               `Urban Population 2020`)) %>%
         pull(`Federal Portion`) - `Federal Portion`,
       `Nov 1` = fmap(population %>%
                        transmute(State,
                                  population = Percent_by_Nov *
```

```
                                                `Urban Population 2020`)) %>%
          pull(`Federal Portion`) - `Federal Portion`,
        `Dec 1` = fmap(population %>%
                         transmute(State,
                                   population = Percent_by_Dec *
                                                `Urban Population 2020`)) %>%
          pull(`Federal Portion`) - `Federal Portion`,
        `Jan 1` = fmap(population %>%
                         transmute(State,
                                   population = Percent_by_Jan *
                                                `Urban Population 2020`)) %>%
          pull(`Federal Portion`) - `Federal Portion`
) %>%
  select(- `Federal Portion`) %>%
  filter(`Oct 1` + `Nov 1` + `Dec 1` + `Jan 1` != 0) %>%
  gather(key = "deadline", value = "amount", -State) %>%
  ggplot() +
  aes(factor(State,
             levels =
               State[order(amount[deadline == "Oct 1"] #+
                             #.5 * amount[deadline == "Nov 1"] +
                             #.25 * amount[deadline == "Dec 1"] +
                             #.125 * amount[deadline == "Jan 1"]
                           )]),
      weight = -amount/1e6) +
  geom_bar() +
  facet_grid(~ factor(deadline,
                      levels = c("Oct 1", "Nov 1", "Dec 1", "Jan 1"))) +
  coord_flip() +
  labs(x = "", y = "")

fmap_plot(population_1) +
  labs(x = "",
       title = "Figure 8.1: Loss of federal Medicaid funds if Census operations continue
                   to select deadlines under scenario 1 (similar quality to 2010)",
       y = 'estimated loss in millions of dollars compared to a "complete" 2020 Census')

fmap_plot(population_2) +
  labs(x = "",
       title = "Figure 8.2: Loss of federal Medicaid funds if Census operations continue
                   to select deadlines under scenario 2 (better quality than 2010)",
       y = 'estimated loss in millions of dollars compared to a "complete" 2020 Census')

fmap_plot(population_3) +
  labs(x = "",
       title = "Figure 8.3: Loss of federal Medicaid funds if Census operations continue
                   to select deadlines under scenario 3 (worse quality than 2010)",
       y = 'estimated loss in millions of dollars compared to a "complete" 2020 Census')
```