Using IPUMS data in R with ipumsr

October 12, 2021 (10:00 a.m. - 11:00 a.m. CDT)

QUESTIONS AND ANSWERS

This document provides answers to the questions received during the live webinar. To help with navigation, we have grouped related questions under topical headings.

For more user support, email IPUMS at ipums@umn.edu.

How to access webinar materials

Will the recording be shared with the registered participants? (Share the presentation markdown repo, too, please!)

Yes, the recording is posted on the IPUMS Tutorials page, and the files used to create the presentation, including all the example code from the presentation, are available in this GitHub repository.

Downloading IPUMS data

Can we download the data directly to a server?

Currently, you can only download your IPUMS data using a web browser with a graphical user interface (GUI), so it might be difficult to download directly to a server unless you are running your web browser on that server. Once the IPUMS USA data extract API is live, however, you will be able to download your data programmatically using programs such as R, Python, and curl, so it should be easier to download data directly onto a server.

Is it possible to download the DDI codebook file for extracts that are no longer available on the My Data page?

No, both the data and DDI codebook files are removed from your My Data page 72 hours after they become available, so you'll have to resubmit your extract request to get access to the DDI for an extract older than 72 hours.

Instead of downloading data files from the website, for example from IPUMS USA, can you download data directly into RStudio with R code?

Currently you can only create extracts and download your IPUMS data by browsing the website, but once the IPUMS USA data extract API is publicly available (expected in early 2022), you will be able to define and submit extracts and download your data directly from your R session.

Reading IPUMS data

Do you need to unzip the extract data file (the file with the ".dat.gz" file extension) before reading it with ipumsr?

No, you do not need to unzip the data file.

Can I analyze IPUMS data with SAS?

Yes, R is certainly not the only way to analyze IPUMS data. Check out this tutorial on opening your IPUMS data with SAS.

Can ipumsr only read fixed-width (.dat or .dat.gz) IPUMS data files?

ipumsr can read fixed-width formatted (.dat or .dat.gz) data file or CSV formatted (.csv or .csv.gz) IPUMS data files, so if you plan to read your data file with ipumsr, be sure to request either the "Fixed-width text (.dat)" or the "Comma delimited (.csv)" option when creating your extract. ipumsr cannot read a Stata, SPSS, or SAS formatted IPUMS data file, but you can read such files with other R packages, such as haven.

I encountered a problem reading data probably due to the large file size. The error message says: Error: cannot allocate vector of size xxx.x Mb. I wonder whether there is a way to address that? Thank you a lot!

If your extract is too large to read into memory, check out the ipumsr vignette on working with big data and this helpful walkthrough from Kyle Walker (author of the tidycensus R package) on using a local database to analyze a big IPUMS dataset. If you want to be able to explore your extract without reading the whole file in at once, don't forget about the `n_max` argument to `read_ipums_micro()`, which allows you to read in only a specified number of records (e.g. `read_ipums_micro(my_ddi, n_max = 10000)`).

Installing ipumsr

I got the following error when I tried to install the package? Any ideas?

```
Error in read.dcf(file.path(pkgname, "DESCRIPTION"), c("Package",
"Type")) :
  cannot open the connection
In addition: Warning messages:
1: In unzip(zipname, exdir = dest) :
  write error in extracting from zip file
2: In unzip(zipname, exdir = dest) :
  write error in extracting from zip file
3: In read.dcf(file.path(pkgname, "DESCRIPTION"), c("Package",
"Type")) :
  cannot open compressed file 'sp/DESCRIPTION', probable reason 'No
such file or directory'
```

Errors like these depend a lot on your specific setup, so if you see something like this, it's best to email the IPUMS user support team at ipums@umn.edu and provide as much detail as possible, including the exact commands you submitted and the full error message you received. Sometimes the error message will contain clues, though -- for example, this error message mentions the package "sp" (trying to open file 'sp/DESCRIPTION'), so you could try restarting

your R session and submitting `install.packages("sp")` to attempt to reinstall that package.

Should I install ipumsr from CRAN, or should I install the development version from GitHub?

Almost everyone should install ipumsr from CRAN using `install.packages("ipumsr")`. You should only install the development version from GitHub if you have a specific reason to do so, such as wanting to try out new features that are being actively developed. The downside of using the development version is that some functionality might not be thoroughly tested or documented, and the interface of new functions (such as the function name and arguments) might change before they make it into the package version on CRAN.

The API functions previewed in this webinar are an example of a new feature you might want to try out before it's available on CRAN, but the usual caveats apply: These new functions might contain undiscovered bugs, and the interface of these functions might change rapidly with little warning. In fact, the API functions won't work at all right now, because the API is not publicly available yet, but we do expect to open it up to beta testers before the end of 2021. To install the current development version (as of 10/18/2021), first make sure that you have installed the remotes package (use `install.packages("remotes")` if you don't have it yet), then use `remotes::install_github("mnpopcenter/ipumsr", ref = "api-alpha-dev")` to install the package from the "api-alpha-dev" branch. If you get an error because that branch no longer exists, try `remotes::install_github("mnpopcenter/ipumsr")`.

Working with value labels

Does ipumsr store labeled variables as haven_labelled objects (from package haven) or in the style of the package sjlabelled?

ipumsr stores labeled variables as haven_labelled objects.

Is there a user-friendly guide that describes some of these tips for working with value labels?

ipumsr includes bundled "vignettes" (extended examples) on selected topics, including reading in your data, working with value labels, working with IPUMS geographic data, and working with large IPUMS datasets.

The label helper functions seem like they'd be useful with tidycensus data too -- am I understanding that correctly?

The label helper functions in ipumsr (such as `lbl_na_if()`, `lbl_collapse()`, and `lbl_relabel()`) are designed to manipulate haven_labelled objects. As far as we know, the tidycensus R package does not store data in haven_labelled objects, so these helper functions are probably not readily applicable to tidycensus data. But if we misunderstood your question, please reach out to IPUMS Support at ipums@umn.edu!

Is there a difference between the ipumsr function lbl_relabel() and the case_when() function from the dplyr package?

Yes, these functions are different, and not interchangeable, though they look similar because they both use two-sided formulas to recode values. For one thing, in `lbl_relabel()`, the new value to assign is on the left-hand side of the formula, and the conditional that indicates which values to recode is on the right-hand side. The opposite is true for `case_when()`. Most importantly, however, `case_when()` cannot easily recode haven_labelled objects without stripping away variable and value labels, which is an example of why we created label helper functions such as `lbl_relabel()`.

Is there a library or something of snippets to help me do common recoding like in the education level example?

The ipumsr vignette on working with value labels includes a few examples of using these functions. There are also examples in the function documentation for the label helper functions, which are all listed on the function reference page under the "Work with value labels" heading. You can also view the function documentation for a given function, like `lbl_relabel()`, by submitting the command `?lbl_relabel` in your R session.

Survey weights

How would you add weights to the code you're showing now?

The code on slides 62 and 71 from the webinar show simple examples of using person weights to summarize IPUMS USA data (the slides are available on the IPUMS Tutorials page). For questions about more advanced usage of weights, check out the survey and srvyr R packages, or contact IPUMS Support at ipums@umn.edu. And if you have any ideas for how ipumsr could make using IPUMS weights easier, create an issue on the ipumsr GitHub repository.

Does this webinar cover how to estimate proper standard errors using the replicate weights? If not, will this added functionality in the new version of the package? Current documentation is lacking in how to do so in R.

This webinar does not explain how to use replicate weights in R, nor is there active development of new ipumsr functions to work with replicate weights, but we'd love to provide better support in this area! There is some potentially helpful code on the page for this GitHub issue created by ipumsr package developer Greg Freedman Ellis, and we would encourage users who want more guidance on this topic to comment on that GitHub issue with additional requests. A particularly helpful request would be if there is some documentation that is available from IPUMS for other statistical packages, but not for R, and you just want us to translate that into R. We don't always notice these holes in our documentation, so if you can point them out, we appreciate it!

Can I import IPUMS datasets, for example ACS, taking into consideration the survey design? I want to use the survey package as I use the svy commands in STATA.

Yes, you can use the survey R package with IPUMS data. There are some examples of how to do this on the page for this GitHub issue created by ipumsr package developer Greg Freedman Ellis, but please reach out to IPUMS User Support at ipums@umn.edu if you have additional questions.

How do I convert an IPUMS dataset into a survey object, as used by the survey and srvyr R packages?

There is some guidance on this question on the page for this GitHub issue created by ipumsr package developer Greg Freedman Ellis, and we would encourage users who want more guidance on this topic to comment on that GitHub issue with additional requests. If there is enough interest, we could create a vignette on this topic -- let us know what would be helpful, and be as specific as possible.

API

When do you expect the API to be available?

We expect the IPUMS USA data extract API to be publicly available in the first few months of 2022, depending on what issues come up during beta testing. We are currently conducting internal testing of the API client tools, and plan to send an email to all users requesting beta testers before the end of 2021.

How do we contact you to become a beta tester for the IPUMS USA data extract API?

Email the IPUMS user support team at ipums@umn.edu, and provide your name and the email address at which we should contact you when beta testing begins.

Will we need to sign in to our IPUMS account to use the IPUMS USA data extract API?

To use the IPUMS USA data extract API, you will need to be a registered IPUMS USA user. If you are not yet registered for IPUMS USA, register here. Once registered, you will need to create an API key associated with your IPUMS account. You will include that API key to authenticate your API requests. And one more reminder: *The IPUMS USA data extract API will not be publicly available until the first few months of 2022, so your API key will not work with the IPUMS USA API until after that public launch.* However, the same API key can be used for the IPUMS NHGIS API, which is already publicly available.

Is it right to say that the API is not well-suited to analyze large data, since waiting for an extract can take quite a long time (e.g. a day)?

It is true that large IPUMS extracts can take hours to process, and even small extracts do not process instantaneously, which means that the IPUMS data extract API will not be well-suited for some use cases that require data on demand. However, we still expect that the data extract API will be useful for creating and sharing the definitions of large extracts, because it will allow for the creation and downloading of those extracts programmatically, without the need to interact with a graphical user interface.

Other topics

Does IPUMS publish K-12 standardized test score results?

We are not aware of any IPUMS datasets that contain information about standardized test scores. Please reach out to ipums@umn.edu for more targeted help from the user support team regarding your data needs.

Where is the ipumsr GitHub repository?

https://github.com/mnpopcenter/ipumsr

Are ipumsr functions compatible with the pipe operator?

Yes, ipumsr functions are generally compatible with the pipe (`%>%`, or `|>` in base R versions 4.1 and greater) operator. For any specific questions about using the pipe with ipumsr, email us at ipums+cran@umn.edu.

Is there any code on the GitHub page regarding building the family interrelationships?

No, the code used to build family interrelationship variables is not easily accessible to IPUMS users. All IPUMS data collections that include family interrelationship variables have documentation describing how these variables are constructed (for example, see the documentation for IPUMS USA variable SPRULE). If you have specific questions about the family interrelationship variables, email IPUMS Support at ipums@umn.edu.

How do I install R and RStudio?

Check out this guide to installing R and RStudio from the RStudio team.

Do I have to use R to analyze IPUMS data?

No, you can analyze IPUMS data with any statistical software that can read a comma-delimited data file. In addition to R, IPUMS provides particular support for Stata, SPSS, SAS, and Excel. For more information on getting your IPUMS data into one of those stats packages, check out the relevant tutorials under the "Other Helpful Tutorials" heading on the IPUMS Tutorials page.

Can I use IPUMS Terra with ipumsr?

Yes, ipumsr can read data from IPUMS Terra. This is covered in the intro to ipumsr vignette and the documentation for the ipumsr functions that read IPUMS Terra data -- try navigating to the ipumsr function reference page and searching for "read_terra". Email us at ipums@umn.edu if you have additional questions!

Does ipumsr help with linking survey respondents (for example in CPS) either across time or across different datasets?

ipumsr does not provide any special support for linking respondents across time or different datasets, but check out this documentation on linking from IPUMS CPS, the "Linking CPS (to CPS and ATUS)" webinar posted on the IPUMS Tutorials page, or email the IPUMS user

support team at ipums@umn.edu with more specific questions. And if you have any ideas for how ipumsr could make this easier, let us know by emailing user support or creating an issue on the ipumsr GitHub repository.

When dealing with IPUMS shapefiles do we need to define coordinates in our analysis, e.g., in hotspot analysis?

The webinar, on slides 69-73, provides an example of an analysis in which there is no need to define coordinates, if we are understanding your question correctly. However, we may be able to better answer your question if you email the IPUMS user support team at ipums@umn.edu with more details.

Can the ipumsr package be used to conduct a hotspot analysis that uses kriging interpolation with IPUMS DHS data?

The ipumsr package does not include special support for hotspot analysis or kriging interpolation, but it might still be useful in preparing your IPUMS DHS data for such an analysis. Please reach out to IPUMS user support at ipums@umn.edu if you have more specific questions on this topic.

How often are the data updated? Is there a way to see how often it's updated?

All IPUMS data collections keep a revision history that describes changes to the data over time. The revision history can be accessed with the "Revision History" link on the left sidebar of the data collection homepage, under the "Documentation" heading. For example, see the IPUMS USA revision history.

What packages do I need to conduct panel analyses?

There are many R packages to help conduct panel analyses, but the plm package seems like a good starting point. If you have more questions, contact the IPUMS user support team at ipums@umn.edu.