NATIONAL
AGRICULTURAL
STATISTICS
SERVICE

U.S. Department of Agriculture
1400 Independence Avenue, S.W.
Washington, D.C. 20250-2000

Research Proposal Guidelines for Projects Requesting Access to National Agricultural Statistics Service Data

Persons wishing to conduct research on restricted National Agricultural Statistics Service (NASS) data in a virtual data enclave must submit a research proposal to NASS through the Standard Application Process (SAP). The purpose of this document is to provide guidance for the various sections of the SAP. Generally, this document applies to datasets that list NASS as a Source in the SAP metadata catalog. NASS partners with the Economic Research Service (ERS) for certain data assets. Proposals using data that lists both NASS and ERS as Sources in the SAP metadata catalog will be reviewed and approved by both agencies during the review process.

NASS and designated agents are required to protect the confidentiality of data collected under the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2018, Title III of Pub. L. No. 115-435, codified in 44 U.S.C. Ch. 35 and other applicable Federal laws. All research must have a statistical purpose and may not be used for regulatory, enforcement, or investigative purposes.

The Five Safes framework breaks down the decisions surrounding data access and use into five related but separate dimensions: safe projects, safe people, safe data, safe settings and safe outputs.

More information about accessing NASS restricted microdata is available on the NASS website.

SAP Proposal Submission

Researchers who wish to conduct research using NASS data are required to read the NASS Data Lab Handbook.

One member of the research team will submit the proposal online using the SAP, and the submitted proposal will be reviewed by NASS. The SAP will ask for information about the research team and the proposed project. Question verbiage and layout in the SAP may not be exactly as shown below. If researchers are uncertain about how to respond to an item, refer to this guide or contact the NASS Data Lab and Data Access Group at SM.NASS.Data.Lab@usda.gov.

Data

The first step to beginning a project proposal using NASS data is to select all required datasets by clicking the "Request Access" button in the SAP metadata catalog. Researchers are encouraged to review all information available for each dataset in the metadata catalog on the SAP website and reference it when selecting the data and years available for each dataset. To check whether datasets in the SAP metadata catalog can be accessed together, click on "Data Access" and verify that the Access Modality indicates USDA Virtual Data Enclave. After selecting all datasets needed for the project, click "Start Application."

Data collected by NASS that are not listed are unavailable or not prepared for researcher use. Contact the NASS Data Lab and Data Access Group for additional information.

Research Team

The proposal must include all members of the research team, including collaborators who do not plan to access NASS data directly via an enclave account, but will participate in planning, oversight, or dialogue about the project prior to disclosure review. List the name of the principal investigator (PI) and all other researchers associated with the project. NASS does not allow students to be designated as the PI. Students seeking data access must designate their primary advisor as the PI. Researchers can be citizens of the United States, permanent residents, or foreign nationals. Foreign nationals must complete an additional background check prior to data access. All researchers, including U.S. citizens, must access NASS data from within the United States.

For each member of the research team, the application form requires name (full legal name), affiliation (or place of employment), title (faculty, grad student, research scientist, etc.), email address, phone number, citizenship (U.S. citizen or not), whether the researcher currently has active Special Sworn Status, and if the researcher will access the confidential data through an enclave account (Data Access=Yes) or only participate through collaboration (Data Access=No). For student team members, a graduation date must be entered.

The email address must be an institutional email address (e.g., a work or school email address), not a personal email address.

If new collaborators or researchers are needed to join or leave the project after the proposal has

been approved, the NASS Data Lab and Data Access Group will assist with adding or removing them from the project via addendum.

Research Description

This portion of the project application will focus on the project details. Each field or section must be written so that it can be understood by a competent statistician who is not necessarily a specialist in or on the topic within the field of study. Keep in mind that the audience for a NASS research project proposal is not the same audience researchers typically address when writing research proposals or journal articles. NASS has its own set of expectations and requirements, and therefore researchers often find that they must make significant revisions to preexisting proposals written for other purposes (e.g., grant applications).

Project Title

Provide a descriptive title for the project.

Project Duration

NASS encourages researchers to carefully assess the time period for which they request data access and ensure the intended research can be completed in this timeframe. Requests for extensions beyond the specified end date undergo careful scrutiny, must be justified, and generally are granted only for circumstances beyond the control of the researchers (e.g., unexpected illness). A typical project duration is 12 to 36 months. Note that researchers may discover that it takes longer than anticipated to become familiar with the data and computing environment.

Funding

List sources of funding for data access fees, as well as funding for the research project overall. NASS does not charge researchers for use of the data but does require the researcher to pay for services of the data enclave provider. The enclave provides a secure environment for researchers to access data while administering confidentiality protections. General fee information for enclave access can be found on the NASS website.

Timeline

Upload a table with a timeline, including a column for each year and a row for each task. A generic timeline template is available in the SAP application for download. The timeline must include activities such as cleaning data, merging data, and conducting specific analyses, as well as an estimated timeframe for outputs, such as disclosure review, publications, presentations, or any public use products. Refer to the timeline below as an example.

Task	**year**		
	1	2	3
Add approved outside data to enclave workspace	X		
Assess goodness of fit, quality of estimates for analysis	X		

Conduct data analysis for models of industry flow	X		
Draft paper of research findings		X	
Model entrance of movement of commodities through US		X	
Request export review sent to NASS for release of research results		X	
Submit journal report to conference		X	
Complete paper and submit to journal.			X
Graduate			X

Research Question(s)

Provide an overall description of the research questions. This section must describe the purpose of the project including any key concepts that will inform the research methods or models and a general description of the sample(s).

Demonstrated Need

Discuss why the project needs non-public data. Explain in detail why the project's research questions can only be addressed using the requested restricted-use microdata. Be specific.

Study Population

Briefly describe the study population or universe and how it relates to the research question.

Project Abstract

This section must be 150-250 words in length, reference major datasets and key methods; discuss expected findings. Researchers may include one or two sentences about why the research is important to their field(s) of study. It must capture the essence of the project proposal, similar to a scientific journal abstract.

Time, Geographic, and Other Units Requested

This field is required for all NASS applications. Refer to the Provisioned By fields in the SAP Metadata catalog for required items. Generally, for each requested dataset, list the year(s) of data needed. The years of requested data need to be clearly justified along with linkage plans, particularly for proposals requesting a time series.

Work Location(s)

The SAP lists work locations used by all agencies, but only USDA Virtual Data Enclave can be selected for projects that access NASS data.

Data linkages

Upload a document summarizing how all the data fit together, how they will be linked, and their expected overlap, including expected sample sizes (if applicable). State the datasets to be linked, the unit at which the linkages will occur (e.g., operation, geographic, etc.), the purpose of the linkage (e.g., geographic context), and information as to how the linkages are to be performed.

It is important to describe how and at what unit of observation or aggregation datasets will be linked. Linking survey data to other survey data may lead to small sample sizes (and therefore potential disclosure issues) if the survey samples do not significantly overlap. Also, linking a single survey over time can lead to similar concerns. When linking confidential data to externally provided contextual variables, detail the level of geography (e.g., county, state, etc.) at which this link will be performed.

User-Provided Data

If the project team plans to bring in non-NASS data to the project, enter information about each dataset. The list must be comprehensive. Enter the dataset name, a short description of the data, and the approximate size of the file. The application will ask for the data source and whether the data is publicly available or proprietary. If available, include the link where the data can be downloaded.

Be prepared to provide this data to NASS prior to workspace access. Future requests for user-provided data not listed in the proposal will require additional justification and approval and may result in an ingestion fee.

Software Requirements

The virtual data enclave has statistical software ready to use in the computing environment, such as SAS, R, Python, etc. Available software applications can be found in the SAP metadata catalog; click on Data Access, then on USDA Virtual Data Enclave under Access Modality. Consult with the NASS Data Lab and Data Access Group about other software needs. If the project team requires statistical software other than that which is currently available in the USDA Virtual Data Enclave, enter details in this section. Accommodations for unique software requirements may be available; provide justification, priority, and the inability of available packages to do the work. Fees will apply.

Note that the virtual data enclave environment does not have internet access.

Variables Requested

NASS requires a variable selection spreadsheet to be submitted for applications requesting access to Census of Agriculture data. It is not required for any other datasets. The spreadsheet of available variables can be found on the **NASS website** or downloaded directly through **this link**. Follow the directions on the Information Sheet tab to select project variables. Variables requested need to be clearly justified in the project proposal.

Since the SAP application does not currently allow spreadsheets to be uploaded, this spreadsheet must be emailed to **SM.NASS.Data.Lab@usda.gov** after you submit your application. Include the application number in the subject of the email. Delay in receiving the spreadsheet will delay the review of the application.

Methodology

This section describes the statistical equations that will be estimated, what key variables are needed and how they will be measured, how the data will be used, and how all datasets fit together. It is important in this section to describe the methods used to answer the research questions.

All requested datasets need to be discussed in the methodology section. The purpose for each requested dataset and its role in the analysis must be clear. Data crafting is an essential element of research design. The proposal needs to be as detailed as possible in describing how measures will be constructed and how the data will fit together to construct the focal sample(s). Generally, this is accomplished by discussing how variables are constructed and from which datasets they are retrieved.

It is important for researchers to convey to the application reviewers that they have read the publicly available documentation about the datasets they are requesting and have thought through implications of the data in relation to their research approach. Sources of information about datasets that are linked to the SAP metadata catalog include the website link for general survey information and the supplemental documentation, including survey questionnaires and data dictionaries, which can be downloaded from the catalog.

The unit(s) of observation for the analysis and the groups of units for which researchers will carry out analysis are both significant. For each, be specific. Is the unit of observation a farm operation? Alternatively, will the unit of observation be at a more aggregated level (e.g., county)? Will the analysis use some combination of units? Additionally, for what groups of units (what "levels") will researchers specify their models? Will researchers estimate models at national or sub-national levels? If so, at what levels? In addition, will researchers run separate models by demographics, crops, etc.? If the project uses user-provided data, how big are the samples that will be analyzed? Information that describes the units of analysis and the groups of units/levels of samples the project will be using are necessary for NASS reviewers to assess both the feasibility of the project and the risk of disclosing confidential information.

Write out the equations that will be estimated. Researchers do not need to give the exact functional form they plan to use, but they must write out, in as much detail as possible, the general equation and relevant variables and discuss how the data will fit into the equation. Researchers do not need to list every specification, nor every variable they will be using from each dataset, but they must discuss how the left- and right-hand sides of each equation will be

measured (i.e., if they are dichotomous, categorical, or continuous, etc.). Specify the dataset(s) from which these measures will come and how they will be constructed.

List of References

Provide a brief discussion of the relevant literature. Provide a targeted discussion of the literature that focuses on the key datasets as well as the key concepts to be examined in the analysis. This will help demonstrate scientific merit of the research question(s) as well as demonstrate a deeper knowledge of the requested data. Include a list of publications referenced in the proposal.

Project Products

Describe the deliverables that the research team anticipates producing, such as journal articles, conference presentations, technical reports, PhD dissertations, government reports, and other expected products. Include the names of journals targeted for publication.

Requested Output

In this section, researchers must describe the output they expect to ask to release from the secure data enclave. This description is important in assessing both the substance of the proposal and the risk of disclosing confidential information. Output from the enclave must focus on model-based results (regression coefficients, standard errors, and the like) and be consistent with the approved project objectives. Data mining is not permitted. Summary statistics (variable means, etc.) are allowed, but only to the extent that they support model-based output. For example, the kinds of output expected to be released are regression estimates and tables similar to those found in an article in a peer-reviewed academic journal. Researchers who desire disclosure of large volumes of tabular output must request a special tabulation from NASS rather than requesting access to microdata. In order to protect sensitive and confidential data, any products created from internal data are confidential until they have undergone disclosure review and have been approved for release.

NASS looks for several things in evaluating proposals for disclosure risk. The project application must clearly state the proposed methods and output. Reviewers understand that, because the research team is conducting research, it is unlikely that they know all the details in advance. Nevertheless, reviewers need enough detail to assess whether the proposed project can succeed without posing undue risk. In particular:

- Reviewers need to determine whether there may be possible "thin cells" in the output; for this purpose, a cell is the group of observations (e.g., farm operations) underlying any estimate researchers may release. To evaluate this, reviewers need clear and accurate information on the types of output to be requested (e.g., model tabulations, graphs), the units of analysis underlying the output (e.g., all farm operations, or groups of these), and the groups (levels) for which researchers will request output (e.g., crops, gender, geography, and possible crosses of these). For example, detailed demographics-by-geography cells become thin very rapidly.
- Reviewers need to know whether variables included in models are discrete or continuous; both are allowable, but discrete variables (especially dummy (0, 1) variables) define "cells" that reviewers must carefully consider. Note that including indicator variables in the analysis creates additional “cuts” at the sample (i.e., cross tabulations), but only if the estimated coefficients associated with the indicator variables are reported. Because of this potential issue, the review team encourages including statements in the proposal, as appropriate, for example: “Specifications will include detailed farm type controls. However, we will not release these fixed effect coefficients; we will only note in our results tables that we included them.”
- Reviewers need to know whether researchers will request tabulations, graphs, or maps; these can have particular difficulty satisfying disclosure standards.
- Reviewers need to know whether researchers plan to produce sub-state estimates.
- Reviewers will also look for types of output that are relatively unfamiliar, such as from new statistical techniques. They must try to assess their disclosure risk and welcome any insights researchers may have on this.

Disclosure risk is a complex topic, and researchers will need to discuss their proposed output with their research team as they develop their proposal.

Agency Benefits

List the proposed benefits to NASS, explaining how the proposed work will achieve those benefits.

Documentation

NASS requires one document beyond the application itself. The MOU is an agreement between NASS, the institution/agency, and the researchers. The MOU must be signed by the Principal Investigator/Project Lead who is an employee of the institution and by a designated Senior Official of the Organization who is authorized to enter into contractual agreements. Signatures must be digital or wet. No electronic or typed signatures are accepted.