Recovery of Restricted Decennial Census Microdata for Puerto Rico and Island Areas, 1970, 1980, 1990

Gina Erickson, US Census Bureau

Todd Gardner, US Census Bureau

John Sullivan, US Census Bureau

Abstract

Modern data production often must contend with challenges related to measuring the characteristics of varied groups and disseminating complex data. These challenges are present in the data collection and dissemination for Puerto Rico and the Island Areas. The U.S. Constitution mandates that the Census Bureau conduct a census of the population every ten years. This includes enumerating the population of US territories (commonly referred to as Island Areas by the Census Bureau) and Puerto Rico. Puerto Rico has been enumerated in each Decennial Census since 1910 with U.S. Territories and Commonwealths following in 1920 (Guam and American Samoa), 1940 (U.S. Virgin Islands) and 1970 (Commonwealth of the Northern Mariana Islands). However, research-ready restricted use decennial microdata have not previously been produced for Puerto Rico and the Island Areas for 1970, 1980 and 1990. We discuss the process to produce these data and highlight differences between decennial data for Puerto Rico and the mainland U.S.

Introduction

The complications of modernity abound in demographic data collection, dissemination and retention. Modern empiricism often relies on measurement categories based on the characteristics of dominant groups, produces statistics intended to guide the administration of state policies and disseminates data in ways that may prioritize some places over others. Each of these topics are present in the data collection and dissemination for Puerto Rico and the Island Areas. In this paper, we discuss the recovery of restricted decennial microdata for the population of Puerto Rico (though the project will eventually include details on data recovery for the Island Areas, our efforts thus far – and this paper – focus on Puerto Rico).

The U.S. Constitution mandates that the Census Bureau conduct a census of the population every ten years. This includes enumerating the population of US territories (commonly referred to as Island Areas or Outlying Areas by the Census Bureau) and Puerto Rico. Puerto Rico has been enumerated in each Decennial Census since 1910 with U.S. Territories and Commonwealths following in 1920 (Guam and American Samoa), 1940 (U.S. Virgin Islands) and 1970 (Commonwealth of the Northern Mariana Islands). However, research-ready restricted use decennial microdata have not previously been produced for Puerto Rico and the Island Areas for 1970, 1980 and 1990.

The goal of the Decennial Census Digitization and Linkage Project (DCDL) is to fill a gap in digitized data and the longitudinal data infrastructure by creating and linking research-ready data from the censuses and surveys conducted by the US Census Bureau. This paper describes our efforts, decades after data collection, to locate the unlabeled data, harmonize it with existing restricted decennial microdata and make it available for research. Difficulty in locating appropriate documentation and differences in enumeration procedures and instruments between Puerto Rico and the mainland created challenges for determining record layouts and harmonizing data fields. The research-ready data from both the sample edited file "long-form" and high-density file "short-form" adds 3-4 million person records in each decennial year to the hundreds of millions of records that will soon be included in the Census Bureau's longitudinal linkage infrastructure through the DCDL project. The work described here and the larger DCDL project, will ensure that linked person records covering the population of Puerto Rico are available for research use.

Background

Different places require different enumeration strategies. In Puerto Rico, variations in census methodology, geography and language use are present (p. 215, 1990 Procedural History). Among the differences are terms used to describe geographic delineations (barrio), differences in mailing addresses, relevant questions about language proficiency and migration history. Differences in enumeration instruments not only account for language differences between Puerto Rico and the mainland, but also include different questionnaire items and response categories to address characteristics unique to the population of Puerto Rico.

Restricted microdata files for research exist for stateside censuses for these years but not for Puerto Rico and US Island Areas. Though the data were collected, processed and used to create PUMS for Puerto Rico, they were not made into research-ready files nor included in the available decennial census data holdings made available to external researchers.

Data and Methodology

We located restricted microdata files for Puerto Rico and the Island Areas on a Census Bureau server and transferred the files to the Integrated Research Environment (IRE). The data files consisted of record-per-line character strings and had no column delimiters, variable labels, record layouts and no associated documentation. The files we located are listed below. Our recovery is ongoing and is currently focused on files from Puerto Rico.

- 1970
    - Puerto Rico High Density File (Short Form/Hundred Percent File)
    - Puerto Rico Sample Edited Detail File (Long Form/Sample File)
- 1980
    - Puerto Rico High Density File (Short Form/Hundred Percent File)
    - Puerto Rico Sample Edited Detail File (Long Form/Sample File)
    - American Samoa Census
    - Guam Census
    - US Virgin Islands Census
    - Commonwealth of Northern Mariana Islands (CNMI) Census

    - Trust Territory of the Pacific Islands (TTPI) Census
- 1990
    - Puerto Rico High Density File (Short Form/Hundred Percent File)
    - Puerto Rico Sample Edited Detail File (Long Form/Sample File)
    - American Samoa Census
    - Guam Census
    - US Virgin Islands Census
    - Commonwealth of Northern Mariana Islands (CNMI) Census
    - Palau Census

We aimed to produce research-ready datasets that internal and external researchers can use in the Federal Statistical Research Data Center (FSRDC) network. To do this, we needed to understand how the raw data were split into different record types and different data fields. With no information to go on from the directories in which we found the data files, we collected documentation from a variety of sources.

We drew on five documentation sources, some available in all decennial years and others only available in some (e.g. documentation only for long-form PR PUMS);

1. Record layouts from the currently available restricted microdata for the mainland US.
2. Record layouts from publicly available PUMS data for Puerto Rico.
3. Publicly available decennial questionnaire forms for Puerto Rico.
4. Scanned completed questionnaires from Puerto Rico (via the DCDL project).
5. Internal Census communications about the contents of the data files.

Each source helps us understand the content of the Puerto Rico restricted microdata files, however, no source of documentation directly addresses the exact files we located. Documentation sources often vary or provide conflicting information about the order of fields, the number and labels of response categories and the total length of a data record. The complex web of overlapping and conflicting documentation made it necessary for us to use multiple documentation files to “triangulate” what we believe to be an accurate record layout.

Fortunately, the data files were organized into single-record-per-line format, that is, each line in a data file contains a string of numerals and characters that correspond to a single data record. Furthermore, the record types in the Puerto Rico files follow exactly the scheme used in the mainland US data (1 = Geographic Record, 2 = Group Quarters Household Record, 3 = Person Record, 4 = non-GQ Household Record, 5 = non-GQ Person Record). This record type structure is nested such that person information is nested within household records which are nested within geographic records. Linking keys between these levels allow one to situate an individual within a specific household in a specific census geography with household characteristics described in the household file and with geographic characteristics described in the geography record.

We split a given data file into separate record types and then manually parsed and verified proposed record layouts. This involved referencing multiple documentation sources at once. We would work our way through a data record of a given type, comparing the order and range of values in the data to the order and response categories in the different documentation sources. Because we did not have a single source that corresponded directly to the data we often began with the record layout for the restricted microdata from the mainland US. We lined fields up with documentation from the PUMS and the questionnaire forms as needed. We began this process from an anchor point (a field in the data that we had very strong confidence corresponded to a specific field in the documentation) and then worked

forward and backward from there. As we identified a field, we tabulated its value across all data records to ensure it was within the range of response categories denoted in the documentation and made logical sense given the values of other fields that we had identified in the record and given the values of other record types in which a record was nested (e.g. for example, if 5 person records follow a household record, we would make sure that the household field that we thought was the total number of persons in a household was also 5. The process was akin to completing a multidimensional crossword puzzle.

Below is an excerpt of a set of imaginary records describing a three-person household. The first line is the household record, the following three correspond to a household head, a spouse and a child. The brackets are added here to denote the different fields of a record. By matching the data to the proposed record layout, we can verify that we have correctly parsed the data. For example, the household record begins with a record type value of 4 and has a value of 3 for the number of persons in the housing unit. Likewise, 3 records with record type 5 follow this record (corresponding to the three members of the household.

Example Household and Person Records

[4][72][00435][003][07][001]

[5][72][0001][00][1][727][035]

[5][72]0002][01][2][727][034]

[5][72][0003][02][2][727][009]

Example Corresponding Record Layout

Household

[Record Type] [State Code] [Household ID] [Number of Persons in Unit] [Home Value] [Age of Head]

PER

[Record Type] [State Code] [Person Number in HH] [Relationship] [Sex] [Race] [Age]

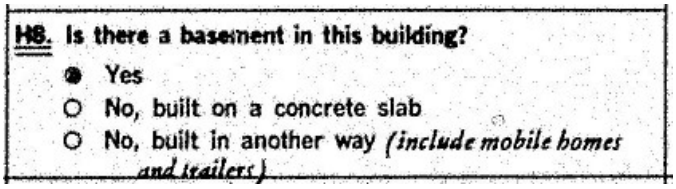
In none of the three decennials we examined, did the data align exactly with the order and response categories of any single documentation sources. For example, a record type may appear to follow the record layout in the mainland US up to a point, but then diverge and align more closely with the PUMS file for Puerto Rico (e.g. when a questionnaire item that is only asked of Puerto Rico – or not asked in Puerto Rico – appears). Likewise, tabulated values may align with PUMS for some fields that have slightly different response categories in Puerto Rico than in the mainland US, but then the records show response categories that do not perfectly align with any documentation source or appear to be auxiliary fields that we cannot locate in any documentation source. In this piecemeal fashion, we reconstructed and verified, as best we could, record layouts. Further, when possible, we used the exact same variable names and labels as are used in the restricted decennial microdata for the mainland.

The figures below show examples of differences in questionnaire items asked of respondents in Puerto Rico and the mainland US. These questionnaire differences made parsing record layouts with restricted documentation from the mainland and PUMS documentation from Puerto Rico challenging. The figure below shows a selection of housing items from the 1970 Decennial Census for the mainland (Item H8 and

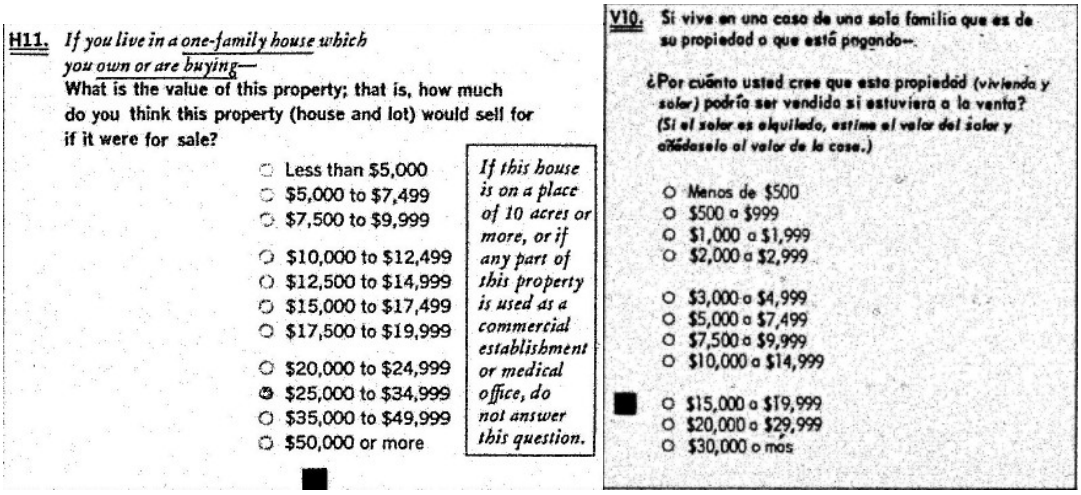
H11) and for Puerto Rico (Item V10). Item H8 asks respondents about the presence of a basement in the housing structure. This item is completely absent from the Puerto Rico questionnaire. Absence of an item "throws a wrench" in our ability to sequentially move from one field to the next when building a record layout, matching response values and length to other sources of documentation. The excerpts below also show another type of "wrench" that we encountered during this process. Item H11 shows the home value questionnaire item for the mainland US and item V10 shows the same item for Puerto Rico. The number of response categories is the same, but their labels are different.

1970 Comparison of Select Household Questionnaire Items

H8. Basement


H8. Is there a basement in this building?
- ● Yes
- ○ No, built on a concrete slab
- ○ No, built in another way *(include mobile homes and trailers)*

H11 & V10. Home Value


H11. *If you live in a one-family house which you own or are buying—*
What is the value of this property; that is, how much do you think this property (house and lot) would sell for if it were for sale?

- ○ Less than $5,000
- ○ $5,000 to $7,499
- ○ $7,500 to $9,999
- ○ $10,000 to $12,499
- ○ $12,500 to $14,999
- ○ $15,000 to $17,499
- ○ $17,500 to $19,999
- ○ $20,000 to $24,999
- ● $25,000 to $34,999
- ○ $35,000 to $49,999
- ○ $50,000 or more

If this house is on a place of 10 acres or more, or if any part of this property is used as a commercial establishment or medical office, do not answer this question.

V10. Si vive en una casa de una sola familia que es de su propiedad o que está pagando—

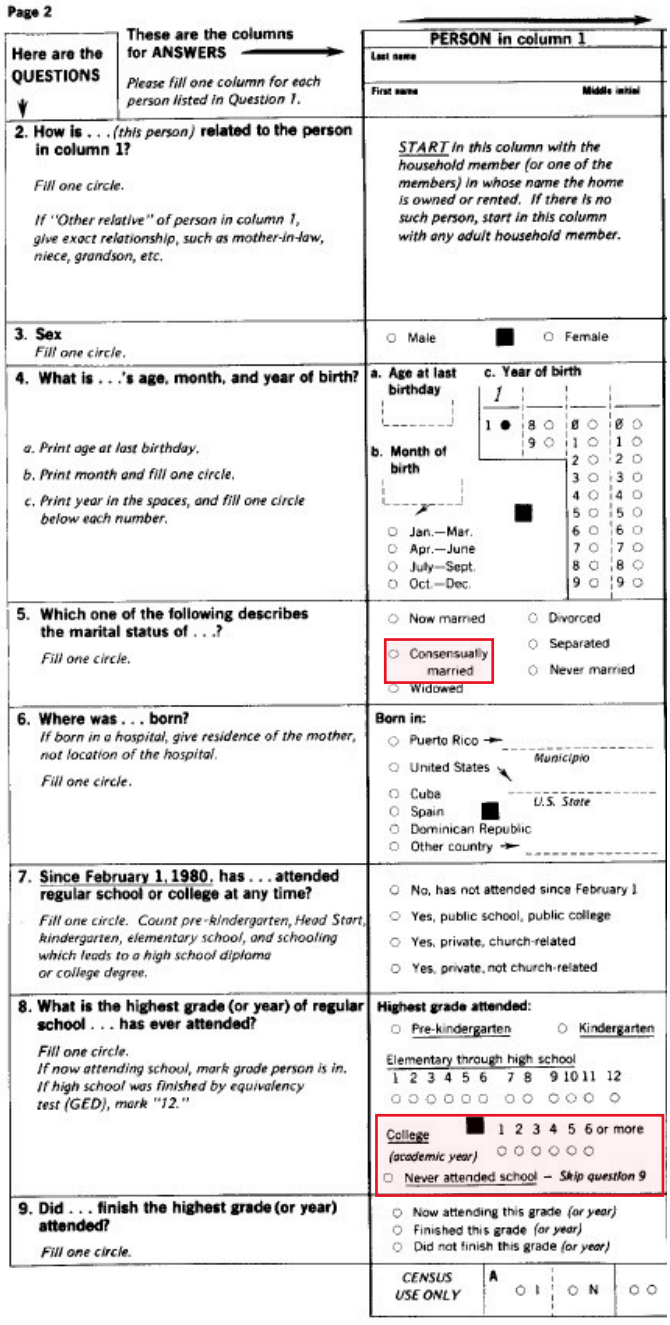
¿Por cuánto usted cree que esta propiedad *(vivienda y solar)* podría ser vendida si estuviera a la venta? *(Si el solar es alquilado, estime el valor del solar y añádaselo al valor de la casa.)*

- ○ Menos de $500
- ○ $500 a $999
- ○ $1,000 a $1,999
- ○ $2,000 a $2,999
- ○ $3,000 a $4,999
- ○ $5,000 a $7,499
- ○ $7,500 a $9,999
- ○ $10,000 a $14,999
- ○ $15,000 a $19,999
- ○ $20,000 a $29,999
- ○ $30,000 o más

The figures below show the "Person 1" section of the 1980 Decennial Census for the Puerto Rico and the mainland. The red boxes highlight two differences between the questionnaire forms. First, the marital status field on the Puerto Rico form includes one response category ("consensually married" equivalent to "common law marriage" or a union recognized publicly, but not recorded in a legal contract or document) that is not present in the mainland form. Second, the education fields differ in the number of years of college they include.

1980 Puerto Rico Person Questionnaire Items

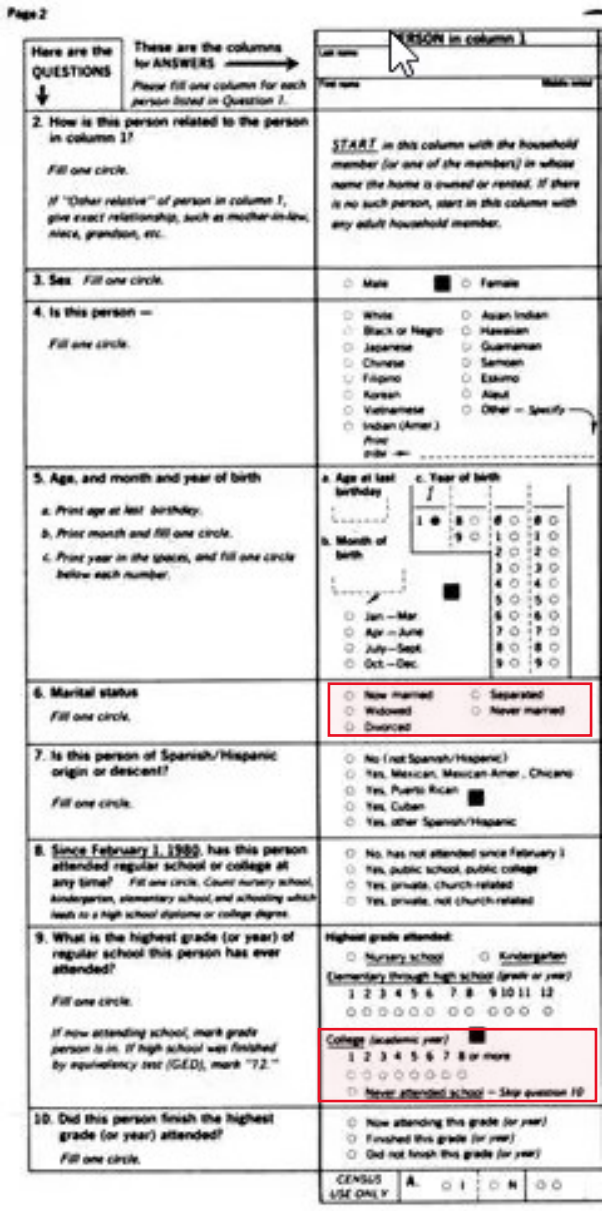
Puerto Rico Questionnaire Form (English translation)



Page 2

Here are the QUESTIONS	These are the columns for ANSWERS *Please fill one column for each person listed in Question 1.*	PERSON in column 1 Last name First name — Middle initial
2. How is . . . *(this person)* related to the person in column 1? *Fill one circle.* *If "Other relative" of person in column 1, give exact relationship, such as mother-in-law, niece, grandson, etc.*		*START in this column with the household member (or one of the members) in whose name the home is owned or rented. If there is no such person, start in this column with any adult household member.*
3. Sex *Fill one circle.*		○ Male ■ ○ Female
4. What is . . .'s age, month, and year of birth? *a. Print age at last birthday.* *b. Print month and fill one circle.* *c. Print year in the spaces, and fill one circle below each number.*		a. Age at last birthday b. Month of birth ○ Jan.—Mar. ○ Apr.—June ○ July—Sept. ○ Oct.—Dec. c. Year of birth 1 ● 8 ○ 9 ○ ; Ø ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ ; Ø ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○
5. Which one of the following describes the marital status of . . .? *Fill one circle.*		○ Now married ○ Divorced ○ Consensually married ○ Separated ○ Never married ○ Widowed
6. Where was . . . born? *If born in a hospital, give residence of the mother, not location of the hospital.* *Fill one circle.*		**Born in:** ○ Puerto Rico → *Municipio* ○ United States → *U.S. State* ○ Cuba ○ Spain ○ Dominican Republic ○ Other country →
7. Since February 1, 1980, has . . . attended regular school or college at any time? *Fill one circle. Count pre-kindergarten, Head Start, kindergarten, elementary school, and schooling which leads to a high school diploma or college degree.*		○ No, has not attended since February 1 ○ Yes, public school, public college ○ Yes, private, church-related ○ Yes, private, not church-related
8. What is the highest grade (or year) of regular school . . . has ever attended? *Fill one circle.* *If now attending school, mark grade person is in.* *If high school was finished by equivalency test (GED), mark "12."*		**Highest grade attended:** ○ Pre-kindergarten ○ Kindergarten Elementary through high school 1 2 3 4 5 6 7 8 9 10 11 12 ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ College ■ 1 2 3 4 5 6 or more *(academic year)* ○ ○ ○ ○ ○ ○ ○ Never attended school — *Skip question 9*
9. Did . . . finish the highest grade (or year) attended? *Fill one circle.*		○ Now attending this grade *(or year)* ○ Finished this grade *(or year)* ○ Did not finish this grade *(or year)*
		CENSUS USE ONLY A ○ I ○ N ○ ○

Mainland US Questionnaire Form



1990 Individual Responses from Puerto Rico

Ancillary to our attempts described above to make research-ready restricted microdata files for Puerto Rico, we also located individual responses to the 1990 Decennial from Puerto Rico. These responses will be scanned and processed as part of the DCDL linkage project.

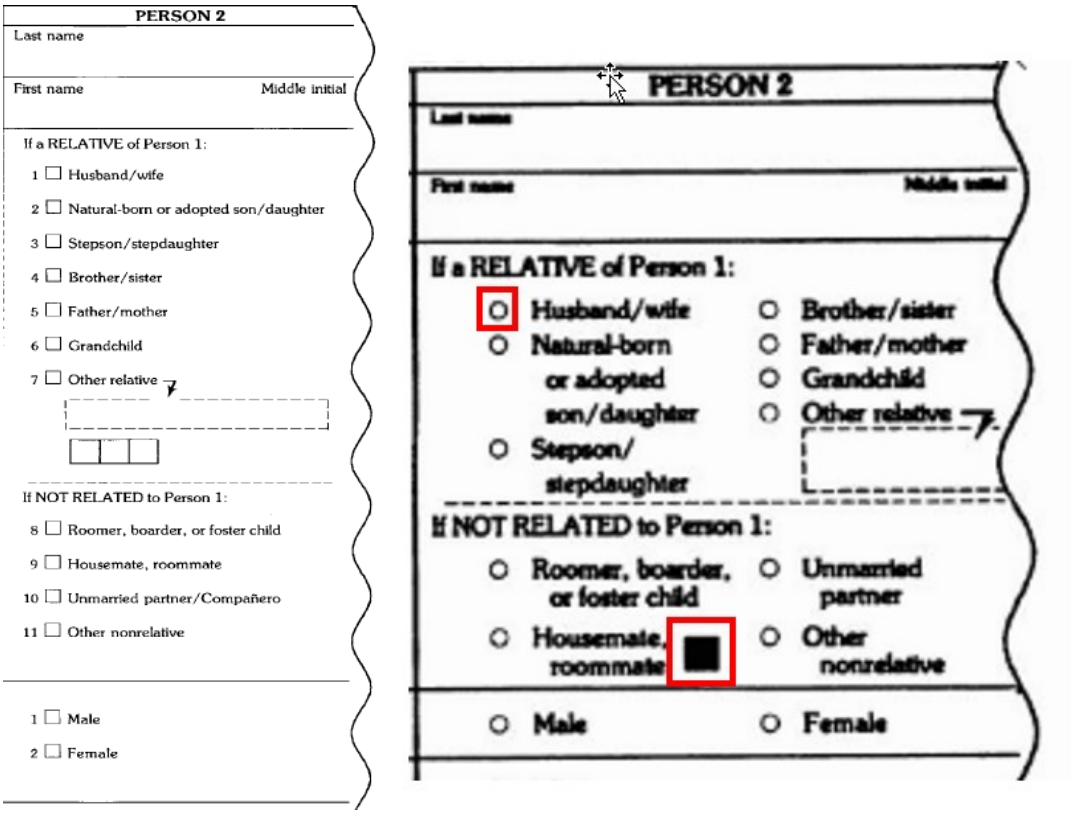
Census Bureau enumeration operations in Puerto Rico frequently differed from those in the mainland US. Responses to the 1990 Decennial Census in Puerto Rico were enumerated with an Advanced Census

Report (ACR) and other enumeration instruments, with responses recorded by enumerators were on forms unique to Puerto Rico.

The 1990 procedural history document states that "Questionnaires for Puerto Rico were keyable, but unlike those of the mainland, were not FOSDIC readable" (1990 Procedural History). The figure below compares the “Person 2” column from the form used by enumerators in Puerto Rico to the questionnaire mainland respondents completed and returned by mail. Red boxes highlight that the mainland form contains circles designed to be filled in for reading by the FOSDIC and black registration squares for FOSDIC alignment. In contrast, the Puerto Rico form lacks FOSDIC registration marks and uses square check boxes instead of circles.

Comparison of Puerto Rico and Mainland Response Instrument, 1990

Puerto Rico *Mainland US*



As such, the collection of 1990 decennial microfilm reels that hold all completed questionnaires from the mainland US do not contain images of any census responses from Puerto Rico. Further reading of the 1990 Procedural History reveals that completed enumeration forms from Puerto Rico were sent to the Jacksonville, FL processing office where they were hand-keyed by clerks. Processing at the Jacksonville office suggests that Puerto Rico responses would have been stored on reel with IDs following the convention 31XXXXXX (for Short Form) or 36XXXXXX (for Long Form) 3 is the "prefix" for Jacksonville reels (indeed, these reels are not present in the collection of microfilm reels).

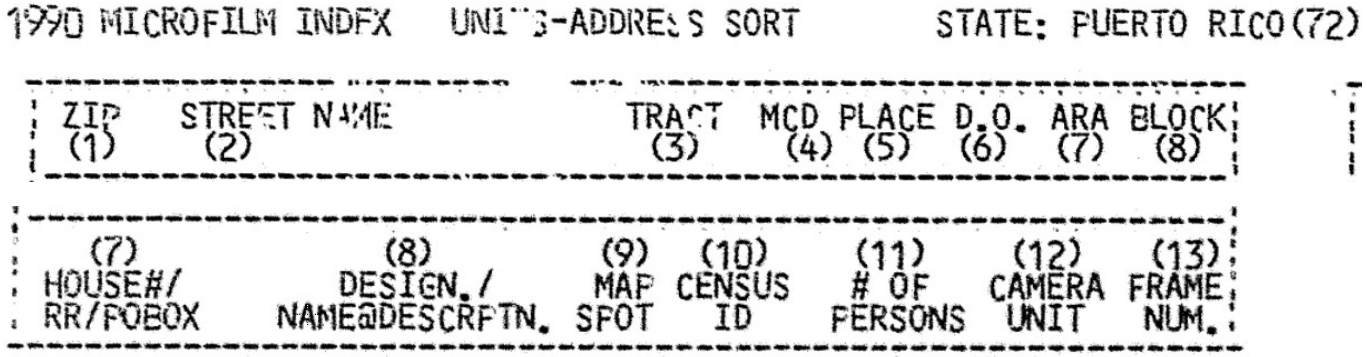
The keyed records containing names, person and household level information were printed to microfiche. After consulting the Age Search division of the Census Bureau, we learned that the Puerto Rico microfiche are stored at the Census Bureau’s National Processing Center (NPC). The records are

organized into three “stacks” (A, B and C) with each stack containing separate envelopes of microfiche for each Municipio (Puerto Rico county equivalent). Stacks A and B appear to be the 1990 Address Control File records. Stack C contains the person/household microfiche were scanned in their entirety in September 2025.

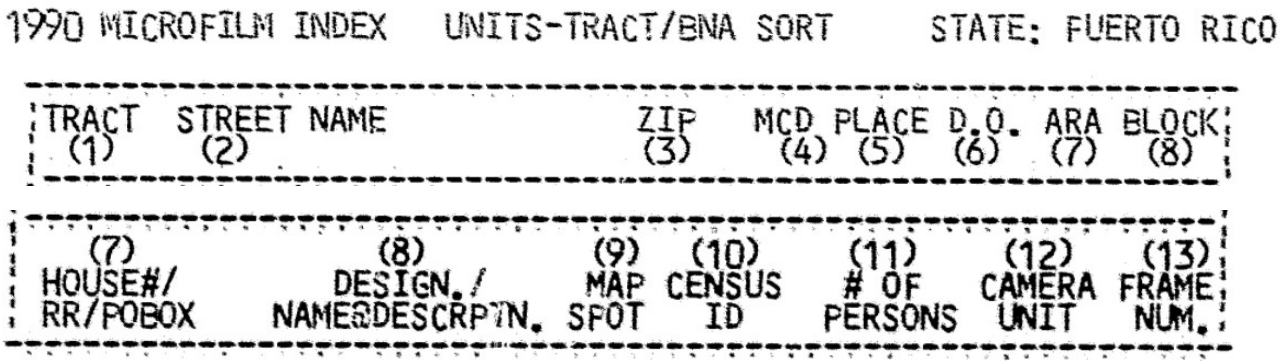
To verify that Stacks A and B contained the same ACF information (but had a different sort order), I found the same records in both the A and B scans and in the raw ACF microdata. The information from Stack A and B appears to be identical to that in the ACF. However, unit number preceded *hholder_name*, but for the first unit within a sort-level the unit number in the Stack A and B scans appears not to have been printed out on the microfiche (this information is complete in the versions of the ACF on IRE).

The last page of an envelope of microfiche (in any stack) contains an inventory of microfiche pages in that envelope listed in an order that does not always match the order in which they were imaged to the microfiche or scanned from the microfiche (this shouldn’t really matter). The bottom of the last page in a microfiche envelope contains a code that we think is the reel code as it matches the reel IDs from the ACF. The figures below show column contents for Stacks A and B.

Stack A: 1990 Microfilm Index Units-Address Sort.



Stack B: 1990 Microfilm Index Units-Tract/BNA (Block Numbering Area) Sort



Stack C: Housing Unit Data

The person-level microfiche consists of approximately 70,000 pages of Stack C records. PNUM is not necessarily assigned after the records are alphabetized (like in the 1990 group quarters microfiche records). The first page of a stack (or series of stacks) that contains column definitions says that PN is the

"sequence in which the person was reported on the questionnaire." DCDL project staff has developed programs to extract information from scans of very similar microfiche from the 1990 Group Quarters records. We are in the process of adapting these programs to produce individual level data (with names and addresses) from the 1990 Puerto Rico microfiche.

Results

We were able to create partial record layouts for the 1970, 1980 and 1990 HDF files and the 1990 SEDF for Puerto Rico. The record layouts include key demographic information, but we were not able to identify some fields (these are noted in internal documentation available with the data files). Each data file follows the dissemination conventions for the mainland US and is split into three separate files (Geography, Household and Person), with keys included to link these files to one another. The sample data files were particularly challenging for reconstruction and to date we have only produced a record layout for the 1990 decennial.

We verified the data and our record layouts by comparing record counts in the data to published tabulations from decennial summary files. We made comparison tabulations at multiple levels of geography to ensure we had correctly reconstructed the data. Currently, we have made available parsed data files, record layouts and documentation for the files listed below. These files can be requested for research use through the FSRDC network.

1990

- Puerto Rico High Density File
- Puerto Rico Sample Edited Detail File

1980

- Puerto Rico High Density File

1970

- Puerto Rico High Density File

Discussion and Conclusions

Though the Puerto Rico and Island Area restricted microdata recovery project is ongoing, newly available data are ready for researchers to use now and more cross-sectional and longitudinal data from Puerto Rico are coming soon.

Decisions about data collection, storage and priorities about data production and dissemination shape research opportunities decades after they have been made. Approaches to managing complex enumeration needs of Puerto Rico resulted in different collection procedures and instruments. This resulted in data files with content differences, which in turn made our efforts to curate the data challenging and, ultimately, imperfect. As the DCDL project incorporates these newly recovered records from Puerto Rico into the Census Bureau's linked data infrastructure, further attention should be paid to how these issues may propagate to create consequences that influence data quality and even the questions that researchers can ask of the data.