Melad, Kris Ann M.

Working Paper

Harmonizing Philippine census data across decades (1970-2020)

PIDS Discussion Paper Series, No. 2024-44

Provided in Cooperation with:
Philippine Institute for Development Studies (PIDS), Philippines

Suggested Citation: Melad, Kris Ann M. (2024) : Harmonizing Philippine census data across decades (1970-2020), PIDS Discussion Paper Series, No. 2024-44, Philippine Institute for Development Studies (PIDS), Quezon City, https://doi.org/10.62986/dp2024.44

This Version is available at:
https://hdl.handle.net/10419/311665

DECEMBER 2024

DISCUSSION PAPER SERIES NO. 2024-44

Harmonizing Philippine Census Data across Decades (1970–2020)

Kris Ann M. Melad

CONTACT US:

RESEARCH INFORMATION DEPARTMENT
Philippine Institute for Development Studies

18th Floor, Three Cyberpod Centris - North Tower
EDSA corner Quezon Avenue, Quezon City, Philippines

publications@pids.gov.ph
(+632) 8877-4000

https://www.pids.gov.ph

Harmonizing Philippine Census Data across Decades (1970–2020)

Kris Ann M. Melad

PHILIPPINE INSTITUTE FOR DEVELOPMENT STUDIES

December 2024

Abstract

This research harmonizes Philippine Census of Population and Housing (CPH) data from 1970 to 2020 to address data consistency challenges across five decades. The study systematically reconciles evolving variable definitions, classification systems, and measurement scales to create a unified longitudinal dataset. Key harmonization challenges include accommodating education system changes such as the K-12 reforms, tracking administrative boundary modifications through the years, managing expanding data scope across census years, and handling historical data preservation issues, particularly for the 1970 and 1980 censuses. The research involved creation of translation tables and crosswalks for major classification systems including the Philippine Standard Geographic Classification (PSGC), Philippine Standard Occupational Classification (PSOC), and Philippine Standard Industrial Classification (PSIC). Variable-specific harmonization protocols and guidelines for researchers using the harmonized data are also documented. The harmonization process standardized core demographic variables across all periods while preserving more detailed classifications where possible, though some variables necessarily lost granularity when harmonized to their lowest common denominator. Beyond producing a consistent dataset for longitudinal analysis, this study contributes to PIDS's agenda of strengthening statistical systems for evidence-based policymaking. The paper concludes with recommendations for improving future census data collection and harmonization practices to support effective policy development in the Philippines.

Keywords: census, data harmonization, variable standardization, data translation, Philippine Statistical System

Table of Contents

List of Tables

List of Figures

Harmonizing Philippine Census Data across Decades (1970–2020)

Kris Ann M. Melad[1]

1. Introduction

1.1. Background

The 2020–2025 research agenda of the Philippine Institute for Development Studies (PIDS) underscores the critical importance of accurate and consistent statistical estimates. As the government's primary think tank for socioeconomic policy, PIDS emphasizes the need for statistical systems to become more "relevant and responsive to the demand for evidence-based policy decisions." This emphasis reflects a growing recognition that effective policymaking relies on the ability to understand and analyze trends over time, drawing insights from reliable and comparable data.

However, despite the Philippines' regular data collection efforts, inconsistencies across different periods pose a challenge to achieving this goal. Discrepancies in variable definitions, classification schemes, and measurement scales hinder accurate comparisons of data points, increasing the risk of misinterpretation and compromising the reliability of longitudinal analyses.

Internationally, significant efforts have been made to address similar challenges in data harmonization. The Integrated Public Use Microdata Series (IPUMS), for example, harmonizes international datasets, allowing researchers to seamlessly compare censuses across time and countries. Other notable initiatives include the Survey Data Recycling (SDR) project, which maximizes the utility of existing survey data including documentation and metadata, and the Consortium of European Social Science Data Archives (CESSDA), which provides integrated data resources for researchers. In addition, the Harmonized Household Income and Expenditure Surveys (HHIES) focuses on income and expenditure surveys in the Arab region, while the World Bank's Microdata Library facilitates access to harmonized datasets from household and business surveys worldwide. These efforts highlight the value of harmonized datasets in facilitating meaningful trend analysis and cross-country comparisons.

The Philippine Census of Population and Housing (CPH), conducted every five years by the Philippine Statistics Authority (PSA), serves as a vital data source for policy research and planning. This comprehensive enumeration captures information on demographics such as age, sex, and education, as well as housing characteristics like construction materials and occupancy. Government agencies rely on this data to guide resource allocation, design targeted policies, and plan for local and national development. Beyond government, the private sector utilizes census data for market research and investment planning. The CPH is thus an indispensable resource for understanding the country's population, socio-economic conditions, and development trajectory.

Inspired by international harmonization initiatives and aligned with PIDS's call for strengthened data systems, this study aims to address the challenge of data inconsistencies in

[1] Acknowledgment is given to Ms. Jhanna Uy for her initial work on the data preparation and material compilation for this research.

the Philippines. By harmonizing census data from 1970 to 2020, the study seeks to enable robust longitudinal analyses and enhance the evidence base for policy research. Specifically, it aims to answer the question: How can changes in variable definitions, classification systems, and measurement scales across census periods be reconciled to enable valid comparisons and trend analysis of key indicators in the Philippine Census data?

The harmonization of Philippine census data from 1970 to 2020 presents several significant challenges that must be carefully addressed. First, variable definitions and classification systems have evolved substantially over the five decades, reflecting changes in both international standards and local data needs. For instance, educational categories have been modified to accommodate the K-12 curriculum reforms, while occupational classifications have been updated to reflect emerging industries and job types consistent with time. Second, geographic boundaries have undergone numerous changes, with the creation, splitting, and renaming of administrative regions, provinces, municipalities, and barangays that require careful reconciliation of location codes across different periods. Third, the scope and depth of data collection have expanded significantly, with earlier censuses like those in 1970 and 1980 having more limited variables compared to recent ones. Furthermore, the digitization and preservation of historical census records pose additional challenges, as some older datasets are incomplete or stored in outdated formats that require conversion and validation. These technical and methodological variations across census years necessitate a systematic approach to data harmonization that can bridge these differences while maintaining the integrity and comparability of the data.

To address these methodological and technical challenges, this study employs a systematic harmonization framework that carefully balances data preservation with standardization needs. This approach not only allows us to overcome the identified obstacles but also ensures that the resulting harmonized dataset maintains its utility for policy analysis and research purposes. Through careful documentation of the harmonization process and development of clear protocols for handling various data inconsistencies, the study provides both immediate analytical tools and a foundation for future data integration efforts.

The outputs of this study include a consistent dataset that supports longitudinal analyses and reveals patterns and trends over time. Additionally, the research documents the harmonization process and propose recommendations for future improvements in census data collection and harmonization practices. Overall, this research contributes to the broader agenda of enhancing statistical systems to support evidence-based decision-making, ultimately promoting more effective socioeconomic policies in the Philippines.

1.2. Objectives of the study

The objective of this research is to develop and implement a set of methodologies for harmonizing key variables in Philippine census dataset from 1970 to 2020 to allow longitudinal analysis and identification of trends in the data over time.

Specifically, it aims to:

- identify and document changes in Philippine census methodology across decades;
- identify and implement methodologies for harmonizing data points from different census years into a common format; and
- develop recommendations for future Philippine census data collection and harmonization.

2. History of Philippine Censuses

The earliest documented census in the Philippines occurred in 1591 during the Spanish colonization. This effort, based on accounting of taxpayers (*tributos*), estimated the population to be around 666,612 in Luzon and Visayas and additional 75,000 to 150,000 in Mindanao (Concepción 1977). The first official census of the Philippine population was conducted by the Spanish colonial government in 1877, pursuant to a royal decree (Gannett 1905; Concepción 1977). The recorded population then was 5.6 million. A slight decline in population was noted in the succeeding census of 1897 where the population estimate was pegged at 5.5 million.

Following the Spanish-American war, the American occupation conducted its first census in 1903, with subsequent censuses conducted in 1918 and 1939. These censuses employed enumerators who visited households to gather information on demographics, housing conditions, and economic activity using standardized questionnaires (United States Census Bureau, n.d.). Compared to the Spanish censuses, there has been significant improvement in data collection methodology while the census data also became an increasingly important basis in the governance of the country during this time.

The Bureau of the Census and Statistics (BCS) took over the responsibility of conducting the censuses in the 1940s after it was established through Commonwealth Act No. 591. The BCS conducted censuses every 10 years until 1970, when the Philippine Statistical System (PSS) was established, transitioning to a 5 year-schedule. The BCS was later renamed to the National Census and Statistics Office (NCSO) in 1974 and National Statistics Office (NSO) in 1987 (DBM 2009).

In more recent history, the PSA was established through the Philippine Statistical Act of 2013 by merging four major statistical agencies, including the NSO into a single entity. The PSA now oversees a wide range of statistical activities, from censuses on population, housing, agriculture, industries, to surveys on a wide array of statistics and monitoring indicators. It also administers civil registration functions and supervises aspects and standards of data collection, processing, analysis, and dissemination in the country (PSA).

Currently, the Census of Population and Housing (CPH) is done by household enumeration wherein trained enumerators visit households to collect data through face-to-face interviews using standardized questionnaires. The census primarily utilizes two types of questionnaires for data collection:

1. **Common Household Questionnaire (Form 2)**: This is the basic questionnaire administered to all households during the census. It gathers core information on

demographics like age, sex, and education, along with housing characteristics such as materials used and tenure status.

2. **Sample Household Questionnaire (Form 3)**: A smaller subset of households is selected to receive a more detailed questionnaire in addition to the common household questionnaire. This additional questionnaire asks further specific aspects like citizenship, language fluency, literacy, educational attainment, and employment details.

Additionally, there may be supplementary questionnaires depending on the specific census year. For example, in some census years, the CPH gathers data not only on individual households but also on the broader characteristics of the communities they reside in. This information is collected through an instrument referred to as the Barangay Schedule (Form 5). The data collected through the Barangay Schedule provides valuable insights into the social and economic conditions of different communities across the Philippines and is the primary data source in determining urbanity of communities.

The most recent CPH was conducted on May 1, 2020, reporting a population of 109 million Filipinos. This result showed a a slight decrease in population growth rate compared to the previous census in 2015. There were over 28 million housing units nationwide, with most being occupied. For illustrative purposes, the list of questionnaire items in the 2020 CPH are listed in Appendix 1.

3. International Data Harmonization Efforts

There are numerous data harmonization efforts globally, with some organizations and experts developing specific guidelines and frameworks tailored to the unique needs of different data types and the repositories they manage.

One of the most prominent data harmonization efforts is the Integrated Public Use Microdata Series (IPUMS) which harmonizes microdata from various national and international surveys and censuses. The initiative focuses on integrating demographic, employment, health, and other types of data into a uniform format for researchers to conduct comparative studies. IPUMS's approach emphasizes standardizing data documentation and variable coding, while also storing the raw form of the data sources for transparency.

The Survey Data Recycling (SDR) Framework, proposed by Słomczyński & Tomescu-Dubrow (2019), emphasizes the importance of maximizing the use of existing survey data. This framework underscores the value of recycling and harmonizing data from various sources across different political and cultural contexts. The framework provides harmonization techniques for variables and the evaluation of measurement equivalence across different surveys thereby addressing the challenges of data comparability. The SDR emphasizes the importance of metadata ("data about the data") that captures information about the source surveys and the decisions made during harmonization.

Peter Granda and Emily Blasczyk (2016) at the University of Michigan contributes to data harmonization efforts by providing an extensive set of principles for executing cross-cultural and comparative survey research in their Cross-Cultural Survey Guidelines (CCSG). These comprehensive guidelines cover all research phases, from conceptualization and design to data collection and analysis. These guidelines acknowledge the importance of considering cultural differences in surveys. They recommend using standardized questions whenever possible,

while also emphasizing strict quality control to reduce bias and ensure reliable comparisons across cultures.

Nihar Sayyed (2023) emphasizes the iterative process of data harmonization, due to the dynamic nature of data sources, structures, and business requirements. She also notes the application of machine learning in automating data harmonization as well as compliance with privacy laws and ethics.

Collectively, these efforts and frameworks are consistent in advocating the importance of data harmonization in maximizing the potential of existing surveys for comparative research. The tools and methods presented in these available resources allow researchers to combine and standardize data. These tools and general principles will be applied in this data for its own data harmonization effort.

4. Methodology

4.1. Data Harmonization Principles

This research employed a retrospective, or ex-post, data harmonization approach to reconcile inconsistencies in Philippine census data for each decade starting 1970 to 2020. This process aims to create a consistent dataset that allows longitudinal and trend analysis across different periods of censuses.

The harmonization was guided by the following main principles:

- **Identifying the lowest common denominator of detail**: This involved determining the most basic level of detail that exists consistently across all census datasets from 1970 to 2020. Through this process, the core variables that can be reliably compared over time are identified and constructed.

- **Preserving meaningful detail:** Although the least common denominator has been identified across census rounds, meaningful details or metadata about each individual census data source have been retained as much as possible. This involved retaining the raw variables in their original form as well as creating new variables or categories to document variations present in specific years.

- **Systematic documentation:** Throughout the harmonization process, all decisions made regarding variable transformations and data conversions have been carefully documented. This documentation will serve as a transparent record for data users and will be used to facilitate future updates or revisions to the harmonized dataset. Coding of variable changes have been done systematically similar to the IPUMS approach.

- **User access and transparency:** To support user access and transparency, the output of the data harmonization process includes a "user's database" in addition to the harmonized dataset. This database contains all relevant information related to the harmonization process, including original (raw) data sources, auxiliary data used for

- conversions or translation tables, detailed explanations of any data transformations or adjustments made, and metadata associated with the variables in the harmonized dataset

4.2. Data Sources

This study utilized data on the CPH every ten years from 1970 to 2020. This include the following sources of information:

1. CPH datasets

- CPH Form 2: Common Household Questionnaire: This form gathers basic demographic and housing information from all households enumerated in the Census.

- CPH Form 3: Sample Household Questionnaire: This form collects detailed demographic, social, and economic information from a representative sample of households.

2. CPH Technical Notes and Documentation:

   To ensure accurate interpretation of the data, technical notes and documentation published by the PSA were also consulted. These resources provide detailed information on questionnaire content, variable definitions, data collection methodologies, and any changes implemented across Census years.

3. Philippine Statistical Classification Systems:

   The following statistical classification systems were used to categorize and analyze the data:

- Philippine Standard Geographic Classification (PSGC): This system provides standardized geographic codes for regions, provinces, cities, municipalities, and barangays and is used to ensure consistency in geographic identifiers across datasets of the PSA. The PSGC is structured hierarchically, with each geographic unit assigned a unique numeric code. The PSGC database is regularly updated by the PSA to reflect changes in administrative boundaries, the creation of new local government units, and other geographic adjustments. Although these updates are documented in separate reports, there is no comprehensive historical mapping of geographic code changes for each area over time (horizontal changes). This gap creates challenges when integrating datasets collected in different periods because aligning geographic identifiers across census years often requires additional effort and assumptions. This data harmonization effort provides a consistent framework to address this issue and ensures that data from various census years can be integrated seamlessly.

- Philippine Standard Occupational Classification (PSOC): The PSOC organizes occupations into a hierarchical structure with increasing levels of detail, including major groups, sub-major groups, minor groups, and unit groups. The most recent version, the 2012 PSOC, aligns with international standards such as the International

Standard Classification of Occupations (ISCO-08) to ensure compatibility with global labor statistics and facilitate international comparisons.

- Philippine Standard Industrial Classification (PSIC)**:** The PSIC categorizes establishments and occupations by industry It is structured into sections, divisions, groups, and classes, providing a detailed framework for classifying economic activities. The latest version, the 2009 PSIC, adheres to the United Nations' International Standard Industrial Classification of All Economic Activities (ISIC, Rev. 4), ensuring consistency in economic reporting and alignment with global standards.

- Philippine Standard Classification of Education (PSCED)**:** The PSCED establishes a standardized framework for categorizing educational attainment, programs, and qualifications. It is organized by educational levels, fields of study, and qualifications to support detailed analysis of education trends. The 2017 PSCED reflects recent developments in the Philippine education system, including the introduction of the K-12 curriculum and technical-vocational education reforms, ensuring alignment with international classification systems like the International Standard Classification of Education (ISCED).

Although the PSOC, PSIC, and PSCED were harmonized as part of this study, the detailed classifications of occupations and industries were not consistently available in all CPH datasets. Consequently, these harmonized classifications could not be fully integrated into all rounds of census data. Efforts are currently underway to secure updated datasets with detailed information / codes from the PSA. Despite this limitation, the crosswalks / translation tables of the PSOC, PSIC, and PSCED are included in the data package, allowing researchers to use them with other applicable datasets, such as national surveys like the Labor Force Survey, and for broader analytical purposes. This inclusion ensures that users have access to robust and standardized classification systems to support diverse research and policy needs.

4.3. Harmonization Framework

The data harmonization process followed a structured approach to ensure consistency and comparability across datasets from different census years (Figure 1). This process involved several key stages, beginning with a thorough appraisal of available documentation. Documents from each CPH round were reviewed to understand the methodologies, enumeration protocol, sampling, coding systems, and data structures over time.

An inventory of variables was then conducted to assess all variables available in the datasets. This stage included an evaluation of variable coding, data management rules, data distributions, and data quality. Particular attention was given to ensuring the comparability of variables across census periods, identifying inconsistencies, and noting potential gaps in the data.

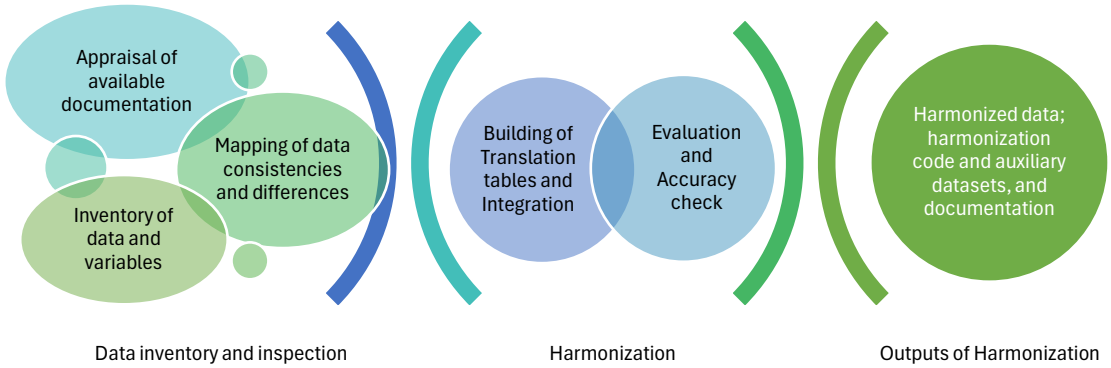
Following the inventory, variables were mapped across datasets to identify those that were completely identical, partially identical, or entirely unique (i.e., collected once only). This mapping accounted for differences in scale, definitions, and sample coverage to get clear understanding of the relationships between variables in different datasets.

To align variable definitions and codes, schema crosswalks and translation tables were developed. Conversion protocols were established to standardize different coding schemes, ensuring consistency across datasets. This step formed the foundation of the harmonization process by reconciling differences and creating a unified framework for analysis. Translation

tables were also constructed for statistical classification systems although not all were integrated in the harmonized data due to unavailability of detailed codes of variables.

The harmonization plan was then implemented, followed by evaluation of the harmonized data. Descriptive statistics, trends, and patterns in the harmonized data were compared against the original data sources to verify accuracy and to maintain fidelity. These were done to assess the impact of harmonization choices and ensure that the results remained robust.

Figure 1. Data Harmonization Process



Source: Author's rendition of the research process

All harmonization procedures were systematically documented to promote transparency and reproducibility. The documentation included detailed translation tables, conversion protocols (in the form of Stata do files), and validation of results. This ensures that future researchers clearely understand the contents and original of the harmonized data as well as build on the harmonization efforts in succeeding updates.

In all, the harmonization process produced key outputs that enhance the utility of the census datasets. These include both original and harmonized datasets, accompanied by a user database containing harmonization code, translation tables, and metadata for seamless integration and analysis. Comprehensive documentation of methods and evaluation processes was also developed to ensure transparency and reproducibility for future research and policy use.

5. Discussion

5.1. Review of Census Methodology

The sampling and enumeration procedures for the Philippine Census have evolved significantly over the decades. Each census round introduced methodological advancements to address growing data demands, and changes in population size, and geographic complexity. In general, the Philippine Census fully enumerates all households in the country using a common questionnaire to collect basic demographic, socio-economic, and housing information. Additionally, a sample of households is selected to answer an expanded questionnaire, which gathers more detailed data for in-depth analysis.

The Philippine Census regularly collects comprehensive data on various dimensions, including the demographic characteristics of the population, such as size, composition (e.g., age, sex, marital status), and geographic distribution. Socio-economic data are also gathered, covering birth registration, literacy, school attendance, employment and type of occupation and other indicators such as citizenship, ethnicity, and disability. At the household level, information on land ownership, ownership of household conveniences, tenure status and mode of acquisition of the housing unit was documented. Data on housing included details on structural characteristics (e.g., materials of the roof and walls), facilities, and the geographic location of housing units.

The following briefly describes the features of the different Census Rounds, particularly the sampling rates for the expanded household questionnaire:

1970 Census of Population. The 1970 Census utilized a dual-questionnaire approach for the household enumeration with long-form (1970 PH Form 2A) and short-form (1970 PH Form 2) questionnaires, which correspond to the sample and common household questionnaires. Approximately 5% of households, or one in every 20 households, were selected to answer the long-form questionnaire. This questionnaire collected detailed data on labor force participation, literacy, fertility, and vocational skills. The remaining 95% of households completed the short-form questionnaire, which focused on basic demographic characteristics. Enumeration was conducted on a de jure basis. Institutional residents, overseas workers, and military personnel were enumerated using specialized forms at their place of confinement or service. This census used static sampling rates and relied on individual households as the sampling unit.

1980 Census of Population. The 1980 Census employed systematic sampling to enhance geographic representation and data reliability. Enumeration Areas (EAs), which consisted of approximately 300 households, served as the primary sampling frame. One in every five households in each EA was selected to complete the sample household questionnaire (PH Form 3), resulting in a 20% sampling rate. A random start within each EA ensured unbiased representation. This census retained the static sampling rates of 1970 but improved household selection with the use of random starts.

1990 and 2000 Census of Population and Housing. The 1990 Census introduced systematic cluster sampling. Each enumeration area which consisted of 300 to 400 households were subdivided into clusters of five (5) households. The sampling rate depended on the size of the municipality where the EA is located. In municipalities with fewer than 500 households, all households were included. In municipalities with 501 to 1,500 households, one in every five households was sampled, resulting in a 20% sampling rate. In municipalities with more than 1,500 households, one in every 10 households was sampled, resulting in a 10% sampling rate. The 2000 Census followed the same systematic cluster sampling used in the 1990 CP.

2010 and 2020 Census of Population and Housing. The 2010 and 2020 censuses refined the cluster sampling method introduced in the last two decades. Municipalities with fewer than 500 households were fully enumerated, while municipalities with more than 500 households were sampled at a 20% rate. Like in the past rounds, clusters of five households were formed using consecutive serial numbers, and random starts ensured unbiased selection.

In summary, static sampling rates used in earlier censuses, such as the fixed 5% for long-form households in 1970 and 20% in 1980, were replaced by dynamic sampling rates starting in 1990. While earlier censuses emphasized geographic representation at the EA level, later censuses adopted stratified sampling methods based on municipality size. This shift ensured proportional coverage of both urban and rural areas while optimizing resources and maintaining data quality. The 1980 Census introduced systematic sampling with random starts to ensure unbiased household selection. The 1990 Census advanced this methodology by implementing systematic cluster sampling, which improved stratification and enhanced proportional representation across different populations and geographic strata. These innovations collectively reflect the progressive refinement of enumeration and sampling methods in the Philippine Census, ensuring accuracy, efficiency, and adaptability to evolving data needs.

5.2. Inventory of available data

The CPH datasets included in the research vary across census years (Table 1). For 1970 and 1980, only Form 3 (Sample Households) data were available, with 253,158 and 430,050 household records, and 1,651,506 and 2,260,602 person records, respectively. Form 2 (Common Households) data were not available for these years.

Table 1. Summary of data sources

Census Year	Form 2 (Common households)		Form 3 (Sample Households)	
	Household record[a]	Person Record[b]	Household record[a]	Person Record[b]
Number of data observations				
1970[c]	*No data*	*No data*	253,158	1,651,506
1980[c]	*No data*	*No data*	430,050	2,260,602
1990	11,554,870	61,087,698	1,155,917	6,013,913
2000	15,275,046	76,313,481	1,511,718	7,417,810
2010	21,745,707	92,097,978	4,133,649	18,824,651
2020	29,706,049	108,667,043	5,223,870	21,322,739
Number of variables				
1970	*No data*	*No data*	*No data*	54
1980	*No data*	*No data*	28	39
1990	23	22	34	44
2000	17	22	46	35
2010	16	29	54	42
2020	25	31	67	41

Notes: a/ Refers to subset of CPH data containing household-level information
b/ Refers to subset of CPH data containing information of household members
c/ the PSA is currently converting archived copies of old census data to digital format. This may or may not include full enumeration data of the 1970 and 1980 census rounds.
Source of basic data: CPH 1970 to 2020, PSA

From 1990 onwards, data availability improved significantly. Form 2 and Form 3 datasets were available, with Form 2 containing 11,554,870 household records and 61,087,698 person records in 1990, increasing to 29,706,049 household records and 108,667,043 person records in 2020.

The number of variables expanded over time as census questionnaires became more detailed in subsequent data collection rounds. Earlier censuses, such as those in 1970 and 1980, contained fewer variables, ranging from 28 to 54. In later censuses, particularly for Form 2 (Common Households) and Form 3 (Sample Households), the number of variables increased significantly, reaching 31 to 67 by 2020. The 2020 census represents the most comprehensive data set, across all forms with the highest number of observations and variables.

5.3. Standardization of Variable Definitions

The harmonization process involved standardizing variable definitions across multiple census years to ensure consistency and comparability. This standardization was particularly complex due to evolving data collection methods and changing data needs over the five decades (Table 2).

For core demographic variables, such as age, sex, and marital status, standardization was relatively straightforward as these maintained consistent definitions across census rounds. However, other variables required more extensive harmonization. For example, household composition variables like "relationship to household head" expanded from 9 categories in 1970 to 26 categories in 2020. These were reconciled by creating a simplified set of harmonized categories that preserved key relationship distinctions while maintaining consistency across all periods.

Educational attainment categories underwent significant changes, particularly with the introduction of the K-12 curriculum and the adoption and subsequent expansion of the PSCED categories in recent rounds. The harmonization process created standardized education levels that could be consistently interpreted across all census years. Similarly, employment-related variables required careful standardization. The "class of worker" variable, for instance, was harmonized to five main categories based on the 1970 census classification, which served as the lowest common denominator across all years.

Housing characteristics also required substantial standardization. Variables like "construction materials for roof and walls" and "toilet facilities" had varying levels of detail across census years. The harmonization process created standardized categories that captured the essential distinctions while maintaining consistency across all periods. For example, roofing materials were standardized into five main categories: galvanized iron/aluminum, tile/concrete/brick/stone, asbestos, makeshift/salvaged materials, and others.

5.4. Translation Tables or Crosswalks for Coding Schemes

To address inconsistencies in coding schemes across census years, comprehensive crosswalks were developed for key classification systems. The Philippine Standard Geographic Classification (PSGC) crosswalk maps location codes from 1990 to 2020, enabling consistent geographic analysis across these periods. For 1970 and 1980 data, geographic harmonization was limited to the provincial level due to the unavailability of detailed municipal identifiers.

Occupation codes underwent significant changes, with varying levels of detail across census years. The 1970, 1980, 2000, and 2010 censuses used 3-digit occupation codes, while the 1990 census used 4-digit codes, and the 2020 census used single-digit major occupation groups. Crosswalks were developed to map these different classification systems to a common framework, though some granularity was necessarily lost in harmonizing to the lowest common denominator.

Industry classifications similarly varied in detail, from section-level codes in 2020 to 4-digit industry classes in 2010. Crosswalks were created to align these different classification schemes, with undefined codes from earlier periods (particularly 1970) being grouped under appropriate major categories in the harmonized dataset.

For educational classifications, crosswalks were developed to account for changes in the education system, including the transition to K-12. These crosswalks map educational attainment across different periods while maintaining the ability to track educational progress consistently over time.

Crosswalks were created using a combination of manual review and automated algorithms (e.g., fuzzy matching) to ensure accuracy, and validation checks were performed by comparing harmonized codes against official PSA classifications. These crosswalks are included in the data package to enable researchers to apply them for other datasets, such as the Labor Force Survey or Family Income and Expenditure Survey and other national surveys collected by the PSA or other agencies.

Table 2. Overview of the Harmonization Process by variable group

Variable Group	Data availability		Description of Data Pre-Harmonization	Harmonization Process
	Common HHs	Sample HHs		
Geographic Location *(Region, Province, Municipality, Barangay)*	Available in CPH 1990 to 2020	Available up to municipality level in CPH 1970 to 2020	- 1990 to 2020: Location data are available up to the barangay level. - 1970 to 1980: Municipality codes are sequential within provinces, but municipal and barangay identification are not available.	- Crosswalks for PSGC codes were developed to map corresponding codes for all barangays from 1990 to 2020. These crosswalks enable translation of PSGC codes to their versions in 1990, 2000, 2010, and 2020. - PSGC codes for 1970 and 1980 data were harmonized at the province level only, due to the absence of identifying information for municipalities.
Household size	Available in CPH 1990 to 2020	Available in CPH 1970 to 2020	Derived from the number of household members in the enumeration of household members (Person record)	Construction of household size variable based on the number of household members
Relationship to Household Head	Available in CPH 1990 to 2020	Available in CPH 1970 to 2020	Level of detail varied across decades, with the 1970 Census including only 9 relationship-to-head codes, compared to 26 codes in the 2020 CPH.	Recoding of values to least common denominator codes are as follows: 1 Head 2 Wife/Husband 3 Son/Daughter 4 Son/Daughter-in-law 5 Grandchild 6 Other relative 7 Not related 9 Missing Value
Age	Available in CPH 1990 to 2020	Available in CPH 1970 to 2020	Age is reflected as age in years as of last birthday	Minor data recoding needed for age 80 years old as they are lumped together in 2020 CPH.

Variable Group	Data availability		Description of Data Pre-Harmonization	Harmonization Process
	Common HHs	Sample HHs		
Sex	Available in CPH 1990 to 2020	Available in CPH 1970 to 2020	All datasets follow the same coding scheme: Male = 1; Female=2	No data recoding needed
Marital status	Available in CPH 1990 to 2020	Available in CPH 1970 to 2020	Minor differences in codes in Census Rounds Earlier censuses code married and living together under one category.	Recoding of values to least common denominator codes are as follows: 1 Single 2 Married / Common-law/Live in 3 Widowed 4 Divorced/Separated 5 Others 6 Unknown
Religion	Available in CPH 1990 to 2020	Available in all rounds except 1980	Level of detail varied across decades, with the 1970 Census having only 9 religion codes, compared to 129 codes in the 2020 CPH. No religion data was collected in the 1980 sample household data	The least common denominator codes are as follows: 1 Roman Catholic 2 Protestant 3 Iglesia Mi Cristo 4 Aglipayan 5 Islam 6 Buddhist 7 Others 8 None 9 Missing Value However, translation table / crosswalk was constructed to map the code correspondence of the more detailed types of religion in latter decades.
Country of Citizenship	Available in CPH 1990 to 2020	Available in CPH 1970 to 2020	Level of detail varied across decades. Documentation available for the 1970 Census contains only definitions for 11	The least common denominator codes are based on the 1970 dataset as follows: 01 Philippines 04 China

Variable Group	Data availability		Description of Data Pre-Harmonization	Harmonization Process
	Common HHs	Sample HHs		
			codes. The highest number of codes is 203 codes in the 2010 CPH.	05 India 06 Indonesia 14 Pakistan 55 France 60 Netherlands 63 Spain 66 U.K. (Britain) 72 United States 300 Others -9 Missing Value However, translation table / crosswalk was constructed to map the code correspondence of the more detailed countries of citizenship in latter decades.
Ethnicity	Available in CPH 1990 to 2020	Data is only available in 2000 to 2020	Level of detail varied across decades: 148 codes in 2000, 183 codes in 2010, and 290 codes in 2020 CPH	Translation table was constructed to map the correspondence of codes across the three rounds of census. Codes that were clustered in earlier rounds were mapped to the more detailed subcategories in latter rounds, and vise versa. Ethnicity types without direct correspondence in the two other rounds are assigned to their own code.
Disability	Available in CPH 1990 to 2020	Data is only available in 1990 to 2020	Data for 1990 and 2000 CPH uses slightly different sets of codes for the disability types. Typology and questions has been updated in the 2000 and 2020 CPH to “functional disability” categories.	Lowest common denominator for the 4 decades is the reported presence or absence of disability. Translation table is available for the correspondence of codes between 1990 and 2000 types of disabilities.

Variable Group	Data availability		Description of Data Pre-Harmonization	Harmonization Process
	Common HHs	Sample HHs		
Residence from 5 years ago (Province and Municipality)	Available in CPH 1990 to 2020	Available in CPH 1970 to 2020	- 1990 to 2020: Location data are available up to the barangay level. - 1970 to 1980: Municipality codes are sequential within provinces and cannot be identified	Crosswalks/ translation tables were constructed at the municipal level for 1990 to 2020 CPH and at the province level for 1970 and 1980.
Education (HGC)	Available in CPH 1990 to 2020	Available in CPH 1970 to 2020	Level of detail varied across decades with 1970 and 1980 having 2-digit education codes (i.e., undergraduate degree grouped according to field); 1990 CPH report 4-digit codes; 2000, 2010 and 2020 CPH report 3 digit codes;	Crosswalks for PSCED versions were developed to map corresponding codes across decades. Codes that were clustered in earlier rounds were mapped to the more detailed subcategories in latter rounds, and vise versa. Education program types without direct correspondence in the other rounds are assigned to their own code. Additional variable was created for number of years of schooling.
Literacy	Available in CPH 1990 to 2020	Available in CPH 1970 to 2020	All datasets follow similar coding scheme: Yes/ No or Literate or Not	Minor recoding needed to to harmonize codes
Language spoken at home	Not collected	Available in CPH 1970 to 1990 at the individual level. Available in the CPH 2000 and 2020 at the household level	Collected per household member in the 1970, 1980, and 1990 Censuses. Collected at the household level in the 2000 and 2020 CPH. The 2010 CPH did not collect this data. Codes refer to	Data was processed to convert 1970 to 1990 individual data to household level (most reported language spoken by the household members). Crosswalk / translation table for language codes across the years were constructed.
Able to speak Filipino/ English	Not collected	Filipino: Available in 1970, 1980, 1990	Question on whether member can speak Filipino/ Tagalog was asked from 1970 to 1990.	Minor recoding done to standardize missing values for each variable.

Variable Group	Data availability		Description of Data Pre-Harmonization	Harmonization Process
	Common HHs	Sample HHs		
		Spanish:Available in 1970 English: Available in 1970, 1980, and 2000	Question on whether member can speak English was asked in 1970, 1980, and 2000. 1970 Census also asked members whether they can speak spanish.	
Occupation	Not collected	Available in CPH 1970 to 2020	Level of detail varied across decades with 1970, 1980, 2000 and 2010 having 3-digit occupation codes; 1990 CPH report 4-digit codes; while 2020 CPH report only single digit codes or major occupation groups.	Crosswalks for PSOC versions were developed to map corresponding occupation codes across decades. Occupation types without direct correspondence in the other rounds are assigned to their own code.
Industry / Kind of business	Not collected	Available in CPH 1970 to 2020	Level of detail varied across decades with 2020 CPH reporting up to sections only (one digit codes), 1980 and 2000 CPH reporting up to 2 digits, 1990 CPH reporting 3-digit codes, and 2010 CPH reporting industry classes (4-digit) codes. 1970 data report 4-digit codes but only major groups can be identified based on available documentation.	Crosswalks for PSIC versions were developed to map corresponding occupation codes across decades as much as possible. Undefined codes in the data (e.g., 1970) were grouped under major groups in the harmonized data.
Class of worker	Not collected	Available in 1970, 2000, 2010, and 2020 CPH	Response codes are consistent from 2000 to 2020 CPH while 1970 has two fewer response codes since some categories were split/subcategorized in the latter censuses.	Least common denominator is based on 1970 Census with fewer categories: 1 Working for private employer for wage, salary, commission, tips, etc. 2 Working for government or governmentowned or controlled corporation

Variable Group	Data availability		Description of Data Pre-Harmonization	Harmonization Process
	Common HHs	Sample HHs		
				3 In own business, farm, profession or trade for profit or fees without paid employees 4 Employer in own business, farm, profession or trade for profit or fee (with one or more paid employees) 5 Working without pay on family farm or Enterprise
Place of work	Not collected	Available in 1990 to 2020 CPH	1990 to 2020 CPH have codes at the province and municipality levels. Note that 1980 collected information on Place of School OR Work while 2000 CPH also collected information on the Place of School	Crosswalks/ translation tables were constructed at the municipal level for 1990 to 2020 CPH.
Overseas worker indicator	Available in CPH 1990, 2000, 2010, 2020	Available in CPH 1990, 2000, 2010, 2020	Coded as indicator (yes/no) for overseas contract workers	Minor recoding to standardize missing values.
Children ever born/ still alive	Not collected	All years except 2000	1970 and 1980 data collected number of children born alive and still alive by gender. 1990, 2010, and 2020 CPH only report total for both sexes.	Minor processing of data (i.e., summation) for 1970 and 1980 datasets. Counts of 8 or more are coded under one category consistent with the 2020 CPH.
Age at first marriage	Not collected	All years except 2000	Age (in years) reported for all datasets with minor differences for missing labels.	Minor recoding of missing values to make label and codes consistent.
Construction Materials of the Roof	Available in CPH 1980 to 2020	Available in CPH 1980 to 2020	Response categories are consistent from 1990 to 2020. 1980 data has fewer categories.	Least common denominator are the following categories: 1 Galvanized iron/Aluminum 2 Tile/Concrete/Brick/Stone

Variable Group	Data availability		Description of Data Pre-Harmonization	Harmonization Process
	Common HHs	Sample HHs		
				3 Asbestos 4 Other (Cogon/Nipa/ anahaw, Wood/bamboo, etc.) 5 Makeshift/Salvaged materials "Half Galvanized Iron and Half Concrete" code in latter censuses were coded under Tile/Concrete/Brick/Stone category Translation table is available to to allow direct correspondence of 1990 to 2020 codes.
Construction materials of the outer walls	Available in CPH 1980 to 2020	Available in CPH 1980 to 2020	Response categories are consistent from 1990 to 2020. 1980 data has fewer categories (no glass)	Harmonized categories are as follows: 1 Galvanized iron/Aluminum 2 Tile/Concrete/Brick/Stone 3 Wood/Plywood 4 Mixed tile/Concrete/Brick/ Stone and Wood/Plywood 5 Asbestos 6 Bamboo/Sawali 7 Cogon/Nipa 8 Makeshift/Salvaged materials 9 Other (anahaw, etc.) 10 No walls Translation table is available to to allow direct correspondence of 1990 to 2020 codes.
Type of building	Available in CPH 1980 to 2020	Available in CPH 1980 to 2020	Response categories vary across decades, but codes are consistent from 1990 to 2010.	Least common denominator is based on the 1990 CPH response categories: 1 Single house 2 Duplex 3 Multi-unit res. 4 Commercial

Variable Group	Data availability		Description of Data Pre-Harmonization	Harmonization Process
	Common HHs	Sample HHs		
				5 Institution 6 Other
Tenure status of house/ Tenure status of lot	Available in CPH 1980 to 2020 with variations. For 2010 only tenure status of lot is included in common questionnaire	Available in CPH 1980 to 2020 with variations	For CPH 1980 to 2010, the tenure status of the of the housing unit and tenure of the lot are asked separately. For CPH 2020 the tenure status of house or lot is asked in one question.	Response codes for the CPH 2020 were recoded to create separate variables for tenure status of housing unit and tenure status of lot. The following are the LCD categories: 1 Owner 2 Lessee or Sublessee 3 Other Legal Tenure 4 No Tenure/ No consent (squatter, etc.)
Year building was built	Available in CPH 1980 to 2020	Available in CPH 1980 to 2020	Response categories vary across decades with most recent five years recorded as individual codes and earlier years grouped by 5 year durations.	Least common denominator are categories based on 2020 CPH: 2016 – 2020 2011 – 2015 2001 – 2010 1991 – 2000 1981 – 1990 1980 or earlier No response
Floor area of the housing unit	Available in CPH 1980 to 2020	Available in CPH 1980 to 2020	Response categories vary across decades. 1990 codes are similar to the 2000 codes while 2010 codes are similar to the 2020 codes.	Least common denominator categories are as follows: 1 Less than 30 sq. meters 2 30- 49 sq. meters 3 50- 69 sq. meters 4 70- 149 sq. meters 5 150-199 sq. meters 6 200 sq. meters and over
State of repair	Available in CPH 1990, 2000, 2010	Available in CPH 1990, 2000, 2010, 2020	Response categories consistent across years.	No recoding needed List of codes as follows: 1 Needs no Repair/Needs Minor Repair

Variable Group	Data availability		Description of Data Pre-Harmonization	Harmonization Process
	Common HHs	Sample HHs		
				2 Needs Major Repair 3 Dilapidated/Condemned 4 Under Renovation/Being Repaired 5 Under construction 6 Unfinished Construction 9 Not Reported
Acquisition of the housing unit	Not collected	Available in CPH 1990, 2000, 2010, 2020	Response codes vary. 1990 and 2000 CPH follow the same response codes while 2010 and 2020 CPH follow the same response codes.	Least common denominator reconciling codes across the years as follows: 1 Inherited 2 Company Benefit 3 Purchased / Constructed 4 Others (Gift, Lottery) 9 Not Reported
Source of financing of the housing unit, by type	Not collected	Available in CPH 1990, 2000, 2010, 2020	Response codes are consistent across the years	Final response codes as follows: 1 Own Resources/Interest Free Loans From Relatives/Friends 2 Government Assitance (PAG-IBIG, GSIS, SSS, DBP, etc) 3 Private Banks/Foundations/ Cooperatives 4 Employer's Assistance 5 Private Persons 6 Others
Monthly rent	Not collected	Available in CPH 1990, 2000, 2010, 2020	Response codes for range of amounts vary across the years.	Range of amounts reconciled to arrive at the following set of codes: 1 500 and below 2 501 to less than 2000 3 2000 to less than 10,000 4 10000 and over

Variable Group	Data availability		Description of Data Pre-Harmonization	Harmonization Process
	Common HHs	Sample HHs		
Land ownership (residential, agricultural, others)	Available in CPH 1990 and 2020	Available in CPH 1990, 2000, 2010, 2020	Data contains responses to questions on ownerhip of other residential lands, agricultural lands, and other types of land. Agricultural land acquired through CARP was asked starting 2000.	Variables retained are ownership of other residential land, agricultural land (whether through CARP or not), and other types of land.
Fuel for lighting	Not collected	Available in CPH 1980, 2000, 2010, 2020	Response categories were consistent through the years except for the addition of solar panel as its own category/option in 2020 CPH.	Solar panels are recoded under the others category. Final codes are as follows: 1 Electricity 2 Kerosene (gaas) 3 Liquefied Petroleum Gas (LPG) 4 Oil (Vegetable, Animal, etc) 5 Others
Fuel for cooking	Not collected	Available in CPH 1980, 2000, 2010, 2020	Response categories were consistent from 1990 to 2020 with minor differences in "missing" codes. 1980 census reports "wood" and "charcoal" as one response code while latter CPH rounds report the responses separately.	Minor recoding needed to reconcile codes. Final responses are as follows: 1 Electricity 2 Kerosene (gaas) 3 Liquefied petroleum gas (LPG) 4 Charcoal, Wood 5 Others 0 None
Kind of toilet facility	Not collected	Available in CPH 1980, 1990, 2000, 2010, 2020	Response categories were generally consistent through the years.	Final responses are as follows: 1 Water-sealed, Sewer Septic Tank, used exclusively by HH 2 Water-sealed, Sewer Septic Tank, shared 3 Water-sealed, Other Depository used exclusively by HH 4 Water-sealed, Other Depository, shared 5 Closed pit 6 Open pit

Variable Group	Data availability		Description of Data Pre-Harmonization	Harmonization Process
	Common HHs	Sample HHs		
				7 Others (pail system etc.) 8 None
Garbage disposal	Not collected	Available in CPH 1990, 2000, 2010	Response categories were generally consistent through the years.	Final responses are as follows: 1 Picked up by garbage truck 2 Dumping in individual pit (not burned) 3 Burning 4 Composting 5 Burying 6 Feeding to animals 7 Others 8 Not Reported
Source of water supply for drinking/ washing clothes/ cooking	Not collected	Available in CPH 1980 to 2020 with variations in specific information collected	1980 and 2010 CPH asked the source of water for drinking, laundry, and cooking. 2020 CPH only asked the source of water for drinking and cooking. 2000 CPH asked the source of water for drinking or cooking, and laundry or washing. 1990 CPH only asked the source of water for drinking.	Data were harmonized based on the availability per year: Water for drinking = 1980, 1990, 2010, and 2020 Water for drinking AND cooking = = 1980, 1990, 2000, 2010, and 2020 Water for laundry & washing = 1980, 1990, 2000. 2010, and 2020 Least common denominator for response codes: 1 Own use, faucet, Community Water System 2 Shared, faucet, Community Water System 3 Own use, Tubed/Piped Deep Well 4 Shared, Tubed/Piped Deep Well 5 Tubed/Piped shallow Well 6 Dug well

Variable Group	Data availability		Description of Data Pre-Harmonization	Harmonization Process
	Common HHs	Sample HHs		
				7 Spring, Lake, River, Rain, etc. 8 Peddler 9 Others (Bottled water, water refilling station)
Household conveniences	Not collected	Available in CPH 1980 to 2020 with variations in specific information collected	Only TV, Refrigerator, and Radio are consistently collected across the years.	Recoding needed to reconcile labels for assets and codes for missing/no responsevalues

Note: List excludes variables that are only available in one or two rounds of data collection only. However, original variables are still retained in the data package.

5.5. Overview of the Harmonization and Patterns in the Census Data

The harmonization process across five decades of Philippine census data revealed both the evolution of data collection practices and the inherent challenges of creating consistent longitudinal datasets. The most successful harmonization occurred with fundamental demographic indicators. Variables like sex maintained perfect consistency across all census years with a simple binary coding scheme, while age data required only minor recoding for the 80+ age group in 2020. Marital status needed more extensive harmonization to reconcile how relationships like "common law" marriages were coded differently over time, ultimately being standardized into five main categories: single, married/common-law, widowed, divorced/separated, and others.

Geographic harmonization revealed a clear divide between pre-1990 and post-1990 data quality. For the period 1990-2020, location data was successfully harmonized down to the barangay level using PSGC crosswalks. However, 1970-1980 geographic harmonization was limited to the provincial level due to the use of sequential municipal codes that could not be mapped to modern geographic identifiers. Migration data, tracking residence from 5 years prior, was harmonized at the municipal level for 1990-2020 and at the provincial level for earlier periods.

The harmonization of socio-cultural variables became increasingly complex over time. Religion expanded from 9 codes in 1970 to 129 codes in 2020, requiring careful aggregation into 8 major categories while preserving detailed mappings in translation tables. Ethnicity data was only available from 2000 onwards, expanding from 148 codes to 290 codes by 2020. Language data underwent a methodological shift from individual-level collection (1970-1990) to household-level collection (2000-2020), requiring careful aggregation of individual responses to create comparable household-level indicators.

Educational variables showed significant evolution in classification systems, reflecting major changes in the Philippine education system. The harmonization addressed changes from pre-K-12 to K-12 curriculum structures, with early censuses (1970-1980) using 2-digit education codes while later ones employed 3-4 digit codes. Additional variables were created to standardize years of schooling across different educational systems. Literacy measurements remained relatively consistent throughout the period, requiring only minor code standardization.

Employment-related variables required complex harmonization approaches. Occupational classifications varied from simple 1-digit codes to detailed 4-digit codes across censuses, while industry classifications showed similar variation, with 2020 using section-level codes whereas 2010 used detailed 4-digit classifications. Class of worker categories were standardized based on 1970 classifications, providing five consistent categories across all periods. Overseas worker indicators were successfully harmonized from 1990 onwards, reflecting increasing attention to international labor migration.

Housing characteristics and amenities demonstrated both consistency and evolution over time. Construction materials for roofs and walls maintained relatively stable categories from 1980 onwards, while utility access (water, electricity, toilet facilities) showed increasing detail in classification over time. Floor area measurements required reconciliation of different category boundaries across census years. Housing tenure required careful harmonization to account for separate house and lot tenure questions in some years versus combined questions in others.

The increase in indicators related to household welfare and quality of life revealed expanding interest in measuring living standards over time. Household conveniences were tracked consistently for basic items (TV, refrigerator, radio) across all periods, while internet access and digital devices were naturally limited to recent censuses. Garbage disposal and sanitation facilities showed increasing detail in classification over time, reflecting growing attention to environmental and public health concerns.

The harmonization process highlighted several key methodological patterns. Nearly all variables showed increasing detail and complexity in more recent censuses, while major classification systems underwent significant changes requiring careful crosswalk development. Core demographic variables maintained the highest consistency across time, while socio-economic variables showed the most variation. The 1990 census emerged as a key turning point, with significantly improved data quality and detail from this period onwards.

The harmonization process emphasized thorough documentation throughout its implementation. Translation tables were created for all major classification systems, and detailed crosswalks documented the relationships between coding schemes across censuses. Variable-specific documentation noted particular challenges or limitations in harmonization, providing clear guidance for researchers on appropriate use of harmonized variables. This comprehensive effort successfully balanced the need for standardization with the preservation of meaningful distinctions in the data, though some variables necessarily lost granularity when harmonized to their lowest common denominator across census years.

5.6. Limitations of the Data Harmonization

While this study has addressed many harmonization challenges in Philippine Census data from 1970 to 2020, some limitations remain that users should consider when analyzing the harmonized dataset:

Variable Detail Resolution: While core variables like educational attainment and household relationships have been standardized across census years, some variables required significant aggregation to achieve consistency. For example, occupational categories expanded from 3-digit codes in earlier censuses to 4-digit codes in 1990, then returned to broader classifications in 2020, requiring harmonization to broader groupings.

Incomplete Historical Records: The research incorporated available sample household data (Form 3) from 1970 and 1980, but the Common Household Questionnaire (Form 2) data for these years remain unavailable. This limits the comprehensiveness of certain analyses for these early periods to sample household data only.

Geographic Boundary Changes: The harmonized dataset provides consistent geographic codes aligned with PSGC standards from 1990 onwards at the municipal level. However, geographic analysis for 1970-1980 remains limited to the provincial level, as these earlier censuses used sequential municipal codes that cannot be mapped to current geographic identifiers.

Classification System Details: While occupation and industry codes have been harmonized using translation tables across census years, the full integration of detailed classifications (4-digit codes) into the harmonized dataset was not possible due to unavailability of detailed codes in the raw CPH datasets. However, the harmonized versions of PSOC, PSIC, and PSCED are included in the data package for use with other applicable datasets.

Variable Coverage Evolution: The harmonization successfully aligned core demographic variables across all census years, but certain variables reflect the evolution of data collection over time. For instance, ethnicity data is only available from 2000 onwards, while language collection shifted from individual-level (1970-1990) to household-level (2000-2020).

Documentation Gaps: Though the harmonization process itself is thoroughly documented with translation tables and crosswalks, some aspects of the original data collection methodology in earlier censuses remain unclear due to limited historical documentation.

Census Round Selection: This study focused on decennial census data from 1970 to 2020, excluding midyear censuses (1975, 1995, 2007, and 2015). This selection was methodologically appropriate as midyear censuses did not implement the sample household questionnaire with expanded variables, ensuring consistency in variable coverage across harmonized rounds.

These remaining limitations reflect inherent challenges in historical data harmonization that could not be fully resolved even with robust methodological approaches. However, the harmonized dataset still provides a valuable resource for longitudinal analysis, particularly for core demographic and social indicators that have been successfully standardized. Users should consider these limitations when designing their analyses and interpret results within the context of these constraints.

The documentation accompanying the harmonized dataset includes detailed variable-specific notes to help users understand where harmonization was most successful and where additional caution in interpretation may be warranted. Future updates to the harmonization process may address some of these remaining limitations as additional historical documentation is discovered, or new methodological approaches are developed.

6. Summary and Recommendations

This study successfully addressed many key challenges in harmonizing Philippine Census data from 1970 to 2020, while acknowledging certain inherent limitations that persist. Through careful standardization of variable definitions, alignment of classification systems, and comprehensive documentation of transformations, the research has produced a harmonized dataset that allows meaningful longitudinal analysis of important demographic and social indicators. The harmonized dataset is particularly robust for the period from 1990 onwards, with some limitations in earlier periods due to data availability and documentation constraints. The harmonized dataset offers several valuable applications such as the following potential use cases:

1. Population Management and Urban Planning: The dataset enables analysis of urbanization and migration trends from 1990 onwards, with consistent geographic coding aligned to PSGC 2020 standards. While a more granular geographic analysis for 1970-1980 is more limited, the dataset still provides valuable insights into broad demographic shifts across provinces and regions.

2. Education and Workforce Development: Standardized educational attainment categories allow for consistent tracking of educational progress across all periods. Labor force analysis is most detailed from 1990 onwards, with two-digit occupation

codes available, while broader occupational trends can be analyzed across the full period using harmonized one-digit codes.

3. Health and Social Protection: Core demographic variables and household composition data are consistently available across all periods, supporting long-term analysis of aging trends and household structures. However, specialized indicators like disability status and detailed health metrics are only available from 2000 onwards.

4. Socioeconomic Analysis: Housing characteristics and basic household assets can be tracked consistently from 1980 onwards, though with varying levels of detail across periods.

The study proposes the following recommendations based on the harmonization experience and remaining challenges:

1. Historical Data Recovery: The digitization and recovery of Form 2 (Common Household) data from 1970 and 1980 should be prioritized, along with any additional documentation that could enhance understanding of early census methodologies.

2. Classification System Documentation: The development of comprehensive crosswalks between different versions of classification systems (PSOC, PSIC, PSCED) is needed, including detailed documentation of how categories evolved over time.

3. Geographic Reference System: A historical geographic reference system should be created to track all administrative boundary changes since 1970, enabling more precise geographic analysis across census periods.

4. User Guidelines: Detailed guidelines should be developed for researchers on appropriate uses of different variables across time periods, best practices for handling partially harmonized variables, methods for addressing missing data and classification changes, and techniques for assessing and reporting uncertainty in longitudinal analyses

5. Continuous Improvement: A systematic process should be established for incorporating newly discovered historical documentation, updating harmonization approaches as new methodologies emerge, expanding the scope of harmonized variables where possible, and addressing user feedback and analytical needs

The implementation of these recommendations will enable the Philippine Statistical System to build upon this harmonization effort to provide increasingly robust and well-documented data for longitudinal analysis.

Bibliography

"CESSDA - Consortium of European Social Science Data Archives." n.d. Accessed March 26, 2024. https://www.cessda.eu/.

Concepción, Mercedes B, ed. 1977. "POPULATION OF THE PHILIPPINES." University of the Philippines - Population Institute.

Fortier, Isabel, Parminder Raina, Edwin R. Van den Heuvel, Lauren E. Griffith, Camille Craig, Matilda Saliba, Dany Doiron, et al. 2017. "Maelstrom Research Guidelines for Rigorous Retrospective Data Harmonization." *International Journal of Epidemiology* 46 (1): 103–5. https://doi.org/10.1093/ije/dyw075.

Gannett, Henry. 1905. "The Philippine Census." *Bulletin of the American Geographical Society* 37 (5): 257–71. https://doi.org/10.2307/198467.

Granda, Peter, and Emily Blasczyk. 2016. "CCSG." 2016. https://ccsg.isr.umich.edu/chapters/data-harmonization/#one.

Integrated Public Use Microdata Series (IPUMS) https://international.ipums.org/international/

PSA. 2020 Census of the Population and Housing Data. May 08, 2023. Downloaded from the PSADA Microdata Catalog: https://psada.psa.gov.ph/catalog/231

———.2010 Census of the Population and Housing Data. Sep 13, 2021. Downloaded from the PSADA Microdata Catalog: https://psada.psa.gov.ph/catalog/13

———.2000 Census of the Population and Housing Data. Sep 13, 2021. Downloaded from the PSADA Microdata Catalog: https://psada.psa.gov.ph/catalog/14

———.1990 Census of the Population and Housing Data. Sep 13, 2021. Downloaded from the PSADA Microdata Catalog: https://psada.psa.gov.ph

Nozawa, Katsumi. n.d. "History of the Philippine Statistical System." Accessed March 26, 2024. https://www.ier.hit-u.ac.jp/COE/Japanese/Newsletter/No.13.english/Nozawa.html.

Pan, Ke, Lydia A Bazzano, Kalpana Betha, Brittany M Charlton, Jorge E Chavarro, Christina Cordero, Erica P Gunderson, et al. 2023. "Large-Scale Data Harmonization Across Prospective Studies: The Preconception Period Analysis of Risks and Exposures Influencing Health and Development (PrePARED) Consortium." *American Journal of Epidemiology* 192 (12): 2033–49. https://doi.org/10.1093/aje/kwad153.

Sayyed, Nihar. 2023. "Data Harmonization: Steps and Best Practices." September 7, 2023. https://datavid.com/blog/data-harmonization.

Slomczynski, Kazimierz M, and Irina Tomescu-Dubrow. 2013. "About the SDR Project." *Survey Data Recycling* (blog). November 19, 2013. https://wp.asc.ohio-state.edu/dataharmonization/about/.

University of the Philippines Population Institute. PopArchive. 1980 Census of Population and Housing. Requested via email on January 5, 2024.

_____. 1970 Census of Population and Housing. Requested via email on January 5, 2024.

Appendices

Appendix 1. Data items in the 2020 CPH questionnaires

Table 3. Data Items in the CPH 2020 Common Household Questionnaire

ID variables	Household-level variables	Individual-level variables
• Region • Province/Highly Urbanized City • City/Municipality • Barangay • Urban-Rural Classification • Housing Unit Serial Number • Household Serial Number	• Type of Building • Number of Floors of the Building • Construction materials of the roof and outer walls • Construction and Finishing materials of the floor • State of repair of the building • Year building was built • Floor area of the housing unit • Tenure status of the housing unit/lot • Land Ownership - Other residential land/s, Agricultural, Agricultural acquired from CARP, Other land/s • Presence of operator in crop farming, livestock, etc. • Language/dialect generally spoken at home • Residence Five (5) Years from Now • Household Indicator	• Line Number • Relationship to household head • Sex • Age as of Last Birthday • Birth Registration Status • Copy of Birth Certificate • Marital Status • Religious Affiliation • Citizenship & Country of citizenship • Ethnicity • Functional difficulties, by type • Residence of Mother at the Time of Birth of the Household Member • Residence Five (5) Years Ago • Literacy • Highest Grade/Year Completed • Overseas Worker

Table 4. Additional Data Items in the CPH 2020 Sample Household Questionnaire

Household-level variables	Individual-level variables
• Household-level variables • Acquisition of the housing unit • Source of financing of the housing unit • Monthly rental of the housing unit (range) • Usual manner of kitchen garbage disposal • Kind of toilet facility • Fuel for lighting • Fuel for cooking • Source of water supply for drinking • Source of water supply for cooking • Presence of household conveniences (assets), ICT devices, vehicles • Type of internet access available • Internet use (where, last three months)	• 5 to 24 years old • School attendance • Place of school • 15 years old and older • Usual Activity/Occupation • Kind of Business or Industry • Class of Worker • Place of Work • Females 15 to 49 years old • No. of children born alive • No. of children still living (among born alive) • No. of children born alive last year • Age at first marriage

Table 5. Data Items in the CPH 2020 Barangay Schedule

ID variables	Barangay characteristics	
• Region • Province/Highly Urbanized City • City/Municipality • Barangay	• Former poblacion of city/municipality • [Current] poblacion of city/municipality • Has Street Pattern • Accessible to National Highway, distance • Town/City Hall or Provincial Capitol, distance • Church, Chapel or Mosque, distance • Public Plaza or Park for Recreation, distance • Cemetery, distance • Marketplace, distance • Elementary School, distance • High School, distance • College/University, distance • Library, distance • Hospital, distance • …Barangay Health Center, distance • Fire Station…, distance • Seaport in Operation, distance • Community Waterworks System, distance • Post Office or Postal Service, distance	• Landline Telephone System, distance • Cellular Phone Signal • Public Street Sweeper • By industry: presence of establishments, establishments with >100 employees, establishments with 10 to 99 employees number of by size, and • Households along Estero • …along Riverbanks/Shoreline • …along Railroad • …in Garbage Dumpsite • …under the Bridge • …along Sidewalk or Easement of Roads and Highways • …in other danger areas… • …in Government Land • …in Private Land which they do not own • …Temporary Relocation Area • …Permanent Relocation/ Resettlement Area • In-movers and out-movers per reason