Census Coverage Measurement Initial Housing Unit Matching Activities

Diane Cronkite
U.S. Census Bureau, Washington, DC 20233¹

Abstract

The Initial Housing Unit Matching and Followup operation of the Census Coverage Measurement (CCM) Program consists of the following activities: computer matching, preprocessing, Before Followup Clerical Matching, Initial Housing Unit Followup, and After Followup Clerical Matching. In order to conduct Before Followup Clerical Matching and After Followup Clerical Matching, trained technicians and analysts use the Housing Unit Matching, Review, and Coding System (HUMaRCS) and the HUMaRCS Map Viewing System (MVS). This report summarizes each of the Initial Housing Unit Matching activities, including use of computer matching, the quality control program, and a high-level overview of the clerical matching process.

Key Words: Census Coverage Measurement, CCM, address matching, clerical matching, computer matching

I. Introduction

The Census Coverage Measurement (CCM) survey is an area sample survey of approximately 170,000 housing units² in the United States excluding remote Alaska and group quarters³ and approximately 7,500 housing units in Puerto Rico and of the people who live in those housing units. The major goal of the CCM survey is to provide estimates of the 2010 Census coverage error for the housing unit population and for housing units.

CCM does not include group quarters because our procedures for measuring coverage of persons in housing units are not compatible with the transient nature of much of the group quarters population. Many of the people who stay at group quarters are normally transient meaning they could be there one day, but not the next, making it hard to contact for followup. There is a group quarters research project this decade to look at the possibility of including group quarters as part of the 2020 CCM.

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily those of the U.S. Census Bureau.

¹ Disclaimer

A housing unit is a house, townhouse, mobile home or trailer, apartment, group of rooms, or single room that is occupied as a separate living quarters or, if vacant, is intended for occupancy as separate living quarters. If occupied by someone who has no other usual residence, a camper, tent, boat, van, motel room, or any other similar unit is classified as a housing unit.

A group quarters is a living quarters normally owned or managed by an entity or organization providing housing or services for the residents. People living in group quarters are generally not related to each other. Group quarters include institutions such as prisons and nursing homes, and other places such as college dormitories, monasteries, and convents.

The CCM is independent from the Census. Independence is required to preserve the validity of the coverage estimates. CCM remains independent by employing a different field staff and using different materials than the Census.

The five major components of the CCM survey are:

- Independent Listing (August 2009 December 2009)
- Initial Housing Unit Matching and Followup (January 2010 May 2010)
- Person Interview (August 2010 October 2010)
- Person Matching and Followup (October 2010 April 2011)
- Final Housing Unit Matching and Followup (March 2011 July 2011)

There are two housing unit matching and followup operations. They are very similar, but the final operation allows the CCM to process changes to the Census since CCM received the Census' initial address list.

This paper details the CCM Initial Housing Unit Matching and Followup operation's activities. Those activities are:

- Computer Matching
- Preprocessing
- Before Followup Clerical Matching
- Initial Housing Unit Followup
- After Followup Clerical Matching

The purpose of the Initial Housing Unit Clerical Matching and Followup operation is to collect data for housing unit estimation and to produce a Preliminary Enhanced List (PEL). The PEL is used as the sampling frame for the Person Interview operation of the CCM. It consists of all confirmed and potential housing units that may exist in the sample block cluster or, in some cases, the surrounding blocks that are eligible to receive an interview during the CCM Person Interview. Potential housing units are those housing units from the followup interview without enough information to confirm their existence as housing units or their location relative to the sample block cluster. The PEL contains CCM address and Census addresses. Those Census addresses are the addresses that are not matched to a CCM address. The PEL is produced after the Initial Housing Unit After Followup Clerical Matching activity. Once the PEL is created, there is a subsampling operation to create the Enhanced List or the final list of address for the CCM Person Interview. CCM addresses are eligible to be selected for the P sample and Census addresses are eligible to be selected for the E sample. The P sample refers to either the housing units in a geographically representative sample, or the persons that reside in those housing units, listed or rostered independently from the census by a post-enumeration survey. The E sample is a sample of data-defined census enumerations. For CCM, the E sample is composed of housing units and persons in the same sample block clusters as the P sample.

A block cluster is a single collection block or a group of contiguous collection blocks. It is used as the primary sampling and work unit for the CCM. A surrounding block is a Census collection block that is geographically contiguous to at least one block in a CCM block cluster or that is contained completely within the contiguous blocks.

Before describing the Initial Housing Unit Matching and Followup Operation, there needs to be a small discussion on the Independent Listing Operation. CCM Independent Listing is a paper based field operation, conducted separately from the Census, in order to compile a list of all housing units and potential housing units in each CCM sample block cluster. Each housing unit

listed also gets a map spot placed on a paper map. Potential housing units are housing units that may exist in the future. For example, houses that are under construction or a vacant mobile home site. CCM lists these units to project to Census Day and what will be there for the CCM Person Interview. CCM uses paper listing books and paper maps because they are low risk and have known properties from earlier decades. Both the listing books and the paper maps are sent to the National Processing Center in Jeffersonville, IN. The listing books are keyed and the paper maps are scanned. This is done to get the data ready for later automated operations.

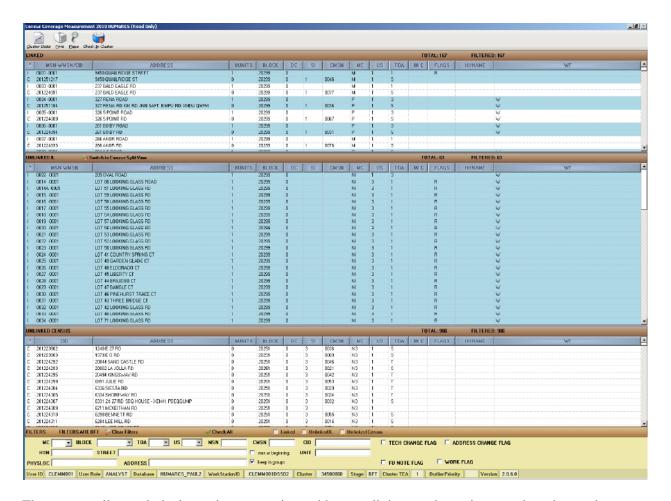
II. Computer Matching

Computer matching uses a Census Bureau developed computer matching software. Before addresses are sent to the computer matching process, addresses from the Independent Listing (CCM addresses) and the Universe Control & Management extract file (Census addresses) are standardized. Data standardization is the process of ensuring that all codes and abbreviations are consistent between address fields. CCM standardizes: capitalization, abbreviations (for example: Street to ST), compresses spaces, removes or converts special characters (for example: convert ñ (tilde n) to n), recodes apartments (for example: A, B, C to 1, 2, 3, etc), and parses data to identify address components (for example: parse "Main St" to "Main" and "St").

The software matches the standardized CCM addresses to the standardized Census addresses (including housing units and group quarters). There are three possible outcomes: matches, possible matches, or nonmatches. When the CCM address matches or possibly matches a Census address, the computer matching software links the two addresses and assigns a match code. Addresses from either file that are nonmatches or unlinked are assigned a match code in the next step – Preprocessing. Computer matching also allows for multiple links to be assigned to identify duplicate addresses. CCM addresses are matched to Census addresses in the sample block cluster and the blocks surrounding sample block cluster.

III. Housing Unit Matching, Review and Coding System

The Housing Unit Matching, Review and Coding System (HUMaRCS) is the computer system used for Initial Housing Unit Preprocessing, Before Followup Clerical Matching, producing forms for Field Followup, and After Followup Clerical Matching. HUMARCS is a GUI interface, designed for 300 users, a client server application that uses Microsoft .NET application, Oracle 10g database, and other third-party development tools, and developed in-house at Census in conjunction with Gunnison, Inc. Below is a screen shot of HUMaRCS. This is NOT real data.



The system allows clerical matchers to review addresses, link records, assign match codes, and code duplicate records. There is a three window display of records. The first window is Linked -- CCM and Census matches and possible matches. The second window is of Unlinked IL – nonmatched CCM addresses. The third window is Unlinked Census – nonmatched Census addresses. The third window contains all nonmatched Census address in the sample block cluster and those blocks that touch the sample block cluster (surrounding blocks). The fields are aligned for ease of reading and reviewing. HUMaRCS allows for filtering. The clerical matcher can filter on fields such as match codes, address fields, and map spot numbers.

At the top of the HUMaRCS screen there is a Maps icon used to launch the map viewer. The map viewer displays both CCM and Census maps side by side. Clerical matchers use this data in the matching process.

HUMaRCS also generates reports for monitoring operations and technicians' work.

IV. Preprocessing

During preprocessing, the results of computer matching are loaded in HUMaRCS for use in clerical matching. The CCM and Census addresses are processed and compared within the sample block cluster and the blocks surrounding the sample block cluster. HUMaRCS will assign additional match codes besides the match and possible match codes from computer matching. Those additional match codes are:

- The CCM address is a nonmatch of a Census address.
- The CCM address is a possible duplicate of another CCM address (within block cluster only).
- The Census address is a nonmatch of a CCM address.
- The Census address is a possible duplicate of another Census address
- The Census address has insufficient information for matching and followup; there is no map spot number or unique address information available.
- The Census address is a group quarter within the block cluster and is a nonmatch to a CCM address.
- The Census address is a housing unit within the surrounding blocks and is a nonmatch to a CCM address.
- The Census address is a group quarter within the surrounding blocks and is a nonmatch to a CCM address.

V. Before Followup Clerical Matching

Both the Before Followup (BFU) Clerical Matching and After Followup (AFU) Clerical Matching activities are preformed at the National Processing Center in Jeffersonville, IN. There are two levels of matchers. They are technicians (staff of 80) and analysts (staff of 15). The technicians and analysts use the HUMaRCS software to review addresses from preprocessing. Technicians are less experienced users, while analysts are highly-trained staff with more experience.

The purpose of BFU is to:

- Identify those CCM Addresses that do not match Census Addresses.
- Identify those Census Addresses that do not match CCM Addresses.
- Identify addresses in CCM or Census that are duplicated.

Technicians will review the preprocessing results and assign match codes to CCM and Census addresses. They do this review map spot by map spot. Technicians will:

- Review computer matches. There may be computer matching errors. If the technician finds an error, the match will be changed to possible match or a nonmatch. The review of the matches might also help when assigning other match codes.
- Review of possible matches. When reviewing the computer linked possible matches, technicians determine if the possible match should be a match, nonmatch, or remain a possible match. All remaining possible matches will be reconciled in the field during the Initial Housing Unit Followup activity.
- Review nonmatches. When reviewing the computer nonmatches, technicians attempt to locate a match, possible match, or duplicate addresses within the block or the surrounding blocks. All remaining nonmatches are sent to the field during the Initial Housing Unit Followup activity to verify the addresses do refer to a housing unit.
- Perform duplicate searches. Technicians search for duplicate addresses for both CCM and Census addresses to identify possible duplicates on both lists. All possible duplicates will be sent to the field during the Initial Housing Unit Followup activity to verify whether or not the two addresses do refer to the same housing unit.

Technicians assign following match codes during BFU. They are:

- The CCM and Census address match.
- The CCM and Census address possibly match. There is not enough information to assign a match with confidence.

- The CCM address is a nonmatch to a Census address.
- The CCM address is a possible duplicate of another CCM address.
- The Census address is a nonmatch to a CCM address.
- The Census address is a possible duplicate of another Census address.
- The Census address is a group quarter within the block cluster and is a nonmatch to a CCM address.
- The Census address is a housing unit within the surrounding blocks and is a nonmatch to a CCM address.
- The Census address is a group quarter within the surrounding blocks and is a nonmatch to a CCM address.

CCM has a very low tolerance for matching error. The level of matching error must be negligible, i.e., small enough not to affect the CCM estimates. Because of this low tolerance, technicians' work will go through quality control (QC). Analysts perform QC on technicians' work. Analysts are highly-trained clerical matchers. Technicians are assigned to either 100 percent or in sample QC. When a technician is in 100 percent QC, the analysts will review all of the technician's work. When a technician is in sample QC, the analysts will review part of the technician's work. The QC logic is built into HUMaRCS. Based on a technician's performance, a technician can flow in and out of 100 percent QC. Meaning if the technician's work meets certain criteria the technician can leave 100 percent and go into sample QC. For QC, analysts review an entire randomly selected cluster to make sure technicians performed their work correctly. They will also review clusters that are not selected for a full QC, but contain must-do conditions. Those must-do conditions are those difficult situations in matching that an analyst should always review. Analysts will also review cases referred by technicians. The results are entered in the HUMaRCS.

VI. Initial Housing Unit Followup

After the Before Followup Clerical Matching activity, certain addresses such as nonmatches, possible duplicates, possible matches, and select matches determined to need additional information are sent to field for the Initial Housing Unit Followup activity. This includes any CCM addresses that were linked to a Census group quarter address. Initial Housing Unit Followup is a field operation. HUMaRCS generates a paper questionnaire with a series of questions for the field interviewer to ask that will confirm the existence or nonexistence of housing units in the sample block cluster or the blocks that touch the sample block cluster at the time of the followup.

The paper questionnaire verifies:

- Is the address a housing unit?
- Is the housing unit in the correct block?
- If there are duplicate housing units, are they the same housing unit?
- If Census listed a group quarter and it matches a CCM housing unit, is it a group quarter, housing unit, or two different units?
- Find additional duplicates missed in Before Followup Clerical Matching (This happens because seeing what's happening on the ground is often easier than looking at addresses and maps on a computer monitor).

The field interviewer can also update paper maps during the field followup.

VII. After Followup Clerical Matching

During After Followup (AFU) Clerical Matching, the technicians and analysts:

- Use HUMaRCS
- Review and interpret the results of the field followup interviews
- Assign match codes
- Update CCM and Census address data
- Update CCM and Census maps
 - Any updates are used in later CCM operations, but no changes are made to actual Census data.

Analysts will perform quality control on the technicians' work. The QC process is the same as in Before Followup Clerical Matching. See section V for details.

Technicians assign the following match codes during AFU:

- Matches between CCM and Census addresses can fall into different categories:
 - CCM and Census addresses that are matched and the units are located in the sample block cluster. This would include housing units that are boarded up.
 - CCM and Census addresses that are matched and the units are located in the blocks that touch the sample block cluster. The CCM housing unit is a geocoding error and the Census unit is a correct enumeration.
 - CCM and Census addresses that are matched and the units are located outside the blocks that touch the sample block cluster. Both the CCM housing unit and the Census housing unit are geocoding errors.
 - CCM and Census addresses that are matched and there is not enough information on the followup form to confirm this unit as a housing unit or to confirm the geocode⁴ with certainty. The followup interview was not conducted, was incomplete, was never received in the National Processing Center (NPC), had contradictory information, or was a noninterview. The status or location of the unit is unknown.
 - CCM and Census addresses that are matched and the units are confirmed to be a group quarter.
 - CCM and Census addresses that are matched and do not refer to a housing unit at the time of the followup interview but are correctly geocoded in the sample block cluster. The followup interview confirmed that the addresses refer to a unit that is under construction, future construction, unfit for habitation, demolished or burned down, is an empty lot/site in a trailer park, or the mobile home moved by the time of the followup interview. These addresses still have the potential to be a housing unit in the sample block cluster at the time of person interview and will be retained for the PEL.
 - CCM and Census addresses that are matched and do not refer to a housing unit at the time of the followup interview but are correctly geocoded in the sample block cluster. These addresses do not have the potential to be a housing unit at the time of the person interview and will not be placed on the PEL. For example, the address could be a merged unit or a non-existing housing unit (e.g., shed, barn, commercial property, or storage of non-household goods).

⁴ A geocode (geographic code) is a code or set of codes used to identify a specific geographic entity. For example, the information needed to geocode an address is the state code, county code, and block number.

- Nonmatched CCM addresses can fall into different categories:
 - CCM addresses that existed as a housing unit at the time of the followup interview and are correctly geocoded in the sample block cluster. The housing unit is not found in the Census in the sample block cluster or search area (surrounding blocks). This would include CCM housing units that are boarded up.
 - CCM addresses that did not refer to a housing unit at the time of the followup interview but is correctly geocoded in the sample block cluster. The followup interview confirmed that the address refers to a unit that is under construction, future construction, unfit for habitation, demolished or burned down, is an empty trailer lot/site in a trailer park, or the mobile home moved by the time of the follow-up interview. These addresses still have the potential to be a housing unit in the sample block cluster at the time of person interview and will be retained for the PEL.
 - CCM addresses that did not refer to a housing unit by the time of the followup interview but is correctly geocoded in the sample block cluster. These addresses do not have the potential to be a housing unit at the time of person interview and will not be placed on the PEL. For example, the address could be a merged housing unit or a non-existing housing unit (e.g., shed, barn, commercial property, or storage of non-household goods). They are removed from further processing for the CCM.
 - CCM addresses that are confirmed to be a group quarter. They are removed from further processing for the CCM.
 - CCM addresses that existed as a housing unit at the time of the followup interview, but is incorrectly listed in the sample block cluster. The address is a CCM geocoding error. They are removed from further processing for the CCM.
 - CCM addresses that are a nonmatch and there is not enough information on the followup form to assign a code with certainty. The followup interview was not conducted, was incomplete, was never received in NPC, had contradictory information, or was a noninterview. The unit status or location of the unit is unknown.
- Nonmatched Census addresses can fall into different categories:
 - Census addresses that existed as housing units at the time of the followup interview and are correctly geocoded in the sample block cluster. The housing units are not found in the CCM in the sample block cluster. This would include Census housing units that are boarded up.
 - Census address is erroneously listed on the Universe Control and Management (UC&M) file, because the address is not a housing unit in the sample block cluster or the surrounding blocks at the time of the follow-up interview. For example, the structure is still under construction (final doors, windows, and floors NOT in place), is unfit for habitation (condemned or open to the elements), has been demolished or burned down, the mobile home has been moved, is a business, is used for storage, or the address is nonexistent.
 - Census addresses that are erroneously listed on the UC&M file, because the address is a group quarter at the time of the followup interview. These addresses were originally listed as housing units.
 - Census addresses that existed as housing units at the time of the followup interview and were geocoded to the sample block cluster, but are actually located in the surrounding blocks to the block cluster. These housing units are considered correctly enumerated because it is located in the search area.

- Census addresses that were geocoded to the sample block cluster, but are actually located beyond the surrounding blocks to the sample block cluster. These addresses were erroneously enumerated in this block cluster because of a geocoding error.
- Census addresses that are a nonmatch and there is not enough information on the
 followup form to assign a code with certainty. The followup interview was not
 conducted, was incomplete, was never received in NPC, had contradictory
 information, or was a noninterview. The status of these addresses or their locations
 are unknown.
- Census addresses that are nonmatched group quarters in the sample block cluster. These addresses were originally listed as Census group quarters.
- Census addresses that are nonmatched and are located in the blocks that touch the sample block cluster and are defined to be part of the search area.
- Census addresses that are a nonmatch and are group quarters located in the blocks that touch the sample block cluster. They are geocoded to a block that touches the sample block cluster.
- Census housing units that have insufficient information for matching and followup: there is no map spot number or unique address information available (e.g., "white house" when there are 6 other "white houses" listed in the same area).
- Duplicate CCM addresses can fall into different categories:
 - CCM addresses that should not have been listed in the CCM. These addresses are a duplicate of other CCM addresses in the sample block cluster. These addresses are removed from further processing for the CCM.
 - CCM addresses that are a duplicate identified in After Followup Clerical Matching but not field-confirmed by Initial Housing Unit Followup. They will be sent to Final Housing Unit Followup for a field interview. They will remain on the PEL and are eligible for a CCM Person Interview. These addresses are either in a cluster that skipped Before Followup Clerical Matching or were an additional CCM duplicates not found in Before Followup Clerical Matching.
- Duplicate Census addresses can fall into different categories:
 - Census addresses that were erroneously enumerated in the Census. These addresses are duplicates of other Census addresses.
 - Census housing units that are a duplicate identified in After Followup Clerical Matching but not field-confirmed by Initial Housing Unit Followup. They will be sent to Final Housing Unit Followup for a field interview if the duplicate or the original address is in the E sample. These addresses are either in a cluster that skipped Before Followup Clerical Matching or were additional duplicates not found in Before Followup Clerical Matching.

VIII. Results of Initial Housing Unit Operations

The Results of the Initial Housing Unit Operations are used to create the sampling frame for the CCM Person Interview. The addresses included in the frame are those addresses that were a match, those CCM and Census addresses that were nonmatches, and those addresses that are potential housing units at the time of the CCM Person Interview.

The data from the Initial Housing Unit operations, along with data collected in the Census and in subsequent CCM operations will be used to produce estimates of coverage error for the 2010 Census.

References

Manual # 2010-D6-02 2010 Census Coverage Measurement Initial Housing Unit Before Followup Clerical Matching Reference Manual

Manual # 2010-D6-06 2010 Census Coverage Measurement Initial Housing Unit After Followup Clerical Matching Reference Manual

DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-B-12R1 – Codes Used In 2010 Census Coverage Measurement Initial Housing Unit Matching

DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-B-10 – 2010 Census Coverage Measurement Glossary