Environment: Data Guide

The purpose of this document is to explain to users the structure of the environmental data curated by Core D. These data are organized in long format, and individual study participants may have multiple rows, each corresponding to a unique address, calendar time, and/or entry type. The following fields will assist users with merging and working with DAP Environment data:

Contents

Record Identification & Metadata Variables

Record Identification Variables

- **dog_id**
  - The dog_id field represents the unique dog identifier, consistent with the dog_id in HLES and CSLB data frames.
- **address_1_or_2**
  - 1 = Primary Address
    - Every respondent must have a primary address corresponding to the location where the dog spends the majority of its time. This field will have a "1" signifying that the geocoded information is for the primary address.
  - 2 = Secondary Address
    - Approximately 10% of respondents report having a secondary address where their dog spends time. Participants reporting secondary addresses will have an additional row, and these rows will have a "2" entered in this field. All columns will correspond with the geocoded secondary address.[1]
- **address_month**
  - This field indicates the month of the entry event. For baseline entries (gm_entry_type = 1), this is the month in which the respondent joined the DAP pack.
- **address_year**
  - This field indicates the year of the entry event. For baseline entries (gm_entry_type = 1), this is the calendar year in which the respondent joined the DAP pack

[1] The proportion of time that the dog spends in its primary vs. secondary residence can be found in the HLES owner contact instrument (variable name "oc_secondary_residence_time_percentage")

Geocoding Metadata Variables

The purpose of the geocoding metadata variables (prefix "gm") is to provide information pertaining to the underlying geocoding process. The following metadata variables are provided:

- **gm_address_type**
  - A value will only appear in this field for addresses which were geocoded automatically through esri's Business Analyst locator ("Level 1"/option 1 in "gm_geocoder"). This field specifies which of the two acceptable types of matches were assigned by the geocoder for this address:
    - 1 = StreetAddress
      - From ArcGIS:
        - "A street address that differs from PointAddress because the house number is interpolated from a range of numbers. Reference data contains street center lines with house number ranges, along with administrative divisions and optional postal code information, for example, 647 Haight St, San Francisco, CA, 94117.
    - 2 = PointAddress
      - From ArcGIS:
        - "A street address based on points that represent house and building locations. Typically, this is the most spatially accurate match level. Reference data contains address points with associated house numbers and street names, along with administrative divisions and optional postal code. The X / Y and geometry output values for a PointAddress match represent the street entry location for the address; this is the location used for routing operations. The DisplayX and DisplayY values represent the rooftop, or actual, location of the address. Example: 380 New York St, Redlands, CA, 92373."
- **gm_match_type**
  - A value will only appear in this field for addresses which were geocoded automatically through esri's Business Analyst locator ("Level 1"/option 1 in "gm_geocoder") or manually matched ("Level 3"/option 3 in "gm_geocoder") in ArcMap. Three options are possible in this field:
    - 1 = "A"
      - Means that the coordinates were automatically found at one of the two precise levels (see "gm_addr_type") using esri Business Analyst geocoder
    - 2 = "PP"

        - Means that the coordinates were manually placed following a search for this address by the Core D RA
    - 3 = "M"
        - Means that the Business Analyst initially returned a "tie" for this address, and that these tied coordinates were resolved based on research and review by the Core D RA
- **gm_state_fips**
    - Two-digit state FIPS code corresponding to geocoded coordinates, attained via spatial join with TIGER US census tract boundary shapefiles corresponding to the 2014-2018 American Community Survey (accessed via IPUMS NHGIS[2]).
- **gm_geocoder**
    - Indicates which geocoding method/level returned the coordinates for this address (See geocoding documentation appendix below for more detailed information on this process). Four possible options:
        - 1 = esri
            - Matched automatically at "Level 1": esri Business Analyst geocoder
        - 2 = SmartyStreets
            - Matched at "Level 2": SmartyStreets, returning Zip9 precision
        - 3 = manual
            - Matched at "Level 3": manual revision by Core D RA
        - 4 = CannotLocate
            - Unable to locate this address, even after manual review
- **gm_entry_type**
    - This field indicates the entry event type corresponding with each row. The possible values are:
        - 1 = Baseline entry
            - Initial gecoding attempt based on respondent address provided in the HLES owner contact survey
        - 2 = Owner Profile Update
            - A value of 2 reflects that the geocoding attempt is based on an updated residential address provided in the "Update My Profile" feature of the Dog Aging Project participant portal. Users can update their profile at any time.
        - 3 = Annual follow-up
            - A value of 3 reflects a geocoding attempt based off new residential address information provided in the annual HLES follow-up survey

[2] Steven Manson, Jonathan Schroeder, David Van Riper, and Steven Ruggles. IPUMS National Historical Geographic Information System: Version 14.0 [Database]. Minneapolis, MN: IPUMS. 2019. http://doi.org/10.18128/D050.V14.0

- 4 = Secondary Data Update
  - A value of 4 reflects a secondary data update entry. For the 2020 secondary data update, these entries always correspond with a month of 12 and year of 2020. Rows with gm_entry_type = 4 take the most up-to-date address for each respondent and link this address to updated secondary data, where applicable (ie, new American Community Survey data).
- 5 = Secondary Data Update + Owner Profile Update
  - A value of 5 reflects incidents when a respondent has provided an updated address in their owner profile portal, *and* this updated address is simultaneously linked to updated secondary data
- 6 = Secondary data update + Annual follow-up
  - Secondary data update and new annual follow-up address provided occurs in the same month-year
- 7 = Manually corrected address[3]
  - A value of 7 reflects a manually corrected address. In these cases, a respondent provides a non-geocodable address in their initial HLES Owner Contact form. We reach out to respondents for clarification, and if they provide corrections, we geocode this corrected address.

- **gm_complete**
  - Indicates whether this record's geocoding processing is complete. All addresses in the Curated Data Release will be marked "complete" (2) except unmatchable addresses (corresponding to *gm_geocoder* = 4), which will be marked "unverified" (1) for this field.
    - 0 = incomplete
    - 1 = unverified
    - 2 = complete

[3] **NOTE**: this procedure was discontinued in October 2020

Contextual Variables

Successfully geocoded respondent addresses are linked to pre-existing environmental indicators. The following four sections describe the contextual variables linked to geocoded respondent addresses.

Census Tract-level Sociodemographic and Economic Indicators

Socioeconomic and demographic information come from the US Census, and are provided at the **census tract** level. We include descriptive tract information; race/ethnicity variables; tract gender breakdown; socioeconomic status (SES) variables; and neighborhood stability variables. Each census variable provided contains the prefix "cv".

Descriptive tract variables:

- **cv_population_estimate**
  - estimated tract population from ACS
- **cv_area_sqmi**
  - square mileage of tract
- **cv_population_density**
  - persons per square mile (cv_population_estimate divided by cv_area_sqmi)
- **cv_data_year**
  - Indicates ACS release year. Current options:
    - 1, 2018 5-year ACS
    - 2, 2019 5-year ACS

Race/Ethnicity Variables: All variables are converted from counts to percent based on estimated tract population (cv_population_estimate)

- **cv_pct_nothispanic_white**
  - Percent Not Hispanic, White Alone
- **cv_pct_nothispanic_black**
  - Percent Not Hispanic, Black Alone
- **cv_pct_nothispanica_ian**
  - Percent Not Hispanic, American Indian and Alaska Native Alone
- **cv_pct_nothispanic_asian**
  - Percent Not Hispanic, Asian Alone
- **cv_pct_nothispanicn_hpi**
  - Percent Not Hispanic, Native Hawaiian and Other Pacific Islander Alone
- **cv_pct_nothispanic_other**
  - Percent Not Hispanic, Some Other Race Alone

- **cv_pct_nothispanic_two_or_more**
  - Percent Not Hispanic, Two or More Races
- **cv_pct_hispanic**
  - Percent Hispanic or Latino (not broken down by race)

Gender: Variable is and converted from count to percent based on estimated tract population

- **cv_pct_female**
  - Percent Female

Socioeconomic/SES Variables:

- **cv_median_income**
  - Median Household Income in the Past 12 Months (In Inflation-Adjusted Dollars for current data year)
- **cv_gini_index**
  - Gini Index of Income Inequality
- **cv_pct_below_125povline**
  - Percentage of tract population below 1.25% of the poverty line
- **cv_pct_jobless16to64mf**
  - Percentage of tract population ages 16-64 unemployed in the labor force AND not in the labor force, males & females combined
- **cv_pct_famsownchild_female_led**
  - Percentage of own children living in households with female householder, no husband present
- **cv_pct_less_than_ba_degree**
  - Percentage of population 25 years or older with less than a Bachelor's Degree
- **cv_pct_less_than_100k**
  - Percentage of Households Earning Under $100,000 in the Past 12 Months (In Inflation-Adjusted Dollars corresponding to that data year)
- **cv_disadvantage_index**
  - Disadvantage Index
  - Calculated by taking z-score of 5 preceding variables and averaging them

Neighborhood Stability:

- **cv_pct_same_house_1yrago**
  - Percent in Same House 1 Year Ago
- **cv_pct_owner_occupied**
  - Percent Owner Occupied Housing Units
- **cv_pct_us_born**
  - Percent Born in United States
- **cv_stability_index**
  - Stability Index
  - Calculated by taking z-score of 3 preceding variables and averaging them

Census Tract-level Air Pollution Indicators

Air pollutant concentration data come from public-use estimates developed by the Center for Air, Climate and Energy Solutions (caces) using v1 empirical models as described by Kim et al. (2018)[4]. The most recent data are from 2015 (available here) and include model-based **census tract** level estimates for six pollutants . From caces:

> "These models provide estimates of outdoor concentrations for six pollutants (four gases: O3, CO, SO2, NO2; two aerosols: PM10, PM2.5) throughout the contiguous U.S. Model estimates are annual-average values for years 1979 – 2015 (SO2, NO2), 1988 – 2015 (PM10), 1990-2015 (CO), 1999-2015 (PM2.5), and the average during May through September of the daily maximum 8-hour moving average for years 1979-2015 (O3). When downloading data, concentrations are listed as the variable "pred_weight"; units are micrograms per cubic meter for PM2.5 and PM10, parts per billion for ozone, SO2, and NO2, and parts per million for carbon monoxide. Data are available at national, state, county, census tract, and census block group levels."

These data are merged to geocoded respondent addresses using their 11-digit tract FIPS code. The following pollutant variables (prefix = "pv") are provided:

- **pv_co**
  - carbon monoxide (ppm)
- **pv_no2**
  - nitrogen dioxide (ppb)
- **pv_o3**
  - ozone (ppb)
- **pv_pm10**
  - particulate matter (Âµg/m3)
- **pv_pm25**
  - particulate matter (Âµg/m3)
- **pv_so2**
  - sulfur dioxide (ppb)
- **pv_complete**
  - Indicates whether this record's pollutant data variables are complete. All addresses should be marked "complete" (2) except AK & HI addresses[5], which will be marked "unverified" (1) for this field. Options are:
    - 0, Incomplete

[4] Kim S.-Y.; Bechle, M.; Hankey, S.; Sheppard, L.; Szpiro, A. A.; Marshall, J. D. 2018. "A parsimonious approach for estimating individual-level concentrations of criteria pollutants over the contiguous U.S." In Preparation.

[5] **NOTE:** Pollutant levels are estimated ONLY for the contiguous US, excluding Alaska & Hawaii. Respondents with addresses in AK and HI will have their pollutant instrument missing and marked "unverified" in RedCap. The field *"pv_complete"* will be given a "1", corresponding to "unverified", and all other geocoded addresses will have a "2" (corresponding to "complete") for this field.

        - 1, Unverified
        - 2, Complete
- **pv_data_year**
    - Indicates the year of data. Currently only one option exists:
        - 1, 2015 caces

County-level Temperature and Precipitation Measures

Temperature and Precipitation variables data come from NOAA's Climate Divisional Database (nClimDiv). These data are provided at the **county** level. NOAA provides county-level climate indicators at two temporal levels. First, annual summaries contain temperature and precipitation averages for each month of each year. Second, "normal summaries" are provided, which are long-term averages over 30-year periods in increments of 10 years. Both annual and normal variables are linked to respondent addresses.

- Global summaries of the month variables: We include monthly summaries for the most recent year available for the following variables:
  - **tp_data_year**
    - Indicates the year that monthly values correspond with based on the calendar year in which the responded lived at a primary or secondary address. Current options:
      - 1, nClimDiv 2019
      - 2, nClimDiv 2020
  - **tp_tmin_annual_XX**
    - Monthly minimum temperature, where "XX" is equal to a two-digit number corresponding to month (ie, 01 = January; 02 = February, etc.)
  - **tp_tmax_annual_XX**
    - Monthly maximum temperature
  - **tp_tmpc_annual_XX**
    - Monthly average temperature
  - **tp_pcpn_annual_XX**
    - Precipitation (inches)
- Normals monthly variables: Monthly long-term averages in 30 year periods in increments of 10 years.
  - **tp_norm_data_year**
    - Indicates the 30-year period that monthly values correspond with. Current options:
      - 1, nClimDiv 1981-2010
  - **tp_tmin_norm_XX**
    - Monthly minimum temperature, where "XX" is equal to a two-digit number corresponding to that month (ie, 11 = November; 12 = December, etc.)
  - **tp_tmax_norm_XX**
    - Monthly maximum temperature
  - **tp_tmpc_norm_XX**
    - Monthly mean temperature
  - **tp_pcpn_norm_XX**
    - Precipitation (inches)

This yields a total of 48 global summaries of the month variables (12 months * 4 variables) as well as 48 normals monthly variables, resulting in 96 total substantive contextual climate variables linked to residential primary and secondary addresses.

NOTE: Not all US regions are included in the NOAA nClimDiv. For such instances, we undergo the following procedures:

- **Hawaii and Washington DC:** NOAA nClimDiv county-level summaries do not include Hawaii counties or Washington, DC . For these regions, we constructed our own climate variables by downloading the individual station-level data from NOAA in each jurisdiction, geocoding them, and averaging out the monthly readings by county. This procedure is repeated for both annual and normals periods.
- **Ad Hoc missing regions in contiguous US:** Some contiguous counties are missing from the NOAA nClimDiv, particularly small counties that are enveloped entirely within larger counties. In these instances, we provide the climate variables for the larger surrounding counties in which these smaller counties are located.

Neighborhood Walkability Indicators

The Walkability variables (prefix = "wv") draw from two separate data sources. First, three variables are provided by Walkscore, a private company that generates a walkability index based on walking routes to nearby amenities and other indicators (see here for more information). We use the walkscoreAPI R package to retrieve Walkscores for each geocoded address based on their geocoded lat/lon coordinates. Second, we include residential density variables from the ACS. We include these in our walkability instrument based on the conclusions drawn by Mooney et al. (2020)[6], who find that residential density is an appropriate proxy measure for walkability. Residential density is calculated at the **census tract** level. The following variables are provided:

- Walkscore Variables:
  - **wv_walkscore**
    - A numeric score from 0-100 indicating walkability
  - **wv_walkscore_descrip**
    - Walkscore's categories for walkability, based on retrieved Walkscore. There are five coded categories:
      - 1 = Walker's Paradise
        - For WalkScores 90-100
      - 2 = Very Walkable
        - For WalkScores 70-89
      - 3 = Somewhat Walkable
        - For WalkScores 50-69
      - 4 = Car-Dependent
        - For WalkScores under 50
      - 5 = NA
        - No WalkScore available for these coordinates
  - **wv_walkscore_date**
    - A date indicating the last time WalkScore updated the calculated score for the data they provide
- Residential Density Variables:
  - **wv_housing_units**
    - Number of housing units in the tract, from American Community Survey

[6] Stephen J. Mooney, Philip M. Hurvitz, Anne Vernez Moudon, Chuan Zhou, Ronit Dalmat, Brian E. Saelens. "Residential neighborhood features associated with objectively measured walking near home: Revisiting walkability using the Automatic Context Measurement Tool (ACMT)." *Health & Place*, Volume 63. doi: https://doi.org/10.1016/j.healthplace.2020.102332.

- **wv_res_density**
  - Housing units per square mile (“wv_housing_units” divided by “cv_areasqmi”)
- **wv_density_data_year**
  - Data source year. Multiple choice. Currently two values exist, more options will be added as each new wave of ACS data becomes available:
    - 1, 2018 5-year ACS
    - 2, 2019 5-year ACS

Geocoding Appendix

The following appendix describes the DAP Core D geocoding process in detail for user reference.

I. Adjudicating Geocoders:

Our first step in the geocoding process was to identify a geocoder which could match respondent addresses with the greatest degree of accuracy. We tested the first several hundred submitted respondent addresses against three popular geocoders: 1) esri's Business Analyst 2019 data, which is similar to their ArcGIS StreetMap Premium project; 2) SmartyStreets, an online API based geocoder; and 3) the "geocode" function in the R package ggplot2, which uses the Google geocoding API.

The first 542 addresses were geocoded against all three services. The esri Business Analyst geocoder returned the most complete results, and was capable of matching 97.8% of the addresses at a high quality (rooftop or street address) level. SmartyStreets was able to match 95.4% of the addresses at a high quality level (Zip9 or higher), and the Google API was able to match 96.5% of the results at some level (accuracy score not available).

A manual review and coordinate distance analysis led us to approach the geocoding procedure in a series of three "levels". The following procedure will be run for primary and (when applicable) secondary residencies where the dog spends time (as indicated in owner contact survey)

- **Level 1 (esri):** First, addresses will be run through the esri Business Analyst geocoder; the "gold standard" based on our pilot analysis. Precisely matched addresses are taken at face value. The ~3% of addresses that cannot be matched at the rooftop or street address level will then proceed to level 2.
    - Addresses matched at level 1 will take the following values for gm_match_type, gm_address_type, and gm_geocoder:
        - Match_type = A if not tied in address locator; Match_type = M if addresses tied & manual match was performed
        - Addr_type = "PointAddress" or "StreetAddress" (the two acceptable "precise" matching scores)
        - Geocoding Method/Level = "esri"
- **Level 2 (SmartyStreets):** Poorly matched addresses (matched at the street name or zip code level in esri) will then be run through the SmartyStreets geocoder. SmartyStreets is preferred because it provides users with an accuracy score. If addresses which have proceeded to level 2 can be geocoded at SmartyStreets' highest accuracy level (Zip9, typically block-level), then we take these coordinates at face value. All others proceed to level 3.
    - Addresses matched at level 2 will take the following values for gm_match_type, gm_ address _type, and gm_geocoder:
        - Match_type/Addr_type = "NA"
        - Geocoding Method/Level = "SmartyStreets"

- **Level 3 (Google maps & manual review):** Our pilot analysis indicated that approximately 1% of addresses proceeded to level 3, and were not able to be precisely matched by either esri or SmartyStreets geocoder. In these instances, we will proceed with a manual review using Google Maps and human research.
    - Addresses matched at level 3 will take the following values for gm_match_type, gm_ address _type, and gm_geocoder:
        - Match_type= "PP"; Addr_type = "NA"
        - Geocoding Method/Level = "manual"
- **Unmatchable Addresses:** In cases addresses *still* cannot be matched (perhaps due to respondent entry errors, rural addresses, or new build development), we will reach out to the respondent and ask them to clarify their location.
    - Unmatchable addresses (both primary and secondary) will take the following values for gm_geocoder & gm_complete:
        - Geocoding Method/Level = "CannotLocate"
        - gm_complete will be marked "Unverified"
    - If respondents provide an updated address after we reach out to them, we will re-run the geocoding procedure for this newly provided address. Such addresses will have a "7" (corresponding to "Manually corrected address") entered in the field "gm_entry_type".
        - **NOTE**: Core D stopped reaching out to respondents to manually correct addresses in October 2020

See metadata documentation above for detailed information on the geocoding metadata variables (prefix = "gm") included.

II. Spatially joining coordinates to census boundaries

Precisely-matched coordinates are then plotted over the TIGER US census tract boundary shapefiles corresponding to the 2014-2018 American Community Survey (accessed via IPUMS NHGIS[7]). Using the spatial join tool in ArcMap, the block group information that each respondent address point falls within is joined to the respondents' study id.

III. Merging to secondary datasets

- *Census Variables (prefix = "cv")*
    - Once each address is assigned a block-group level FIPS code, chosen relevant neighborhood characteristics are merged with each respondents' record using the 11-digit ***tract*** FIPS code as the join key. This creates a matrix of respondent study ID, tract FIPS code, and relevant census variables. For a complete list of census variables selected, see Census Variable documentation above.
- *Pollutant Variables (prefix = "pv")*

[7] Steven Manson, Jonathan Schroeder, David Van Riper, and Steven Ruggles. IPUMS National Historical Geographic Information System: Version 14.0 [Database]. Minneapolis, MN: IPUMS. 2019. http://doi.org/10.18128/D050.V14.0

- Pollutant concentration data come from public-use estimates developed by the Center for Air, Climate and Energy Solutions (caces) using v1 empirical models as described by Kim et al. (2018)[8]. The most recent data are from 2015 and include estimates for six pollutants (four gases: O3, CO, SO2, NO2; two aerosols: PM10, PM2.5) at the **census tract** level. For in-depth description of variables, see pollutant variables documentation above. These data are merged to geocoded respondent addresses using their 11-digit tract FIPS code
    - **NOTE:** Pollutant levels are estimated ONLY for the contiguous US, excluding Alaska & Hawaii. Respondents with addresses in AK and HI will have their pollutant instrument marked "Unverified".

- *Temperature and Precipitation Variables (prefix = "tp")*
    - Temperature and precipitation variables come from the NOAA's Climate Divisional Database (nClimDiv), which makes available four key climate indicators (maximum air temperature, minimum air temperature, average air temperature, and precipitation in inches) for nearly every census **county** in the contiguous US and Alaska, both annually (dating back to 1895) and in long-term 30-year averages in increments of 10 years ("normals"). We include these four monthly variables for the most recent calendar year as well as the four monthly variables for the most recent "normals" period (1981-2010 as of the time of this writing). This results in 96 total climate variables included (12 months * 4 variables * 2 periods). For in-depth description of these variables, see climate variables documentation above.
        - **NOTE:** NOAA nClimDiv county-level summaries do not include Hawaii or Washington, DC counties. For these counties, we construct our own climate variables by downloading the individual station-level data from NOAA in each jurisdiction, geocoding them, and averaging out the monthly readings by county. This procedure is done for both annual and normals periods.
        - **NOTE:** Some contiguous counties are missing from the NOAA nClimDiv, particularly small counties which are enveloped entirely within larger counties. In these instances, we take the climate indicators for the larger surrounding counties in which these smaller counties are located.
- *Walkability Variables (prefix = "wv")*
    - Walkability variables come from two sources: Walkscore and ACS. Walkscore variables include the proprietary "Walkscore" (an integer between 0-100), as well as the corresponding walkability category according to Walkscore. The ACS walkability indicator included is residential density (estimated number housing units in each **tract** divided by **tract** square mile). See neighborhood walkability documentation above for more details.

[8] Kim S.-Y.; Bechle, M.; Hankey, S.; Sheppard, L.; Szpiro, A. A.; Marshall, J. D. 2018. "A parsimonious approach for estimating individual-level concentrations of criteria pollutants over the contiguous U.S." In Preparation.

IV. Residential and Extra-local changes

New DAP pack members' residential addresses will be geocoded twice monthly (on the first and third weeks of each month). Baseline entries will have the "gm_entry_type" = 1 in their *gm_entry_type* field, corresponding with baseline entry.
Given that DAP respondents may individually change residencies, and that environmental conditions are likely to change, we must establish a procedure for updating respondent addresses and secondary datasets as they become available. The Environment Data is structured as a longitudinal project to accommodate these changes. All respondents have a non-repeating "baseline" instrument which includes their study id and DAP pack date. Geocoding metadata and linked environmental variables are then pushed into the month-year event which corresponds to their DAP pack date. Residential and extra-local changes are then treated as follows:

- *Individual Moves (monthly):*
    - "Ad-Hoc" revision: Once a month, on the first of the month, Core D RA will check if any DAP participant user profiles have been updated to reflect a change in primary or secondary residences. Changed addresses will be geocoded, linked to secondary data, and imported as they occur. The *address_month* and *address_year* fields will correspond with the month and year that the user updated their profiles. Ad hoc updates will have the "Entry Type" field = "Owner profile update" in their *gm_entry_type* field.
    - Annual revision: In addition, all respondents will be asked to confirm their information once a year via annual check-in surveys. After these annual check-ins have been completed, we will update any newly changed addresses, geocode them, and link to corresponding environmental indicators. Annual updates will have the "Entry Type" = "Annual Follow-Up" in their *gm_entry_type* field.
- *Data updates (annually):*
    - In January of each year, we will update all geocoded addresses with newly available secondary data. New data will be linked to the most up-to-date respondent address and imported into the December month-year event for the preceding year (ie, December 2020 for the 2021 annual update). Source data year will be indicated in the "data year" fields corresponding to each secondary data category. Data updates will have the "Entry Type" field = "Secondary data update" in their *gm_entry_type* field. If the data update occurs in the same month-year that an owner profile update OR an annual revision occurs, these will be indicated by "Entry Type" = "Secondary data update + Owner profile update" or "Entry Type" = "Secondary data update + Annual Follow-Up" respectively in their gm_*entry*_type field