SafeTab-H v3.0.0 Documentation

Executive Summary

The SafeTab-H algorithm produces differentially private statistics (counts) of households broken down by several dimensions:

- Type or tenure.
- Detailed race, ethnicity, and American Indian and Alaska Native (AIAN) tribe and village groups.
- Geographies, including nation, state, county, tract, place, and American Indian, Alaska Native, and Native Hawaiian (AIANNH) areas.

The data product derived from the output of SafeTab-H is known as the Detailed Demographic and Housing Characteristics File B (Detailed DHC-B).

Goals

1. **Produce tables of statistics (counts)**.
2. **Satisfy differential privacy:** the algorithm used to produce the tables satisfies zero-concentrated differential privacy and also supports pure differential privacy.
3. **Low error:** the algorithm should allow users to tune input parameters to improve the error in statistics.

Problem Specification

Demographic Statistics: The Census Bureau would like to release the following statistics for each population group as part of the Detailed DHC-B:

- T3: Household Type by Detailed Race/Ethnicity of Householder
    - T03001. Household Type (Universe)
    - T03002. Household Type (2 Categories)
    - T03003. Household Type (6 Categories)
    - T03004. Household Type (8 Categories)
- T4: Tenure by Detailed Race/Ethnicity of Householder
    - T04001. Tenure (Universe)
    - T04002. Tenure (3 Categories)

Each of the tabulations (i.e., T3, T4) must be released for a set of predefined population groups. The set of population groups (and tabulations released for each) are enumerated below.

Geographies: Statistics are released for population groups at the following geographic levels:

- USA: corresponds to national level counts where the nation is composed of the 50 states + DC
- State / PR-State: corresponds to populations groups at the state level for the 50 states + DC and Puerto Rico
- County / PR-County: corresponds to counties or county equivalents within the 50 states + DC or in Puerto Rico
- Tract / PR-Tract: corresponds to census tracts within the 50 states + DC or in Puerto Rico
- Place / PR-Place: corresponds to Census Bureau places within the 50 states + DC or in Puerto Rico
- AIANNH areas: correspond to the following areas -
    - 0001 - 4999: Federally recognized American Indian Reservations and Off-Reservation Trust Lands
    - 5000 - 5499: Hawaiian Home Lands
    - 5500 - 5599: Oklahoma Tribal Statistical Areas (OTSAs)
    - 6000 - 7999: Alaska Native Village Statistical Areas
    - 8000 - 8999: Tribal Designated Statistical Areas
    - 9000 - 9499: State recognized American Indian Reservations
    - 9500 - 9998: State Designated Tribal Statistical Areas

Note: SafeTab-H will not tabulate statistics for AIANNH = 9999 (which indicates “not in an AIANNH area”) nor for PLACE = 99999 (not in a Place).

Race and Ethnicity Characteristic Iterations: Statistics must be released for race and ethnicity characteristic iterations for both “*Alone*” as well as “*Alone or in any Combination*” *(AOIAC)*. The list of characteristic iterations is partitioned into “`Common`,” “`Detailed`,” and “`Other`” lists described below:

Race and Ethnicity “`Common`” Characteristic Iterations:

- These are designated in the input file ‘race-characteristic-iterations.txt’ with DETAILED_ONLY = False and COARSE_ONLY = False (see Appendix A)

Race and Ethnicity “`Other`” Characteristic Iterations:

- A list of characteristic iterations that capture “`Other`” detailed and intermediate race and ethnicity characteristic iterations that are not captured in the “`Common`” list
- These are designated in the input file ‘race-characteristic-iterations.txt’ with DETAILED_ONLY = False and COARSE_ONLY = True (see Appendix A)

Race and Ethnicity “`Detailed`” Characteristic Iterations:

- Additional detailed race and ethnicity characteristic iterations

- For instance, these include American Indian and Alaska Native groups that meet a population threshold of 100 or more nationally in the 2010 CPH-T-6 but are not in the "Common" list
- These are designated in the input file 'race-characteristic-iterations.txt' with DETAILED_ONLY = True and COARSE_ONLY = False (see Appendix A)

Depending on the geographic level and which list an iteration comes from, the statistics that need to be released are different as shown in the table below. T3,T4 indicates that SafeTab-H will tabulate both tables T3 *and* T4 at some level of granularity. See the description of the algorithm in the "Approach" section below for more details.

		Characteristic Iterations		
,		Common	Detailed	Other
Geography Levels	USA	T3,T4	T3, T4	N/A*
	STATE			
	COUNTY	T3,T4	N/A**	T3,T4
	TRACT			
	PLACE			
	AIANNH			

* No population T1 counts are produced for Other at USA and STATE levels so SafeTab-H will not produce T3 and T4 counts at those levels.

** No population T1 counts are produced for Detailed at sub-state levels, so SafeTab-H will not produce T3 and T4 counts at those levels.

Consistency Requirements:

- The statistics released will be integral.
- SafeTab-H inherits suppression from SafeTab-P. That is, if a population group does not receive a T1 population count, it will not be output as part of SafeTab-H.

Selected list of potential data inconsistencies in the SafeTab-H output:

- Statistics may contain negative values.
- Levels may not add up.
    - A state's statistics may not match the sum of its corresponding counties

    - A level 1 iteration code's statistics may not match the sum of its corresponding level 2 iteration codes.
- Alone or in any combination iteration code statistics may be less than the corresponding alone iteration code statistics.
- Apart from the suppression inherited from the T1 counts, no consistency is enforced between SafeTab-P and SafeTab-H. For example, it may be possible to see a positive population count in Detailed DHC-A in a geographic area with negative housing units in Detailed DHC-B.

Approach

Privacy loss is measured with respect to adding/removing one household. That is, the privacy-loss parameter (epsilon / rho) quantifies the privacy risks to household units rather than to individuals. Counts in T3 and T4 are linear over a household records dataframe. The Census categorizes the T3 and T4 tables as counts over the *Household Universe.*

Tumult Labs has developed **SafeTab-H** for the household universe (T3 and T4 of Detailed DHC-B). SafeTab-H uses a reader that constructs dataframes from both Census Edited File (CEF)-Person and CEF-unit files with one row for every household and attributes describing household type, tenure, and the householder race and ethnicity. SafeTab-H will also ingest noisy T1 population counts produced by SafeTab-P. SafeTab-H will output statistics about household type and tenure by detailed race of householder for every eligible population group as determined by the algorithm.

The SafeTab-H algorithm will utilize an adaptive procedure as follows:

- Load noisy T1 counts
- For each of the populations groups:
    - Compare the group's T1 count against predetermined thresholds to determine the granularity of T3 and T4 statistics.
    - Compute noisy estimates at the selected granularity
    - Release the estimates as well as less-granular marginals produced by summing over the noisy estimates (as applicable)

For example, population group A which has a small count below the T4 threshold will be tabulated in output table T04001 whereas population group B which has a larger T1 count that exceeds the T4 threshold will be tabulated in table T04002. Population group C, whose T1 count was suppressed, receives no T4 statistics.

For reference,
T04001

TENURE
Universe: Occupied housing units
Total:

T04002

TENURE	
Universe: Occupied housing units	
Total:	
	Owned with a mortgage or a loan
	Owned free and clear
	Renter occupied

SafeTab-H can be instantiated to satisfy “pure”-differential privacy, or zero concentrated differential privacy (zCDP) with a single global privacy loss budget.

SafeTab-H does not support a “non-private” mode.

Assumptions

SafeTab-H makes certain assumptions about the noisy T1 population counts it ingests:

- The T1 counts have been postprocessed by SafeTab-P, including suppression.
- The T1 counts have been through coterminous geography postprocessing.
- The T1 counts have **not** been through any other postprocessing.
- The T1 counts were produced by a SafeTab-P run that used the same GRF-C, race/ethnicity codes, and race/ethnicity iterations as the SafeTab-H run.
- The T1 counts contain data for all runs being attempted (50 states+DC and/or Puerto Rico).

If any of these assumptions is violated, SafeTab-H may not produce correct results.

Performance

We recommend running SafeTab-H on an EMR cluster with 1 primary and 2 core nodes (each an r4.16xlarge instance) and with the spark settings specified in `resources/spark_configs/spark_cluster_properties.conf`. Observed run times, using these recommended settings, for SafeTab-H's validate and execute commands are given below.

Validate subcommand

SafeTab-H `validate` runs successfully on the full geography (GRF-C.txt) inputs, race/ethnicity inputs and the Tumult generated simulated household-records.txt having 100 million records within about 15 minutes.

Execute subcommand

Note: `execute` includes `validate` by default

safetab-h on a cluster: We used a Tumult-generated synthetic data with a household file containing 100 million records, and a T1 output generated from a 300-million person file. We ran SafeTab-H with output validation on the full geography (GRF-C.txt) inputs and race/ethnicity inputs, tabulating both US and PR. The program completed in roughly 3 hours using either the PureDP privacy definition or the Rho zCDP privacy definition.

System Requirements

See safetab_h README.

Testing Plan

See safetab_h TESTPLAN.

Documentation

README - contains instructions for installing, running, and testing SafeTab-H..
LICENSE - software license under which SafeTab-H is distributed

Input/Output Specification

See Appendix A.

Appendix A: File Specifications

This appendix provides details on the formats for the input and output to be used in the 2020 Census Disclosure Avoidance System (DAS) activities supported by Tumult Labs. Input Dataframes refers to python spark dataframe objects created by Census DAS reader programs or by reading synthetic data in csv file format. Output Files refers to files produced by SafeTab-H intended for further use by Census Bureau.

A note of notation:

DataType	Description
StringType(n)	A string with up to n characters
StringType	A string without a character limit
IntegerType(n)	A number with up to n digits
IntegerType	A number without a digit limit

Input Dataframes

GRF-C-df

This dataframe contains a row for each Census block, and indicates which larger geographic areas the block is contained within.

Version and Date

2020-1-14.v1

Column Names and Format Definitions

The schema for this file is maintained by the Census Bureau.

Encoding
UTF-8

Delimiter Character
vertical bar (|)

Comment Character
Not supported.

household-records-df

Representation of custom household records derived from the CEF Unit file that is input to DAS. We assume that *household-records-df* will contain exactly one row for each household (occupied housing unit) in the US. We assume no group quarters facilities are present in the dataframe. Domains for all columns (except HOUSEHOLD_TYPE) match the identically named columns in CEF20_PER, CEF20_UNIT or GRFC.

Version and Date

2023-01-27.v4

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
TABBLKST	State code	StringType(2)	01, 02, 04–06, 08–13, 15–42, 44–51, 53–56, 60, 66, 69, 72, 78
TABBLKCOU	County Code	StringType(3)	001–840
TABTRACTCE	Census Tract Code	StringType(6)	000100–998999
TABBLK	Block Code	StringType(4)	0001–9999
HHRACE	Represents all possible major race categories of householder	StringType(2)	01 = White alone 02 = Black alone 03 = AIAN alone 04 = Asian alone 05 = NHPI alone 06 = SOR alone 07 = White; Black 08 = White; AIAN 09 = White; Asian 10 = White; NHPI 11 = White; SOR 12 = Black; AIAN 13 = Black; Asian 14 = Black; NHPI 15 = Black; SOR 16 = AIAN; Asian 17 = AIAN; NHPI 18 = AIAN; SOR 19 = Asian; NHPI 20 = Asian; SOR 21 = NHPI; SOR 22 = White; Black; AIAN 23 = White; Black; Asian 24 = White; Black; NHPI 25 = White; Black; SOR 26 = White; AIAN; Asian 27 = White; AIAN; NHPI

			28 = White; AIAN; SOR 29 = White; Asian; NHPI 30 = White; Asian; SOR 31 = White; NHPI; SOR 32 = Black; AIAN; Asian 33 = Black; AIAN; NHPI 34 = Black; AIAN; SOR 35 = Black; Asian; NHPI 36 = Black; Asian; SOR 37 = Black; NHPI; SOR 38 = AIAN; Asian; NHPI 39 = AIAN; Asian; SOR 40 = AIAN; NHPI; SOR 41 = Asian; NHPI; SOR 42 = White; Black; AIAN; Asian 43 = White; Black; AIAN; NHPI 44 = White; Black; AIAN; SOR 45 = White; Black; Asian; NHPI 46 = White; Black; Asian; SOR 47 = White; Black; NHPI; SOR 48 = White; AIAN; Asian; NHPI 49 = White; AIAN; Asian; SOR 50 = White; AIAN; NHPI; SOR 51 = White; Asian; NHPI; SOR 52 = Black; AIAN; Asian; NHPI 53 = Black; AIAN; Asian; SOR 54 = Black; AIAN; NHPI; SOR 55 = Black; Asian; NHPI; SOR 56 = AIAN; Asian; NHPI; SOR 57 = White; Black; AIAN; Asian; NHPI

			58 = White; Black; AIAN; Asian; SOR 59 = White; Black; AIAN; NHPI; SOR 60 = White; Black; Asian; NHPI; SOR 61 = White; AIAN; Asian; NHPI; SOR 62 = Black; AIAN; Asian; NHPI; SOR 63 = White; Black; AIAN; Asian; NHPI; SOR
QRACE1	Edited First Race Variable	StringType(4)	1000-8999
QRACE2	Edited Second Race Variable	StringType(4)	1000-8999 or Null if no code. If QRACE2 is Null, then so are QRACE3-QRACE8
QRACE3	Edited Third Race Variable	StringType(4)	1000-8999 or Null if no code. If QRACE3 is Null, then so are QRACE4-QRACE8
QRACE4	Edited Fourth Race Variable	StringType(4)	1000-8999 or Null if no code. If QRACE4 is Null, then so are QRACE5-QRACE8
QRACE5	Edited Fifth Race Variable	StringType(4)	1000-8999 or Null if no code. If QRACE5 is Null, then so are QRACE6-QRACE8
QRACE6	Edited Sixth Race Variable	StringType(4)	1000-8999 or Null if no code. If QRACE6 is Null, then so will QRACE7-QRACE8
QRACE7	Edited Seventh Race Variable	StringType(4)	1000-8999 or Null if no code. If QRACE7 is Null, then so is QRACE8
QRACE8	Edited Eighth Race Variable	StringType(4)	1000-8999 or Null if no code.
QSPAN	Final Edited Hispanic origin variable	StringType(4)	1000-8999, 9950
HOUSEHOLD_TYPE	Recode created by CEF reader. See CEF reader documentation for details.	IntegerType(1)	1 = Family household, Married couple family 2 = Family household, Male householder, no spouse present

			3 = Family household, Female householder, no spouse present 4 = Nonfamily household, Householder living alone 5 = Nonfamily household, Householder not living alone
TEN	Tenure status	IntegerType(1)	1 = Owned with a mortgage or a loan 2 = Owned free and clear 3 = Renter occupied 4 = Occupied without payment of rent

Encoding
UTF-8

Delimiter Character
vertical bar (|)

Comment Character
Not supported.

Sample Records
TABBLKST|TABBLKCOU|TABTRACTCE|TABBLK|HHRACE|QRACE1|QRACE2|QRACE3|QRACE4|QRACE5|QRACE6|QRACE7|QRACE8|QSPAN|HOUSEHOLD_TYPE|TEN

08|500|000300|0060|01|1000|Null|Null|Null|Null|Null|Null|Null|1000|1|2

Additional Inputs

config.json

Description

json file encoding inputs to SafeTab-H as key, value pairs. The key value pairs expected by SafeTab-H 3.0.0 are described below.

Version and Date

2023-01-27.v2

Key Value Names and Format Definitions

Key	Description	Value Format	Legal Values
max_race_codes	Maximum race codes for a single record.	Int	1-8
privacy_budget_h_t3_level_<x>_<geo> (for x in {1,2} and geo in {usa, state, county, tract, place, aiannh, pr_state, pr_county, pr_tract, pr_place})	The privacy loss budget assigned to t3 for geo level <geo> and characteristic iteration level <x>.	Float	0.128, 5.66, etc
privacy_budget_h_t4_level_<x>_<geo> (for x in {1,2} and geo in {usa, state, county, tract, place, aiannh, pr_state, pr_county, pr_tract, pr_place})	The privacy loss budget assigned to t4 for geo level <geo> and characteristic iteration level <x>.	Float	0.128, 5.66, etc

thresholds_h_t3	A list of thresholds for each combination of geography level and iteration level for table t3.
	
	If the corresponding privacy loss budgets are set to 0, the threshold values can be set to anything.
	
	{
	<geo>_<x>: [3 non-decreasing integers] for geo in {usa, state, county, tract, place, aiannh, pr_state, pr_county, pr_tract, pr_place} and x in {1,2}
	}
		{
		Str: List[int]
		}
			See example config below

thresholds_h_t4	A list of thresholds for each combination of geography level and iteration level for table t4.
	
	If the corresponding privacy loss budgets are set to 0, the threshold values can be set to anything.
	
	{
	<geo>_<x>: [1 integers]
	for geo in {usa, state, county, tract, place, aiannh, pr_state, pr_county, pr_tract, pr_place} and x in {1,2}
	}
		{
		Str:
		List[int]
		}
			See example config below

allow_negative_counts	Whether SafeTab-P should round negative counts up to 0.	Bool	{true, false}
run_us	When true, run safetab-p for 50 states + DC	Bool	{true, false}
run_pr	When true, run safetab-p for Puerto Rico. Only records corresponding to “72” will be tabulated for the PR run (if run_pr=true).	Bool	{true, false}
reader	The reader being used. This reader reads the person, and geo files and filters them based on which states are being used. Which states are being used is based on state_filter_us for US runs and is just “72” for PR runs.	Str	{“csv”, “cef”}
state_filter_us	A list of states from the 50 states + DC to include in the US run. PR should not be included in the list. If run_us=true, only records corresponding to these states will be tabulated for the US run. If run_us=false, state_filter_us is ignored.	List[str]	See example config below
privacy_defn	The privacy definition being used, either Pure DP (“puredp”) or Rho zCDP (“zcdp”). Determines how privacy budgets are interpreted.	Str	{“puredp”, “zcdp”}

Encoding

UTF-8

Delimiter Character

JSON uses structured key-value pairs so does not have delimiters.

Comment Character

JSON does not support comment characters.

Sample


```
{
  "max_race_codes": 8,
  "privacy_budget_h_t3_level_1_usa": 0.008,
  "privacy_budget_h_t3_level_2_usa": 0.534,
  "privacy_budget_h_t3_level_1_state": 0.008,
  "privacy_budget_h_t3_level_2_state": 0.534,
  "privacy_budget_h_t3_level_1_county": 0.008,
  "privacy_budget_h_t3_level_2_county": 0.159,
  "privacy_budget_h_t3_level_1_tract": 0,
  "privacy_budget_h_t3_level_2_tract": 2.43,
  "privacy_budget_h_t3_level_1_place": 0,
  "privacy_budget_h_t3_level_2_place": 2.43,
  "privacy_budget_h_t3_level_1_aiannh": 0,
  "privacy_budget_h_t3_level_2_aiannh": 0.159,
  "privacy_budget_h_t3_level_1_pr_state": 0.008,
  "privacy_budget_h_t3_level_2_pr_state": 0.534,
  "privacy_budget_h_t3_level_1_pr_county": 0.008,
  "privacy_budget_h_t3_level_2_pr_county": 0.159,
  "privacy_budget_h_t3_level_1_pr_tract": 0,
  "privacy_budget_h_t3_level_2_pr_tract": 2.43,
  "privacy_budget_h_t3_level_1_pr_place": 0,
  "privacy_budget_h_t3_level_2_pr_place": 2.43,
  "privacy_budget_h_t4_level_1_usa": 0.008,
  "privacy_budget_h_t4_level_2_usa": 0.534,
  "privacy_budget_h_t4_level_1_state": 0.008,
  "privacy_budget_h_t4_level_2_state": 0.534,
  "privacy_budget_h_t4_level_1_county": 0.008,
  "privacy_budget_h_t4_level_2_county": 0.159,
  "privacy_budget_h_t4_level_1_tract": 0,
  "privacy_budget_h_t4_level_2_tract": 2.43,
  "privacy_budget_h_t4_level_1_place": 0,
  "privacy_budget_h_t4_level_2_place": 2.43,
  "privacy_budget_h_t4_level_1_aiannh": 0,
  "privacy_budget_h_t4_level_2_aiannh": 0.159,
  "privacy_budget_h_t4_level_1_pr_state": 0.008,
  "privacy_budget_h_t4_level_2_pr_state": 0.534,
  "privacy_budget_h_t4_level_1_pr_county": 0.008,
  "privacy_budget_h_t4_level_2_pr_county": 0.159,
  "privacy_budget_h_t4_level_1_pr_tract": 0,
  "privacy_budget_h_t4_level_2_pr_tract": 2.43,
  "privacy_budget_h_t4_level_1_pr_place": 0,
  "privacy_budget_h_t4_level_2_pr_place": 2.43,
  "thresholds_h_t3": {
    "(USA, 1)": [5000, 20000, 150000],
    "(USA, 2)": [500, 1000, 7000],
    "(STATE, 1)": [5000, 20000, 150000],
    "(STATE, 2)": [500, 1000, 7000],
    "(COUNTY, 1)": [5000, 20000, 150000],
    "(COUNTY, 2)": [1000, 5000, 20000],
```

```
    "(TRACT, 1)": [5000, 20000, 150000],
    "(TRACT, 2)": [1000, 5000, 20000],
    "(PLACE, 1)": [5000, 20000, 150000],
    "(PLACE, 2)": [1000, 5000, 20000],
    "(AIANNH, 1)": [5000, 20000, 150000],
    "(AIANNH, 2)": [1000, 5000, 20000],
    "(PR-STATE, 1)": [5000, 20000, 150000],
    "(PR-STATE, 2)": [500, 1000, 7000],
    "(PR-COUNTY, 1)": [5000, 20000, 150000],
    "(PR-COUNTY, 2)": [1000, 5000, 20000],
    "(PR-TRACT, 1)": [5000, 20000, 150000],
    "(PR-TRACT, 2)": [1000, 5000, 20000],
    "(PR-PLACE, 1)": [5000, 20000, 150000],
    "(PR-PLACE, 2)": [1000, 5000, 20000]
  },
  "thresholds_h_t4": {
    "(USA, 1)": [5000],
    "(USA, 2)": [500],
    "(STATE, 1)": [5000],
    "(STATE, 2)": [500],
    "(COUNTY, 1)": [5000],
    "(COUNTY, 2)": [1000],
    "(TRACT, 1)": [5000],
    "(TRACT, 2)": [1000],
    "(PLACE, 1)": [5000],
    "(PLACE, 2)": [1000],
    "(AIANNH, 1)": [5000],
    "(AIANNH, 2)": [1000],
    "(PR-STATE, 1)": [5000],
    "(PR-STATE, 2)": [500],
    "(PR-COUNTY, 1)": [5000],
    "(PR-COUNTY, 2)": [1000],
    "(PR-TRACT, 1)": [5000],
    "(PR-TRACT, 2)": [1000],
    "(PR-PLACE, 1)": [5000],
    "(PR-PLACE, 2)": [1000]
  },
  "allow_negative_counts": true,
  "run_us": true,
  "run_pr": false,
  "reader": "csv",
  "state_filter_us": ["01", "02", "11"],
  "privacy_defn": "zcdp"
}
```

race_and_ethnicity_codes.txt

Description

Universe of detailed race and ethnicity codes.

Version and Date

2023-01-27.v5

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
RACE_ETH_CODE	Numeric race or ethnicity code	StringType(4)	1000-9999
RACE_ETH_NAME	Common name of race or ethnicity	StringType	See example records below

Encoding

UTF-8

Delimiter Character

vertical bar (|)

Comment Character

Not supported.

Sample Records

RACE_ETH_CODE|RACE_ETH_NAME

1000|White (Checkbox)

1001|White

1002|White American (By Checkbox)

1003|White American (By Write-in)

1004|White (Edit Generated)

1010|Albanian

1015|Alsatian

1020|Andorran

race-characteristic-iterations.txt

Description

Universe of race characteristic iterations that includes the code associated with the iteration, the name of the iteration, whether the iteration is “Alone” (if ALONE=True) or “Alone or in any Combination” (if ALONE=False). Additionally, this file indicates whether the iteration should be tabulated *only* at the national and state level, or *only* below the state level. Note that SafeTab-P and SafeTab-H share a common race characteristic iteration input schema. Iterations that should be tabulated everywhere will have DETAILED_ONLY=False and COARSE_ONLY=False.

Assumptions

- Iteration codes in this file are distinct from iteration codes in ethnicity-characteristic-iterations.txt.
- Rows with DETAILED_ONLY=True and COARSE_ONLY=True are not permitted.

Version and Date

2023-01-27.v7

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
ITERATION_CODE	Numeric iteration code	StringType(4)	1000-9999
ITERATION_NAME	Common name of iteration (including Alone/Alone or in any Combination)	StringType	See examples below
LEVEL	Level of iteration in the iteration hierarchy	StringType(1)	0 = major races (not be tabulated by SafeTab-P) 1 = regional race groups 2 = detailed race groups
ALONE	True if “Alone”, False if “Alone or in any Combination”	StringType(5)	{“True”, “False”}
DETAILED_ONLY	True if this iteration should be tabulated only with the detailed iterations, i.e. **only** at the state and national level, False otherwise. Null if this is a level 0 iteration (and thus will not be tabulated).	StringType(5)	{“True”, “False”, “Null”}
COARSE_ONLY	True if this iteration should be tabulated only with the coarse iterations, i.e. **only** below state level, False otherwise. Null if this is	StringType(5)	{“True”, “False”, “Null”}

	a level 0 iteration (and thus will not be tabulated).		

Encoding
UTF-8

Delimiter Character
vertical bar (|)

Comment Character
Not supported.

Sample Records
ITERATION_CODE|*ITERATION_NAME*|*LEVEL*|*ALONE*|*DETAILED_ONLY*|*COARSE_ONLY*

1001|White alone|0|True|Null|Null

1002|European alone|1|True|False|False

1003|Albanian alone|2|True|False|False

1004|Alsatian alone|2|True|True|False

…

1070|Other European alone (All geos)|2|True|False|True

…

1060|European alone or in any combination|1|False|False|False

1061|Albanian alone or in any combination|2|False|False|False

Note: Race characteristic iterations are intended to form a hierarchy. At each iteration level, and each value of “Alone” or “Alone or in combination”, a single race code should map to only a single Common iteration, or up to one Detailed and one Other iteration. SafeTab-H checks the mapping and raises an error if a single code is mapped to too many iterations with a particular set of attributes.

ethnicity-characteristic-iterations.txt

Description

Universe of ethnicity characteristic iterations that includes the code associated with the iteration, and the name of the iteration. Additionally, this file indicates whether the iteration should be tabulated *only* at the national and state level, or *only* below the state level. Note that SafeTab-P and SafeTab-H share a common race characteristic iteration input schema. Iterations that should be tabulated everywhere will have DETAILED_ONLY=False and COARSE_ONLY=False.

Assumptions

- Iteration codes in this file are distinct from iteration codes in race-characteristic-iterations.txt.
- Rows with DETAILED_ONLY=True and COARSE_ONLY=True are not permitted.

Version and Date

2023-01-27.v7

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
ITERATION_CODE	Numeric iteration code	StringType(4)	1000-9999
ITERATION_NAME	Common name of iteration	StringType	See examples below
LEVEL	Level of iteration in the iteration hierarchy:	StringType(1)	0 = major races (not be tabulated by SafeTab-P) 1 = regional race groups 2 = detailed race groups
DETAILED_ONLY	True if this iteration should be tabulated only with the detailed iterations, i.e. **only** at the state and national level, False otherwise. Null if this is a level 0 iteration (and thus will not be tabulated).	StringType(5)	{“True”, “False”, “Null”}
COARSE_ONLY	True if this iteration should be tabulated only with the coarse iterations, i.e. **only** below state level, False otherwise. Null if this is a level 0 iteration (and thus will not be tabulated).	StringType(5)	{“True”, “False”, “Null”}

Encoding
UTF-8

Delimiter Character
vertical bar (|)

Comment Character
Not supported.

Sample Records
ITERATION_CODE|ITERATION_NAME|LEVEL|DETAILED_ONLY|COARSE_ONLY

3008|Hispanic or Latino (of any race)|0|Null|Null

3009|Mexican|1|False|False

3010|Central American|1|False|False

3011|Costa Rican|2|False|False

…

3039|Other Hispanic, Latino, or Spanish responses, not specified (National)|2|True|False

3040|Other Hispanic or Latino, not specified (All geos)|2|False|True

Note: Ethnicity characteristic iterations are intended to form a hierarchy. At each iteration level, and each value of “Alone” or “Alone or in combination”, a single race code should map to only a single Common iteration, or up to one Detailed and one Other iteration. SafeTab-H checks the mapping and raises an error if a single code is mapped to too many iterations with a particular set of attributes.

race-and-ethinicty-code-to-iteration.txt

Description

Mapping of characteristic iterations to detailed race and ethnicity codes. There will be one row for every iteration_code that a race_eth_code is associated with.

Version and Date

2023-01-27.v5

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal values
ITERATION_CODE	numeric code of characteristic iteration	StringType(4)	1000-9999
RACE_ETH_CODE	numeric race or ethnicity code	StringType(4)	1000-9999

Encoding

UTF-8

Delimiter Character

vertical bar (|)

Comment Character

Not supported.

Sample Records

ITERATION_CODE|RACE_ETH_CODE

1003|1010

1003|1011

1003|1012

1003|1013

1003|1014

reader_config

Description

Configuration file for CEF Reader program. The CEF Reader program reads its input parameters from a .ini file. The INI configuration file consists of sections, each beginning with a [section] header, followed by key/value entries separated by “=”. The following is a sample of a CEF Reader configuration file.

Sample Section

[paths]

cef_year = 2020

per_dir = <US Person CEF Path>

unit_dir = <US Units CEF Path>

per_dir_pr = <PR Person CEF Path>

unit_dir_pr = <PR Units CEF Path>

grfc_dir = <GRFC Path>

pop_dir = <Population Group Totals Path>

per_file_format = CEF20_PER_%%s.txt

unit_file_format = CEF20_UNIT_%%s.txt

geo_file_format = grfc_tab20_%%s.txt

pop_group_totals.txt

Description

See T1 output format in the SafeTab-P documentation.

Output Files

Output is saved in directories with multiple pipe-delimited csv files. CSV files will have different names matching the part-00000-*.csv pattern each time the program is run. The output directory structure is:

```
output_dir/
  t3/
    T03001/
      part-00000-22220762-ee6f-4503-8618-15dcf9882592-c000.csv
    T03002/
      part-00000-803a65b0-bc01-4314-98e7-2f55f3b748a7-c000.csv
    T03003/
      ...
    T04004/
      ...
  t4/
    T04001/
      ...
    T04002/
      ...
```

Note: Noisy measurements are produced for all geographic entities in schema regardless of whether or not those entities contain housing units or group quarters facilities. This means structural zeros are not directly handled by SafeTab-H. If the structural zeros are removed from the geo_df input, they will not appear in the SafeTab-H outputs. If structural zeros are not removed from the geo_df input, they will appear in the SafeTab-H outputs.

T03001

Description

A folder containing one or more csv files. Each file contains household type statistics. Each file contains a header with attribute names, followed by rows containing a specification of attributes, and the count for those attributes. When the safetab-h algorithm is run with both run_us and run_pr set to True, the T03001/ folder will contain 2 csv files with the prefix part-00000-XXXXX - one with US rows and the other with PR rows. If only one of run_us and run_pr is set to True, the T03001/folder will contain only 1 csv file with corresponding rows. The csv files will have a format as described below.

Version and Date
2022-07-27.v2

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
REGION_ID	entity code corresponding to one of STATE, COUNTY, AIANNH area, etc	StringType(11)	1, 44, 44007, etc
REGION_TYPE	Name of geography level	StringType(9)	{USA, AIANNH, STATE, COUNTY, TRACT, PLACE, PR-STATE, PR-COUNTY, PR-TRACT, PR-PLACE}
ITERATION_CODE	Characteristic iteration code	IntegerType(4)	1000-9999
T3_DATA_CELL	Data cell number from T3 table shell	IntegerType(1)	0 = Total
COUNT	Number of households corresponding to given table cell	IntegerType	2, -52, 670, etc

Encoding
UTF-8

Delimiter Character
vertical bar (|)

Comment Character
Not supported.

Sample Records
REGION_ID|REGION_TYPE|ITERATION_CODE|T3_DATA_CELL|COUNT

1|USA|2790|0|12345

01|STATE|2790|0|13245

01123|COUNTY|2800|0|5543

T03002

Description

A folder containing one or more csv files. Each file contains household type statistics. Each file contains a header with attribute names, followed by rows containing a specification of attributes, and the count for those attributes. When the safetab-h algorithm is run with both run_us and run_pr set to True, the T03002/ folder will contain 2 csv files with the prefix part-00000-XXXXX - one with US rows and the other with PR rows. If only one of run_us and run_pr is set to True, the T03002/folder will contain only 1 csv file with corresponding rows. The csv files will have a format as described below.

Version and Date

2023-07-27.v2

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
REGION_ID	entity code corresponding to one of STATE, COUNTY, AIANNH area, etc	StringType(11)	1, 44, 44007, etc
REGION_TYPE	Name of geography level	StringType(9)	{USA, AIANNH, STATE, COUNTY, TRACT, PLACE, PR-STATE, PR-COUNTY, PR-TRACT, PR-PLACE}
ITERATION_CODE	Characteristic iteration code	IntegerType(4)	1000-9999
T3_DATA_CELL	Data cell number from T3 table shell	IntegerType(1)	0 = Total 1 = Family Households 6 = Nonfamily Households
COUNT	Number of households corresponding to given table cell	IntegerType	2, -52, 670, etc

Encoding

UTF-8

Delimiter Character

vertical bar (|)

Comment Character

Not supported.

Sample Records

REGION_ID|REGION_TYPE|ITERATION_CODE|T3_DATA_CELL|COUNT

1|USA|2790|0|12345

01|STATE|2790|0|13245

01123|COUNTY|2800|0|5543

T03003

Description

A folder containing one or more csv files. Each file contains household type statistics. Each file contains a header with attribute names, followed by rows containing a specification of attributes, and the count for those attributes. When the safetab-h algorithm is run with both run_us and run_pr set to True, the T03003/ folder will contain 2 csv files with the prefix part-00000-XXXXX - one with US rows and the other with PR rows. If only one of run_us and run_pr is set to True, the T03003/folder will contain only 1 csv file with corresponding rows. The csv files will have a format as described below.

Version and Date

2023-07-27.v2

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
REGION_ID	entity code corresponding to one of STATE, COUNTY, AIANNH area, etc	StringType(11)	1, 44, 44007, etc
REGION_TYPE	Name of geography level	StringType(9)	{USA, AIANNH, STATE, COUNTY, TRACT, PLACE, PR-STATE, PR-COUNTY, PR-TRACT, PR-PLACE}
ITERATION_CODE	Characteristic iteration code	IntegerType(4)	1000-9999
T3_DATA_CELL	Data cell number from T3 table shell	IntegerType(1)	0 = Total 1 = Family Households 2 = Married couple family 3 = Other family 6 = Nonfamily Households

			7 = Householder living alone 8 = Householder not living alone
COUNT	Number of households corresponding to given table cell	IntegerType	2, -52, 670, etc

Encoding
UTF-8

Delimiter Character
vertical bar (|)

Comment Character
Not supported.

Sample Records
REGION_ID|REGION_TYPE|ITERATION_CODE|T3_DATA_CELL|COUNT

1|USA|2790|0|12345

01|STATE|2790|0|13245

01123|COUNTY|2800|0|5543

T03004

Description
A folder containing one or more csv files. Each file contains household type statistics. Each file contains a header with attribute names, followed by rows containing a specification of attributes, and the count for those attributes. When the safetab-h algorithm is run with both run_us and run_pr set to True, the T03004/ folder will contain 2 csv files with the prefix part-00000-XXXXX - one with US rows and the other with PR rows. If only one of run_us and run_pr is set to True, the T03004/folder will contain only 1 csv file with corresponding rows. The csv files will have a format as described below.

Version and Date
2023-07-27.v2

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values

REGION_ID	entity code corresponding to one of STATE, COUNTY, AIANNH area, etc	StringType(11)	1, 44, 44007, etc
REGION_TYPE	Name of geography level	StringType(9)	{USA, AIANNH, STATE, COUNTY, TRACT, PLACE, PR-STATE, PR-COUNTY, PR-TRACT, PR-PLACE}
ITERATION_CODE	Characteristic iteration code	IntegerType(4)	1000-9999
T3_DATA_CELL	Data cell number from T3 table shell	IntegerType(1)	0 = Total 1 = Family Households 2 = Married couple family 3 = Other family 4 = Male householder, no spouse or partner 5 = Female householder, no spouse or partner 6 = Nonfamily Households 7 = Householder living alone 8 = Householder not living alone
COUNT	Number of households corresponding to given table cell	IntegerType	2, -52, 670, etc

Encoding
UTF-8

Delimiter Character
vertical bar (|)

Comment Character
Not supported.

Sample Records
REGION_ID|REGION_TYPE|ITERATION_CODE|T3_DATA_CELL|COUNT

1|USA|2790|0|12345

01|STATE|2790|0|13245

01123|COUNTY|2800|0|5543

T04001

Description

A folder containing one or more csv files. Each file contains household type statistics. Each file contains a header with attribute names, followed by rows containing a specification of attributes, and the count for those attributes. When the safetab-h algorithm is run with both run_us and run_pr set to True, the T04001/ folder will contain 2 csv files with the prefix part-00000-XXXXX - one with US rows and the other with PR rows. If only one of run_us and run_pr is set to True, the T04001/ folder will contain only 1 csv file with corresponding rows. The csv files will have a format as described below.

Version and Date

2022-07-27.v2

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
REGION_ID	entity code corresponding to one of STATE, COUNTY, AIANNH area, etc	StringType(11)	1, 44, 44007, etc
REGION_TYPE	Name of geography level	StringType(9)	{USA, AIANNH, STATE, COUNTY, TRACT, PLACE, PR-STATE, PR-COUNTY, PR-TRACT, PR-PLACE}
ITERATION_CODE	Characteristic iteration code	IntegerType(4)	1000-9999
T4_DATA_CELL	Data cell number from T4 table shell	IntegerType(1)	0 = Total
COUNT	Number of households corresponding to given table cell	IntegerType	2, -52, 670, etc

Encoding

UTF-8

Delimiter Character

vertical bar (|)

Comment Character

Not supported.

Sample Records

REGION_ID|REGION_TYPE|ITERATION_CODE|T4_DATA_CELL|COUNT

1|USA|2790|0|12345

01|STATE|2790|0|13245

01123|COUNTY|2800|0|5543

T04002

Description

A folder containing one or more csv files. Each file contains household type statistics. Each file contains a header with attribute names, followed by rows containing a specification of attributes, and the count for those attributes. When the safetab-h algorithm is run with both run_us and run_pr set to True, the T04002/ folder will contain 2 csv files with the prefix part-00000-XXXXX - one with US rows and the other with PR rows. If only one of run_us and run_pr is set to True, the T04002/folder will contain only 1 csv file with corresponding rows. The csv files will have a format as described below.

Version and Date

2022-07-27.v2

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
REGION_ID	entity code corresponding to one of STATE, COUNTY, AIANNH area, etc	StringType(11)	1, 44, 44007, etc
REGION_TYPE	Name of geography level	StringType(9)	{USA, AIANNH, STATE, COUNTY, TRACT, PLACE, PR-STATE, PR-COUNTY, PR-TRACT, PR-PLACE}
ITERATION_CODE	Characteristic iteration code	IntegerType(4)	1000-9999
T4_DATA_CELL	Data cell number from T4 table shell	IntegerType(1)	0 = Total 1 = Owned with a mortgage or a loan 2 = Owned free and clear 3 = Renter occupied

COUNT	Number of households corresponding to given table cell	IntegerType	2, -52, 670, etc

Encoding
UTF-8

Delimiter Character
vertical bar (|)

Comment Character
Not supported.

Sample Records
REGION_ID|REGION_TYPE|ITERATION_CODE|T4_DATA_CELL|COUNT

1|USA|2790|0|12345

01|STATE|2790|0|13245

01123|COUNTY|2800|0|5543