

Introduction to the Sample Files of the General-Purpose Sampling Database for Individual Income Tax Assessment File

I. Sample Data File (One set of sampling data will be provided per case, following the standard sampling method.)

Dataset Code	Dataset	Sample Dataset Name	Year
IBET100	Individual Income Tax Return Assessment File	IBE100_S_Sample_Year	2011-2020
IBET110	Individual Income Basic Tax Assessment File	IBE110_S_Sample_Year	2011-2020
IBET300	Individual Income Tax Assessment Itemized Details File	IBE300_S_Sample_Year	2011-2020
IBET600	Individual Income Tax Assessment Income File	IBE600_S_Sample_Year	2011-2020
IBET610	Individual Income Tax Assessment Basic Tax Income File	IBE610_S_Sample_Year	2011-2020
IBET700	Individual Income Tax Assessment Registration File	IBE700_S_Sample_Year	2011-2020
IIMH001	Individual Income Tax Consolidated Income Data File	IIMH001_S_Sample_Year	2011-2020
EGXH62X	Land Property File	EGXH62X_S_Sample_Year	2011-2020
EGXH61X	Housing Property File	EGXH61X_S_Sample_Year	2011-2020
EGXH63X	Vehicle Property File	EGXH63X_S_Sample_Year	2011-2020
EGXH64X	Investment Property File	EGXH64X_S_Sample_Year	2011-2020

II. Sampling Design (For detailed methods and validation plans, please refer to the “Sampling Methods and Validation Plan of the General-Purpose Sampling Database for Individual Income Tax.”)

(1) Data Description:

1. Individual income tax is filed by household, with assessment data such as assessed income, various deductions, withholding tax, and creditable tax amounts consolidated into the assessment file.
2. Household income data below the taxable threshold is excluded.
3. The sampling is based on the assessment file for Tax Year 2020. The taxpayer identification numbers in this file are used to retrieve sampling data extending back to Tax Year 2011.
4. Income summary and property data are compiled based on the Year 2020 assessment sampling file, linked with the tax registration file. Through this linkage, applicants can access personal data of household members in the sample, and perform data integration and analyses at the individual level using the income summary or property files.

(2) Sampling Methodology and Validation:

1. Stratified simple random sampling at the 95% confidence level:
 - (1) Using “total assessed income” as the standard, extreme values within the lowest 1% around the 20th percentile are the first excluded. Afterwards, the population dataset is divided into 20 equal quantities, with each group representing 5 percentile points.
 - (2) Using the standard sample size formula, the sample size for the 20th percentile is calculated at 95% confidence level, and the same sample size is applied to each stratum (percentile).
2. The representativeness of the sample is assessed through descriptive statistics, cumulative probability distribution plots, and K-S tests.
3. A total of 240,000 records were sampled and randomly divided into 10 subsamples based on total assessed income percentiles, with 24,000 records in each subsample. One subsample was then randomly selected to verify population representativeness.

(3) Sample Size of Sampling Files:

In principle, each application is provided with one small sampling file containing 24,000 records.

III. Available Fields

Fields for Use: Commonly used research fields are provided (see the database user manual for details).

IV. Payment Method:

- (1) Data processing fee, monitoring room equipment usage fees, and management fees should be charged in accordance with the Fee Standards for Fiscal and Tax Information provided by the Fiscal Information Agency, Ministry of Finance.
- (2) The data usage period is one year, with the actual period of use limited to the approved duration of each case.