



TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

Earned Income Tax Credit Interactive Database Documentation

November 2018

Within the Internal Revenue Service (IRS), the Stakeholder Partnerships, Education and Communication (SPEC) office is responsible for outreach and education. As part of its mission, SPEC maintains a dataset of ZIP code level data for low-income tax-returns. This document describes the Urban-Brookings Tax Policy Center Earned Income Tax Credit Interactive Database (www.tpc-eitc-tool.urban.org) which makes the data available as a searchable database with expanded information for additional geographical areas.

For any questions about the data and the tool, please contact Richard Auxier (rauxier@urban.org) and Aravind Boddupalli (aboddupalli@urban.org).

IRS SPEC DATA

The Stakeholder Partnerships, Education and Communication (SPEC) office is the outreach and education arm of the Wage and Investment Division within the Internal Revenue Service (IRS). SPEC manages the Volunteer Income Tax Assistance (VITA) and the Tax Counseling for the Elderly (TCE) programs. As part of their mission, SPEC releases a database of tax returns organized by tax year, market segment (type of tax filer), and ZIP code. These data present information on tax filers who fall into one of the categories covered by SPEC's mission, including low income returns, returns claiming the earned income tax credit (EITC), elderly returns, and returns that used volunteer tax assistance.

The SPEC data presents information for the following market segments:

1. **Total Returns:** All returns with adjusted gross income (AGI) less than \$60,000.
2. **Low Income Returns:** Returns with AGI below the EITC highest income limit in a specific year (\$53,267 for tax year 2015).
3. **EITC Returns:** Any return that claimed an EITC.
4. **VITA Returns:** All returns filed at a Volunteer Income Tax Assistance (VITA) site, as designated by the presence of a Site Identification Number (SIDN) on the return linked to a SPEC-recognized VITA site.
5. **Military VITA Returns:** All returns filed at a military VITA site, as designated by the presence of a Site Identification Number (SIDN) on the return linked to a SPEC-recognized military VITA site.
6. **TCE Returns:** All returns filed at a Tax Counseling for the Elderly (TCE site), as designated by the presence of a Site Identification Number (SIDN) on the return linked to a SPEC-recognized TCE site.
7. **Elderly Returns:** All returns where the primary filer's age is 60 or older.
8. **Low Income Elderly Returns:** All returns where the primary filer's age is 60 or older and his or her AGI is below the highest EITC income eligibility limit.
9. **ITIN Returns:** All returns where the primary or secondary taxpayer, or any of the first four dependents, uses an Individual Taxpayer Identification Number (ITIN).

For each of these market segments, which we refer to as "type of tax filer," the SPEC data contains ZIP code level observations that record the count of returns and dollar amounts related to several tax-return variables. The full list of variables is presented in Table 1. In all market segments, tax returns are only included if their AGI is less than \$60,000. These data are subject to suppression rules to protect taxpayer confidentiality. Data are only included if there are more than 100 total returns in that ZIP code for a specific tax year. Specific information (for example, the count of tax returns receiving the EITC) is only included if there are more than 20 returns with that characteristic in that ZIP code. The data excludes certain ZIP codes such as overseas military and unique ZIP codes such as individual buildings.

The SPEC data for tax year 2015 are missing data for about 3,400 ZIP codes. Most of these ZIP codes are concentrated in Illinois, Michigan, Pennsylvania, and Texas. As a result, the tool does not provide any 2015 data for these four states, except for the available ZIP codes.

DATA METHODS

TPC takes the ZIP code level data provided by SPEC and create files for different geographies. The data are available by state, ZIP code, county, congressional district, upper state legislative chamber, lower state legislative chamber, city or town, and metropolitan division, for each type of tax filer and tax year.

We first standardize the data across tax years, and then create separate files for each geography via allocation factors from Missouri Census Data Center (MCDC)'s MABLE System¹.

Data Cleaning and Standardization

For each year of data, we begin by standardizing the variable names and organization of the SPEC data such that different years have a matching structure. We make sure that empty variables are dropped, and there are no returns with AGI over \$60,000. We drop all ZIP codes outside the 50 states and District of Columbia, and any missing or non-specific ZIP codes.

The following is a full list of variables included in our dataset:

TABLE 1
SPEC Variables
Variable names and descriptions



Variable Name	Variable Description
tax_year	The tax year of the filed tax return
state_name	State name
state_abb	State abbreviation
state_fips	State FIPS code
zip_code	ZIP code
market_segment	The type of tax filer associated with the filed tax return
total_returns	Total returns for the selected type of tax filer
new_returns	Returns that did not file last year
eitc_returns	Returns claiming the EITC
eitc_amount	Total EITC amount recieved
ctc_returns	Returns with CTC
ctc_amount	Total CTC Amount recieved
ctc_refund_returns	Returns with refundable CTC
ctc_refund_amount	Total Amount of refundable CTC

¹ See here: <http://mcdc.missouri.edu/websas/geocorr14.html>.

cdcc_returns	Returns with Child and Dependent Care Credit
balance_returns	Returns with a balance due
balance_amount	Total Amount of balance due
refund_returns	Returns with a refund
refund_amount	Total Amount of refund due
direct_returns	Returns recieving a direct deposit
ral_returns	Returns requesting a Refund Anticipation Loan
rac_returns	Returns requesting a Refund Anticipation Check
self_returns	Returns that were prepared by taxpayer
paid_returns	Returns that were prepared by a paid preparer
volunteer_returns	Returns that were prepared by a volunteer preparer
single_returns	Returns that filed as single
married_j_returns	Returns that filed as either MFJ or qual. widow
married_s_returns	Returns that filed as MFS
hoh_returns	Returns that filed as HOH
f1040_returns	Returns that used form 1040
f1040a_returns	Returns that used form 1040a
f1040ez_returns	Returns that used form 1040ez
cef_returns	Returns with Schedules E, C, or F
education_returns	Returns with Education Credit
studentloan_returns	Returns with Student Loan deduction
wages_returns	Returns with wages greater than \$0
taxable_interest_returns	Returns with taxable interest income greater than \$0
taxable_socsecur_returns	Returns with taxable social security greater than \$0
taxable_socsecur_amount	Total amount of taxable social security reported
taxable_pensions_returns	Returns with taxable pensions greater than \$0
taxable_pensions_amount	Total amount of taxable pensions reported
taxable_ira_returns	Returns with taxable IRA distributions greater than \$0
taxable_ira_amount	Total amount of taxable IRA distributions reported
total_tax_returns	Returns with Total Tax present
total_tax_amount	Total amount of Total Tax reported
eitc_nochildren_returns	Returns with CP09 notice issued
eitc_children_returns	Returns with CP27 notice issued
both_eitc_ctc_returns	Returns claiming both EITC and CTC

split_refund_returns	Returns with a split refund
est_payments_returns	Returns with estimated tax payments
remittance_returns	Returns with a remittance with return
remittance_amount	Total amount of remittances with return reported
vita_returns	Returns prepared by VITA and co-located VITA sites
military_vita_returns	Returns prepared by military VITA sites
tce_returns	Returns prepared by TCE and AARP/TCE sites
irs_prepared_returns	Returns prepared or reviewed by IRS
f1040NR_returns	Returns that used used Form 1040NR
fother_returns	Returns that used all other firms like 1040SS or 1040PR
fscheda_returns	Returns with schedule A
fschedd_returns	Returns with schedule D (capital gains and losses)
paidprepare_electronic_returns	Returns used paid preparer and filed electronically
paidprepare_paper_returns	Returns used paid prepared and filed on paper
selfprepare_electronic_returns	Returns self-prepared and filed electronically (FreeFile included)
selfprepare_paper_returns	Returns self-prepared and filed on paper
selfprepare_freefile_returns	Returns prepared by taxpayer and filed using FreeFile
vita_tce_other_paper_returns	VITA, TCE, and Other returns self-prepared and filed on paper
vita_tce_other_electronic_returns	VITA, TCE, and Other returns self-prepared and filed electronically
paper_returns	Returns filed on paper
electronic_returns	Returns filed electronically
vcoded_returns	Returns that used tax software to print and filed on paper
electronic_repeat_returns	Returns that filed electronically this year and previous year
paper_repeat_returns	Returns that filed on paper this year and previous year
electronic_change_returns	Returns that filed electronically this year and paper previous year
paper_change_returns	Returns that filed paper this year and electronically previous year
taxable_income_returns	Returns with taxable income greater than \$0
agi_0_returns	Returns with AGI below \$5,000
agi_5_returns	Returns with AGI between \$5,000 and \$9,999
agi_10_returns	Returns with AGI between \$10,000 and \$14,999
agi_15_returns	Returns with AGI between \$15,000 and \$19,999
agi_20_returns	Returns with AGI between \$20,000 and \$25,999
agi_25_returns	Returns with AGI between \$25,000 and \$29,999
agi_30_returns	Returns with AGI between \$30,000 and \$34,999

agi_35_returns	Returns with AGI between \$35,000 and \$39,999
agi_40_returns	Returns with AGI between \$40,000 and \$49,999
agi_50_returns	Returns with AGI between \$50,000 and \$59,999

Geographic Allocation of ZIP Code Data

We distribute the ZIP code level data across other geographic levels using the allocation factors calculated by the Missouri Census Data Center's Geographic Correspondence Engine (MABLE). The MABLE system uses sub-ZIP code population estimates to create allocation factors that distribute values among areas.

Specifically, we use MABLE-derived allocation factors to create summaries for the following geographic areas:

1. **States** include All 50 states and the District of Columbia.
2. **Metropolitan Divisions** are composed of related counties of core-based metropolitan statistical areas (MSAs), based on delineations issued by the U.S. Office of Management and Budget in 2013.
3. **Counties** represent the primary legal subdivision of most states; exceptions include Alaska and Louisiana, which are divided into boroughs and parishes, respectively. You can find borough and parish data under counties in the database. County divisions are as of 2014.
4. **Cities and Towns** include incorporated places, such as cities, towns, and villages, as well as census-designated places, which are unincorporated areas delineated by the U.S. Census Bureau for statistical purposes. City and town divisions are as of 2014.
5. **Congressional Districts** represent electoral divisions designed to elect members of the U.S. House of Representatives, based on district boundaries as of the 114th Congress.
6. **State Legislative Districts** are available for both the "lower chamber" (or House) and "upper chamber" (or Senate) districts. Nebraska has a unicameral system and can be accessed through the state senate district data only. Both district boundaries are as of 2014 elections.

Larger geographies may undercount returns because of aggregation and suppression rules. For example, if five ZIP codes each suppress the count for number of returns prepared by military VITA sites (because each ZIP code has fewer than 20 such returns), the county containing only those five ZIP codes will also show "zero" returns.

Differences from IRS Statistics of Income (SOI) Data

The Statistics of Income (SOI) division of the IRS also publishes data on tax filers at the ZIP code, county, and state level. However, those data differ from the SPEC data in certain ways.

- SOI data include all individual income tax returns, regardless of AGI; the SPEC data only includes returns with AGI under \$60,000.
- SOI bases its state, county, and ZIP code data on administrative records of individual income tax returns (forms 1040) from the IRS Individual Master File (IMF) system. Included in these data are returns filed during the 12-month period, January 1 to December 31. The SPEC data includes returns filed from January 1 to June 30, which should capture most low-income returns.
- The SPEC data contains separate information for nine different market segments. Thus, it provides information specific to certain types of filers; for example, the number of EITC returns that also received a child tax credit.

Differences from Brookings EITC Interactive

The Brookings Institution's EITC Interactive served as the model for TPC's EITC Interactive Database. The TPC database extends the Brookings version in several ways:

- It allows users to search within geographic areas, for example for specific ZIP codes or congressional districts;
- It can display data for different types of tax filers rather than just for returns with EITC; and
- It includes all the variables from the SPEC data, for example the number of returns receiving Social Security benefits and the amount received.

There are also some minor data differences from the Brookings EITC Interactive:

- TPC's EITC Interactive Database uses the MCDC's MABLE allocations to assign ZIP code data across geographies. Brookings used Census-block level data and Geographic Information Systems (GIS) software to calculate the proportion of households within each geography's borders. Brookings' process is similar to MCDC's, but the methodologies differ enough to produce marginally different results.
- The SPEC data does not report the number of returns for a tax return item if there are fewer than 20 returns in that category, but related dollar amounts are available for those variables. Brookings used those dollar totals to estimate the number of returns by taking the total amount reported and dividing by the national average amount where number of returns were not reported. This was imputed for returns receiving a refund, returns with a balance due, returns claiming the EITC, and returns claiming the CTC. TPC's EITC Interactive Database does not impute suppressed totals at the ZIP code level for any variable.



The Tax Policy Center is a joint venture of the
Urban Institute and Brookings Institution.



BROOKINGS

For more information, visit taxpolicycenter.org
or email info@taxpolicycenter.org