Research Applied Analytics & Statistics

STATISTICS OF INCOME

Synthesizing the Supplemental Synthetic Public Use File

Victoria Bryant, Chris Rexrode, Derek Gutierrez

Statistics of Income, Internal Revenue Service

October 24, 2024

Transition to Synthetic PUF and Tiered Access

Tier	Access	To Whom
1	Tabular data and reports	Anybody – via website and published reports
2	Synthetic individual income information	Anybody who needs it – upon request to SOI
3	Validation server: Automated system allows researchers to access confidential tax return information in an environment that protects against disclosure	Researchers vetted by SOI with a research plan that could not be completed using tier 1 or tier 2 access.
4	Access to confidential microdata	Researchers approved for access through the Joint Statistical Research Program.

Transition to Synthetic PUF and Tiered Access

Tier	Access	To Whom
1	Tabular data and reports	Anybody – via website and published reports
2	Synthetic individual income information	Anybody who needs it – upon request to SOI
3	Validation server: Automated system allows researchers to access confidential tax return information in an environment that protects against disclosure	Researchers vetted by SOI with a research plan that could not be completed using tier 1 or tier 2 access.
4	Access to confidential microdata	Researchers approved for access through the Joint Statistical Research Program.

Taxpayer Privacy and Confidentiality

Any publicly released tax data must protect the confidentiality of individual taxpayers.

Tabular released data

- Rule of 3
- Rule of 10
- Dominance Rule
- Associated Suppression
- Disclosure by subtraction
- Cross-cell disclosure
- Complimentary disclosure

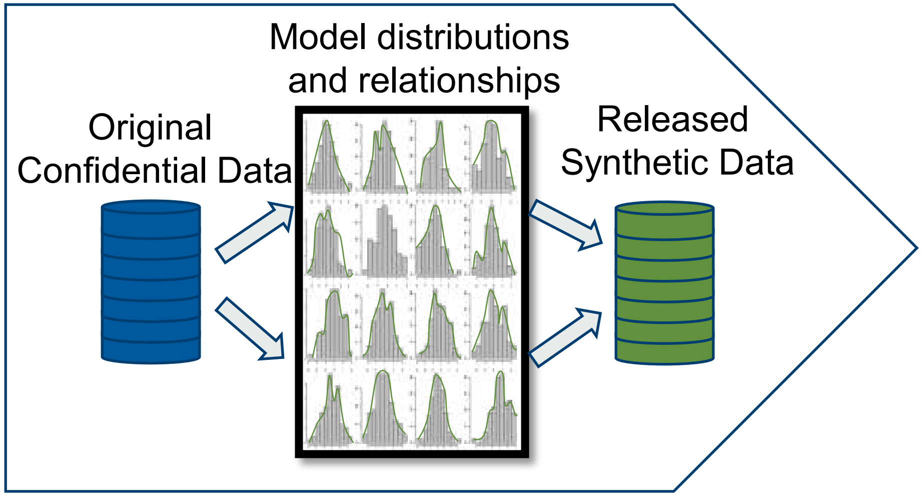
Microdata release

- Subsampling
  - Reweighting
- Aggregation
- Top Coding
- Blurring
  - Multivariate
  - Univariate
  - Rebalancing
- Random Noise
  - Rounding
- Suppression

As the scope of information on individuals that is publicly accessible increases, so too must SOI improve protection techniques.

Synthetic Data – General Approach

General methodology



No real observations are released

- Possibility of expanded demographic and/or tax information
- Possibility of multiple file releases targeting different population subsets

Potential Pitfalls:

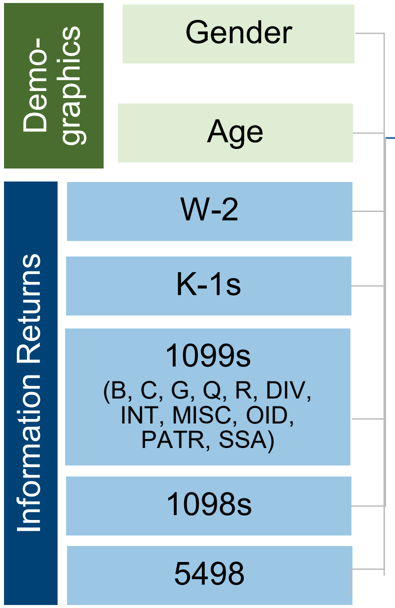
- Model overfitting may result in synthetic data too close to underlying data.
- Database Reconstruction Theorem (Dinur and Nissim, 2003): noisy subset sums can approximate individual records through solving a system of equations.
- Modeler may overcompensate for these concerns resulting in data without enough overlap to confidential data to be statistically useful.

Synthesis Process

Sample Selection

- Individuals receiving an information return in TY2012 with
  - No Form 1040
  - Income below filing threshold
- Based on Continuous Work History Sample (CWHS)
  - A 1-in-1,000 sample
- Additional limitations:
  - Drop late filers
  - Deceased persons
  - Foreign residents
  - Missing or invalid age & gender
- Final sample size ~ 26,000

Record Collection



Synthesis

Classification and Regression Tree (CART) Model

- A decision tree algorithm structured as a sequence of decisions.
- Synthesized categorical variables first followed by continuous.

Synthesis Process, cont.

Subdivide sample into 2 parts

1. Those records with just demographic information
2. Those records with at least one tax amount > 0

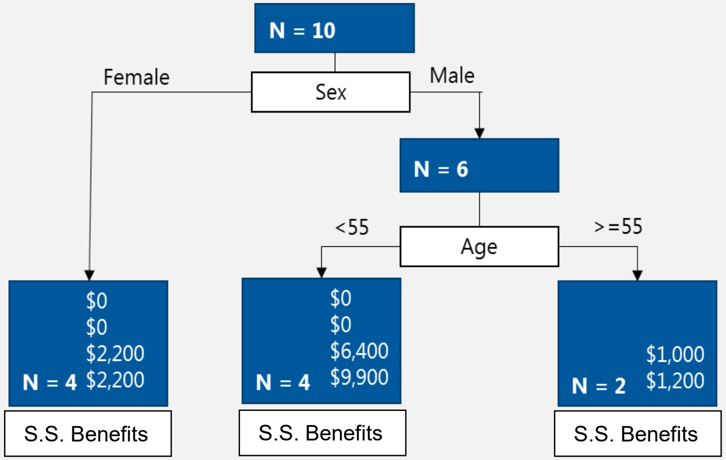
1

- Randomly assigned *gender,* based on proportions of underlying data
- Synthesized *age* based on *gender*
- Assigned zeros to all tax variables

Synthesized zero records

2

- Randomly assigned *gender,* based on proportions of underlying data
- Sequentially synthesize variables using CART starting with *age* conditional on previously synthesized outcome variables
  - Each point is randomly sampled with replacement
  - For continuous variables starting with *Social Security Benefits* then synthesized in order of linear correlation to *Social Security Benefits*.



Obs	Value	Ntile	Optimal KDE Variance	Synthetic Value Distribution
1	$0	1st	$0	0
2	$0	1st	$0	0
3	$6,400	66th	$650	~ N(μ=6,400, σ^2=650)
4	$9,900	98th	$2,300	~ N(μ=9,900, σ^2=2,300)

- Then draw a value from a smoothed KDE distribution
  - ~ N(μ = sampled value, σ2 = “percentile variance”)
  - Variance for a Kernel Density Estimator (KDE) of the percentile of the mean

Synthesized non-zero records

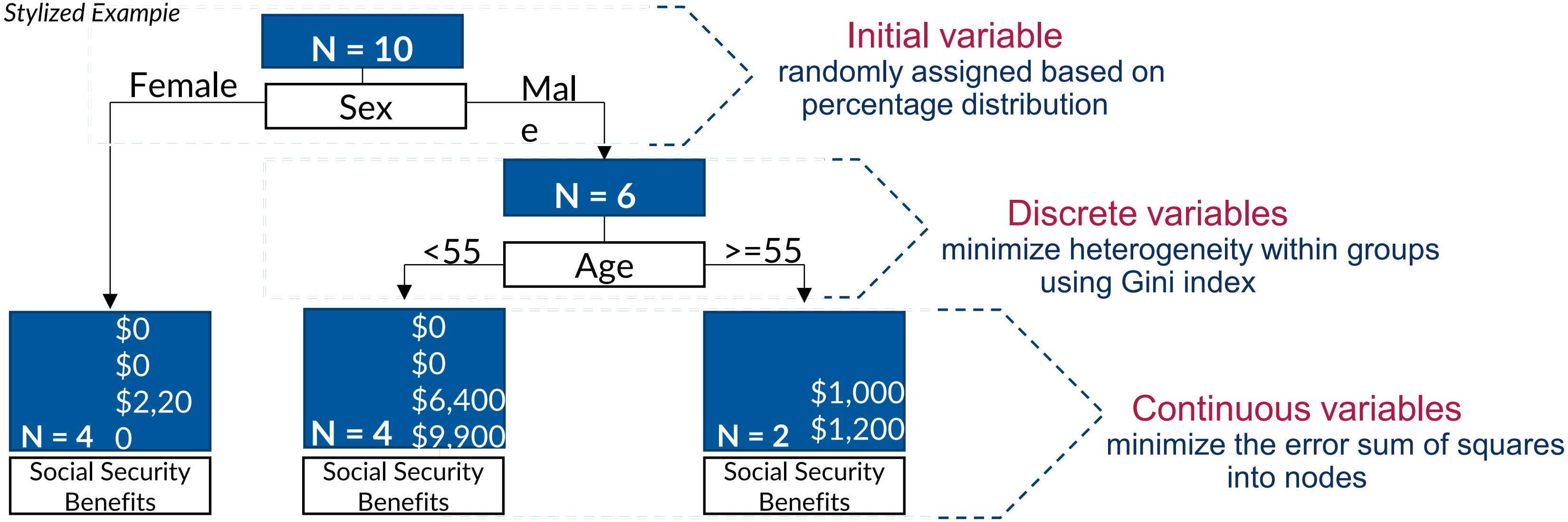
Fully Synthetic File

CART - Methodology

Synthesized non-zero records

Process:

1. Assign *gender* based on the distribution of confidential data.
2. Predict *age* conditional on *gender,* minimizing heterogeneity within groups. Then randomly select value from within those final nodes.
3. Predict *Social Security Benefits* conditional on *gender* and *age,* to minimize Sum of Square Errors.
4. Predict next highest linearly correlated variable(s) conditional on *gender, age,* and *Social Security*



Stylized Example

N = 10

Sex

Female

Mal
e

N = 6

Age

<55

>=55

N = 4

$0
$0
$2,20
0

Social Security Benefits

N = 4

$0
$0
$6,400
$9,900

Social Security Benefits

N = 2

$1,000
$1,200

Social Security Benefits

Initial variable
randomly assigned based on percentage distribution

Discrete variables
minimize heterogeneity within groups using Gini index

Continuous variables
minimize the error sum of squares into nodes

CART – Methodology, cont.

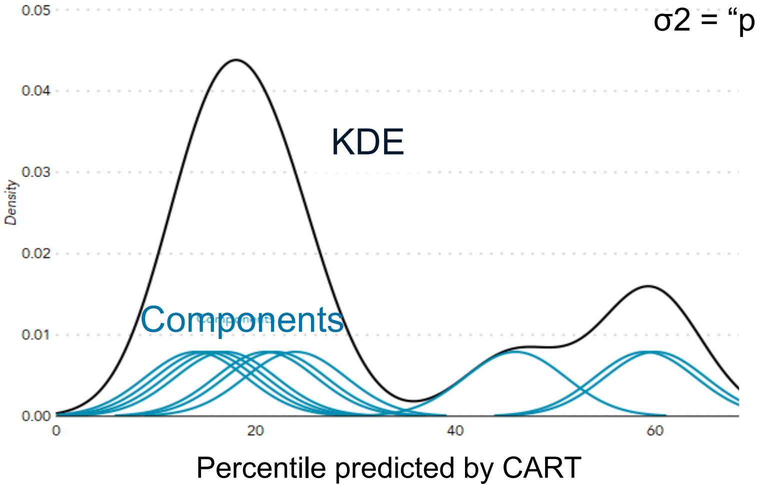
Stylized Example, cont.

Males, < 55

Obs	Value	Ntile	Optimal KDE Variance	Synthetic Value Distribution
1	$0	1st	$0	0
2	$0	1st	$0	0
3	$6,400	66th	$650	$\sim N(\mu=6{,}400, \sigma^2=650)$
4	$9,900	98th	$2,300	$\sim N(\mu=9{,}900, \sigma^2=2{,}300)$

Expanded Stylized Example

~ N(μ = sampled value,
σ2 = “percentile variance”)



Draw a value from a smoothed Kernel Density function for each percentile of values predicted by CART.

Ensuring Privacy

Imposed protocols

- Sample of 1 in 1,000 observations
- Top code age at 85
- Terminal nodes limited to 50
- Kernel Density Estimator with variance σ^2
- Run through simple tax calculator
- Round continuous variables

Validation Metrics

- **Duplicates**
- **Unique-Donors**
- **Unique-Uniques**
- **Row-wise Squared Inverse Frequency**
- **ℓ-diversity of final nodes**

Measuring Quality

Summary statistics

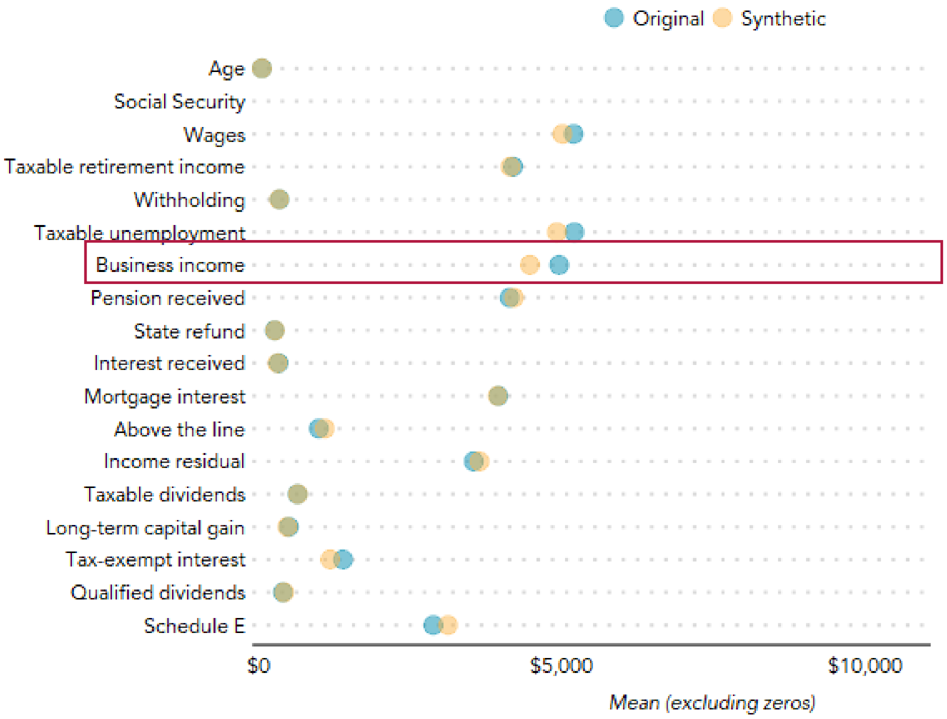
Correlation fit

Kolmogorov-Smirnov (KS) test

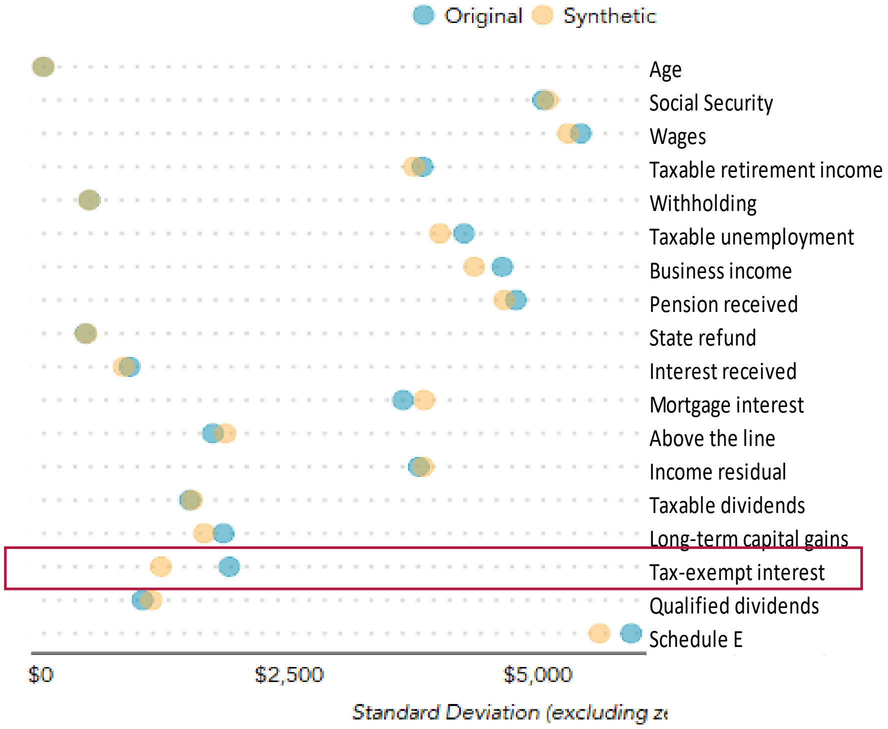
Regression confidence interval overlap

Summary statistics

Means

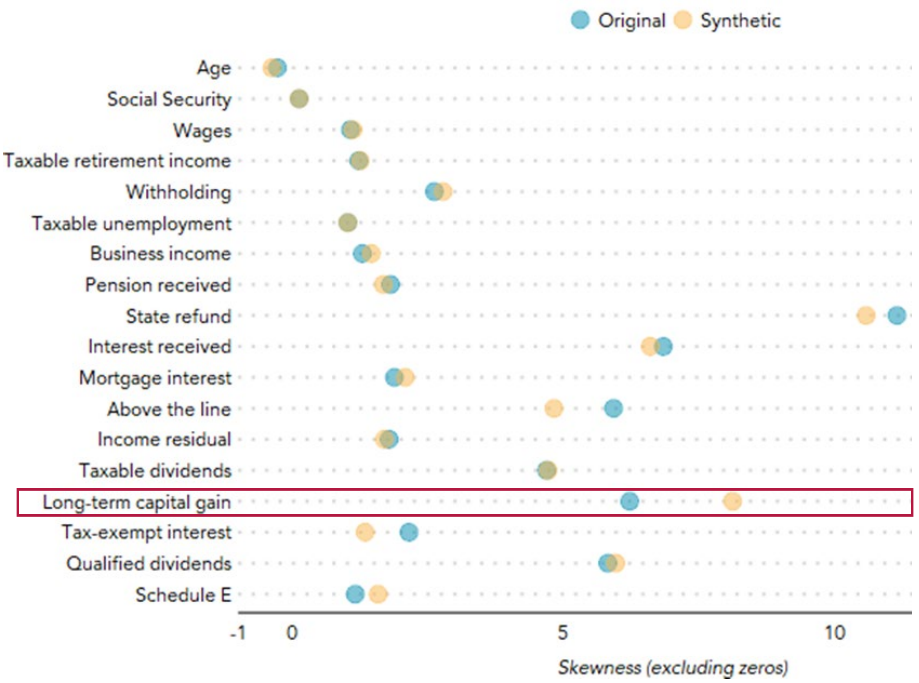


Standard deviations

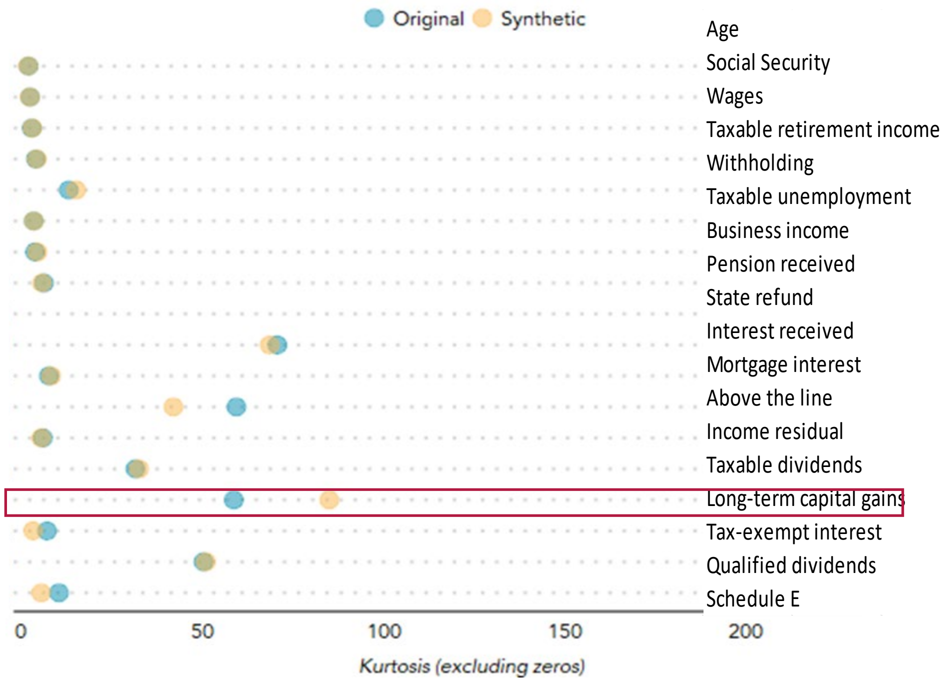


Summary statistics, cont.

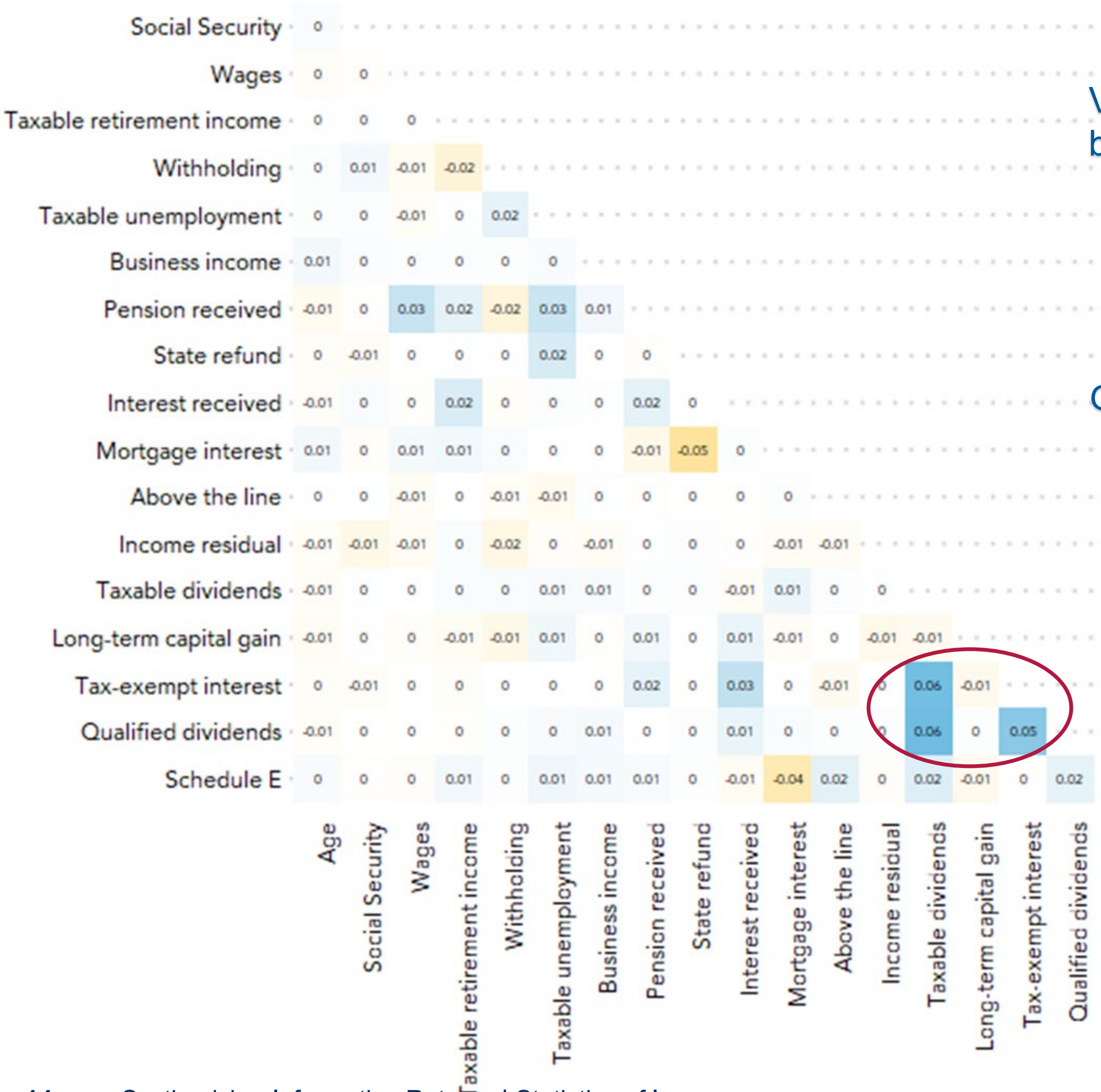
Skewness



Kurtosis



Correlation differences



	Age	Social Security	Wages	Taxable retirement income	Withholding	Taxable unemployment	Business income	Pension received	State refund	Interest received	Mortgage interest	Above the line	Income residual	Taxable dividends	Long-term capital gain	Tax-exempt interest	Qualified dividends
Social Security	0																
Wages	0	0															
Taxable retirement income	0	0	0														
Withholding	0	0.01	-0.01	-0.02													
Taxable unemployment	0	0	-0.01	0	0.02												
Business income	0.01	0	0	0	0	0											
Pension received	-0.01	0	0.03	0.02	-0.02	0.03	0.01										
State refund	0	-0.01	0	0	0	0.02	0	0									
Interest received	-0.01	0	0	0.02	0	0	0	0.02	0								
Mortgage interest	0.01	0	0.01	0.01	0	0	0	-0.01	-0.05	0							
Above the line	0	0	-0.01	0	-0.01	-0.01	0	0	0	0	0						
Income residual	-0.01	-0.01	-0.01	0	-0.02	0	-0.01	0	0	0	-0.01	-0.01					
Taxable dividends	-0.01	0	0	0	0	0.01	0.01	0	0	-0.01	0.01	0	0				
Long-term capital gain	-0.01	0	0	-0.01	-0.01	0.01	0	0.01	0	0.01	-0.01	0	-0.01	-0.01			
Tax-exempt interest	0	-0.01	0	0	0	0	0	0.02	0	0.03	0	-0.01	0	0.06	-0.01		
Qualified dividends	-0.01	0	0	0	0	0	0.01	0	0	0.01	0	0	0	0.06	0	0.05	
Schedule E	0	0	0	0.01	0	0.01	0.01	0.01	0	-0.01	-0.04	0.02	0	0.02	-0.01	0	0.02

Values are the correlation differences between every combination

Presented Difference = Synthetic Correlation − Original Correlation

Generally equal 0

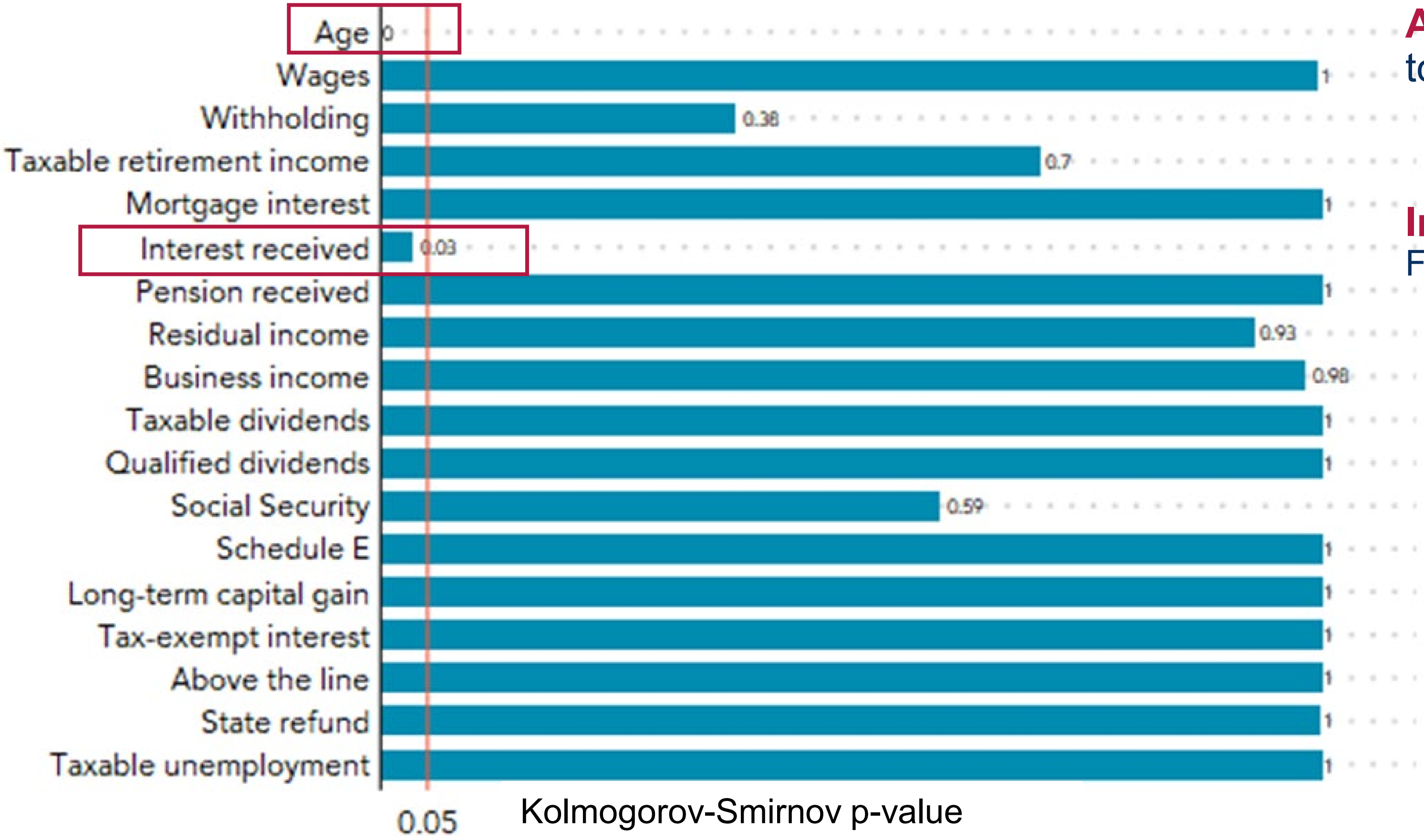
- Tax-exempt interest
- Qualified dividends

Areas for further research

Kolmogorov-Smirnov (KS) test

Purpose: Equivalence of univariate probability distributions

***H_0* =** samples come from the same underlying distribution



Age: Failed due to top coding

Interest Received: Failed due to rounding

Confidence interval overlap

Purpose: Average relative overlap between CIs for each coefficient in identical models.

Wages = f(all other vars)

Interpretation:

1 = Perfect overlap

0 = No overlap, adjacent CIs

< 0 = The distance between CIs

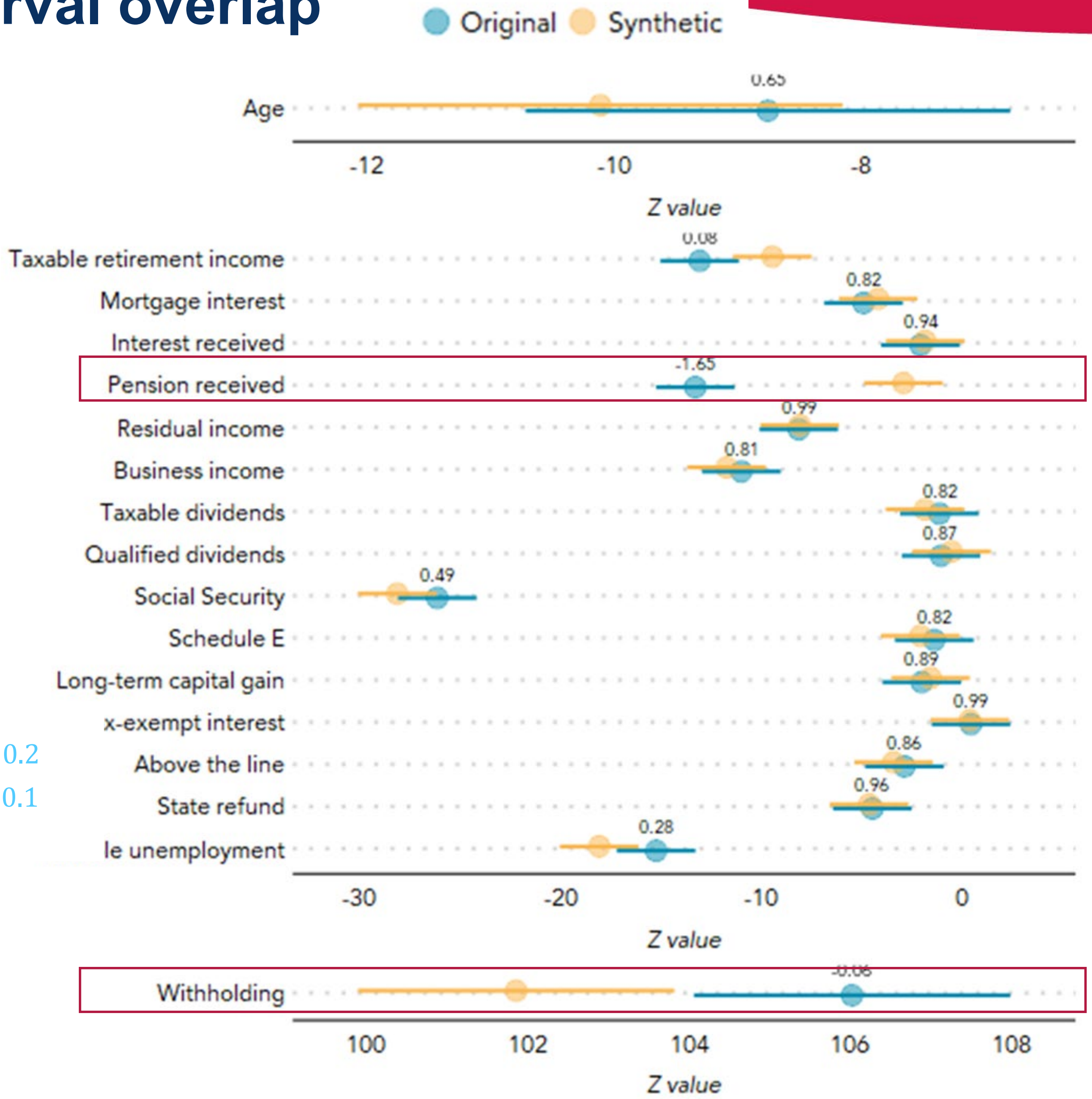
Interval Overlap:

Pension Received = -1.65

$\widehat{\beta_o} = -0.2$
$\widehat{\beta_s} = -0.1$

Withholding = -0.05

$\widehat{\beta_o} = -5.6$
$\widehat{\beta_s} = -5.1$



Thank you

Victoria.L.Bryant@irs.gov