Revenue Is A Poor Indicator For Measuring A Tax-Haven-Based Subsidiaries' Contribution To Tax Reduction

Mingyang Li

May 3, 2018

Abstract

Tax haven is the urban myth of finance world for amateurs: many people claim that setting up a subsidiary in tax haven can help reduce the overall income tax for a company, and that one can use the revenue generated in this tax haven to measure its contribution towards this tax reduction. This report validated these theories with statistical methods. In addition, as a side-product of the project, this paper constructs a comprehensive database from free public filings, which makes advanced researchers to conduct further in-depth investigation. In terms of data science techniques employed during the process, this collection procedure also serves as the key component of this practicum project of Spring 2018.

Contents

1 Background **2**
- 1.1 Social And Political Background . . . 2
- 1.2 Significance . . . 2
- 1.3 Technical Background . . . 2

2 Datasets **3**
- 2.1 On The Collection Of Subsidiary Information . . . 3
- 2.2 List of Datasets Used . . . 6

3 Evaluating Tax Effect On Multinational Corporations **6**
- 3.1 Does Having Haven Subsidiary(-ies) Actually Correlates With Effective Tax Rate? . . . 7
- 3.2 Approximating Effective Tax Rates With Expected Tax Rates . . . 8
  - 3.2.1 Defining Expected Tax Rates . . . 8
  - 3.2.2 Comparing Two Expected Tax Rates With The GETR . . . 9
  - 3.2.3 Which ExTR Leads To Conclusion We Drew From GETR? . . . 9

4 Conclusion **10**

1 Background

1.1 Social And Political Background

Many companies have overseas subsidiaries in territories referred as "tax havens". By routing cash flow to their subsidiaries, these firms can escape from higher tax rates in their home countries. Nevertheless, while setting up offshore subsidiaries is straightforward, maintaining one can be costly. This is because – as Journalist Adam Davidson points out – "following the law requires a team of lawyers and accountants to carefully monitor tax laws in dozens of countries and maintain accounts that stay on the safe side of confusing rules."[4] A more significant reason is that foreign cash holdings can not be easily send back to the headquarters. This could lead to suboptimal uses of cash include parking earnings in the form of cash or short term securities[6], and making unprofitable acquisitions in foreign countries[5]. If we consider running foreign subsidiaries as an investment, how good is the return? Alternatively, will they really help lowering the effective tax rates? If so, can we statistically measure this effect?

1.2 Significance

This article aims to shed some light on the following readers:

1. Multinational corporations not yet incorporating any subsidiary in tax havens. They want to know, "If we set up a subsidiary in a tax haven, will we actually be able to enjoy a lower tax rate?"

2. Financial organizations issuing loans to aforementioned corporations. They would like to predict, "Will my clients grow in business as expected after opening subsidiaries in tax havens, so that our loans actually can be paid back?"

3. Government departments interested in estimating the effect of offshore business to the parent company. They want to know, "Does overseas cash help companies grow within the country more, thus benefiting the nation's economy, than if the money had been held domestically?"

While we can not provide direct answers to these questions due to lack of expertise in related fields, from a data scientist's perspective, we are able to link large heterogeneous datasets and present statistical results that hopefully lead to a correct path to their answers.

1.3 Technical Background

The Securities Act of 1933, commonly known as the "truth in securities" law, aims to protect investors from misleading or deliberately hidden information about public securities for sale. To put this Act into effect, the Securities and Exchange Commission (SEC) requires U.S. companies to disclose key financial information at a regular basis. In compliance with this regulation, companies report their annual financial performance to the SEC in documents known as Form 10-K. [1] They declare subsidiaries – whether domestic or overseas – in Exhibit 21 (EX-21, for short), as part of these filings.

Well designed as the system seems, not all companies have to follow this requirement for transparency. For example, companies offering securities through employee benefit plans, only within the state it's incorporated, or in a private manner, are exempt from this Act. In addition to that, when filer companies decrease their numbers of record shareholders below certain thresholds, their filing obligations may also be suspended. Therefore, readers are advised to keep in mind that SEC is far from an ultimate source for related research.[2]

Hereafter, we follow these conventions:

Definition 1. Subsidiaries, headquarters and corporations.

A "subsidiary" is a company declared in an Exhibit 21.

[1] https://www.sec.gov/about/forms/form10-k.pdf

A "headquarter" is a 10-K filer with at least one subsidiary, denoted by h hereafter.
A "corporation" is a collection of a headquarter together with its all subsidiaries.

Now it is a good time to properly define "tax haven":

Definition 2. A tax haven is a jurisdiction that offers minimum tax liability to foreign businesses or individuals and shares no or little financial information with other authorities. [3]

A common misconception is that a tax haven is simply any country collecting a corporate tax rate lower than that of the US. In fact, few country demands a corporate tax higher than the US tax rate, making almost all countries tax havens if this constraint were true.[2] Therefore, we gathered lists of recognized tax havens from multiple sources. For simplicity, we take the first version of EU's list of non-cooperative tax jurisdictions as our list of tax havens.[1]Trivially,

Definition 3. A haven subsidiary is a subsidiary in a tax haven.

With concepts properly defined, we can continue to introducing datasets.

2 Datasets

2.1 On The Collection Of Subsidiary Information

Consumed more than half of this project's lifetime is the collection of subsidiary information, and it is those various styles of EX-21 that complicated the data extraction. EX-21s can come in either tabular or textual format, and, challengingly, their file extensions does not uniquely identify with their formats – a tabular EX-21 may come in either TXT or HTM format, and so is a textual one[3]. This discrepancy was introduced during the period of 2000~2005, when 10-K filers are encouraged to compose EX-21s in HTM. Intuitively, an HTML file is much easier to parse than a textual document (Figure 1). This is why the developers behind the CorpWatch API, who have been parsing EX-21s since 2007, decided to ignore all filings submitted before 2003.

At Wharton Research Data Services (WRDS), we wish to capture as much information as possible from what CorpWatch leaves behind. Different strategies are used to handle each type of documents:

- With tabular EX-21s, historical reasons is to blame for the chaotic situation. Decades ago, plain texts used to be the final render format for all documents, much similar to PDFs nowadays. It was natural back then for accountants to compose pure ASCII tables, where spaces between columns are literally filled with whitespace characters. Although not as well structured as HTML tables, there are program libraries capable of recognizing tables from plain texts[4]. To enhance the recall, I have developed a statistical procedure that extracts tabular data in a more generalized way.[8] Whichever file extension a EX-21 comes with, as long as it contains tables, there is a good chance we can extract them into DataFrames.

- If an EX-21 does not seem to contain a table, we can treat it as a textual document. They declare subsidiaries with natural English sentences such as "AAA Inc., is a wholly-owned subsidiary of BBB, LP in CCC.", or "As of 2018-09-09, the DDD Co., Ltd. does not have any operating subsidiary.", much more difficult to extract structuralized information from. Our best bet is to feed them into a natural language processor and let it detect all named entities present. We would expect a low precision and low accuracy.

[2]Please see Figure 3 on page 6 for a visualization of this fact.
[3]TXT is the file extension name for plain text files, while HTM – or in its full form, HTML – stands for Hypertext Markup Language (HTML).
[4]For those who are curious, an example would be `astropy.io.ascii`.

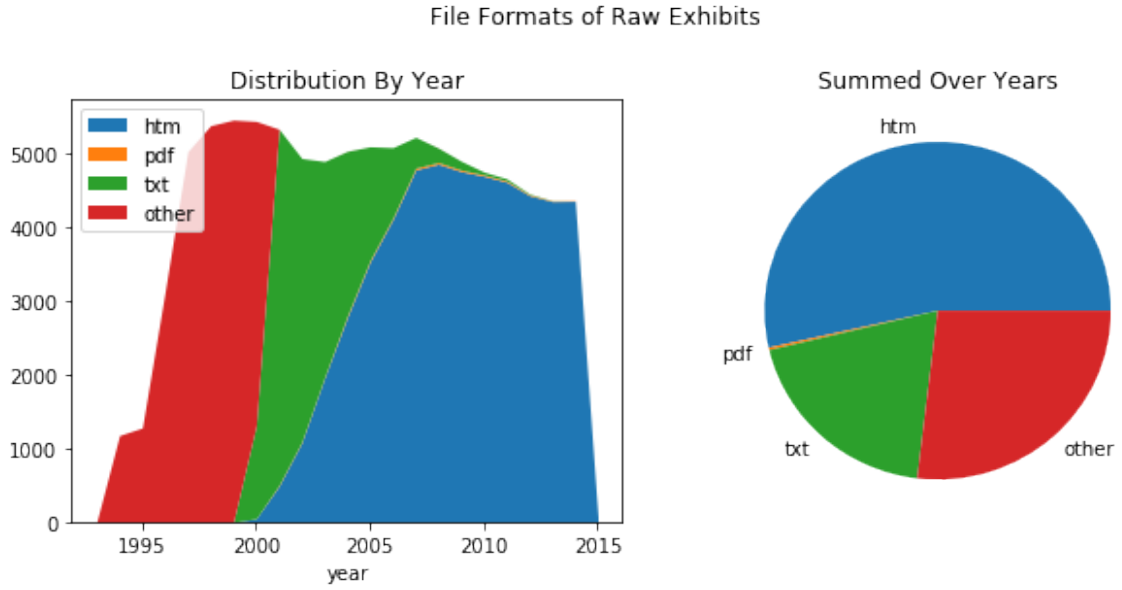

Figure 1: **File formats of raw exhibits.** Please note that this statistic is solely based on the extension names of 10-K files. Due to historical reasons, "other" is in fact an representation of `txt` files. It can be easily observed that, until the year 2000, when filers started to adopt the `htm` formats, EDGAR did not think it was necessary to specify file extensions in their data storage.

After tabular DataFrames are extracted, we would like to detect headers. Usually table extractors can automatically recognize headers if header separators exist (usually a horizontal line), but in corner cases, we need to incorporate external knowledge about the tables. For example, in any column, no matter whether the data is numerical, the column name is almost always a string. We can therefore keep cutting off the first rows, monitoring how the "purity of datatypes" change. The purity is usually maximized within 5 rows, the latter of which composes the header of this table. Presence of one numerically-typed column is sufficient for determining the header of a table. This statistical approach to header detection is further described in my blog post.[7]

Headers detected should be conformed to standard column names. Columns of interest include "subsidiary name", "parent", "voting percentage" and "jurisdiction". For instance, "subsidiaries of filer", "name of subsidiaries" and "SUBSIDIARY NAME" should all be replaced with "subsidiary name". The decision to assign which standard name to a column is based on a majority vote of three classifiers:

1. a editing distance minimizer that finds the closest standard name to the raw name specified in the header detected,
2. a tf-idf + SVM classifier trained on the content of manually labeled columns, and
3. a repetitiveness calculator that
   (a) looks at all columns who would otherwise be labeled as another "subsidiary name" and
   (b) assigns the "parent" label to the most repetitive column which would otherwise be labeled as column.

The first classifier may give up its voting privilege if there is no raw header detected, and, similarly, the third can also waive its vote if there are less than two columns that would be labeled as subsidiary names.

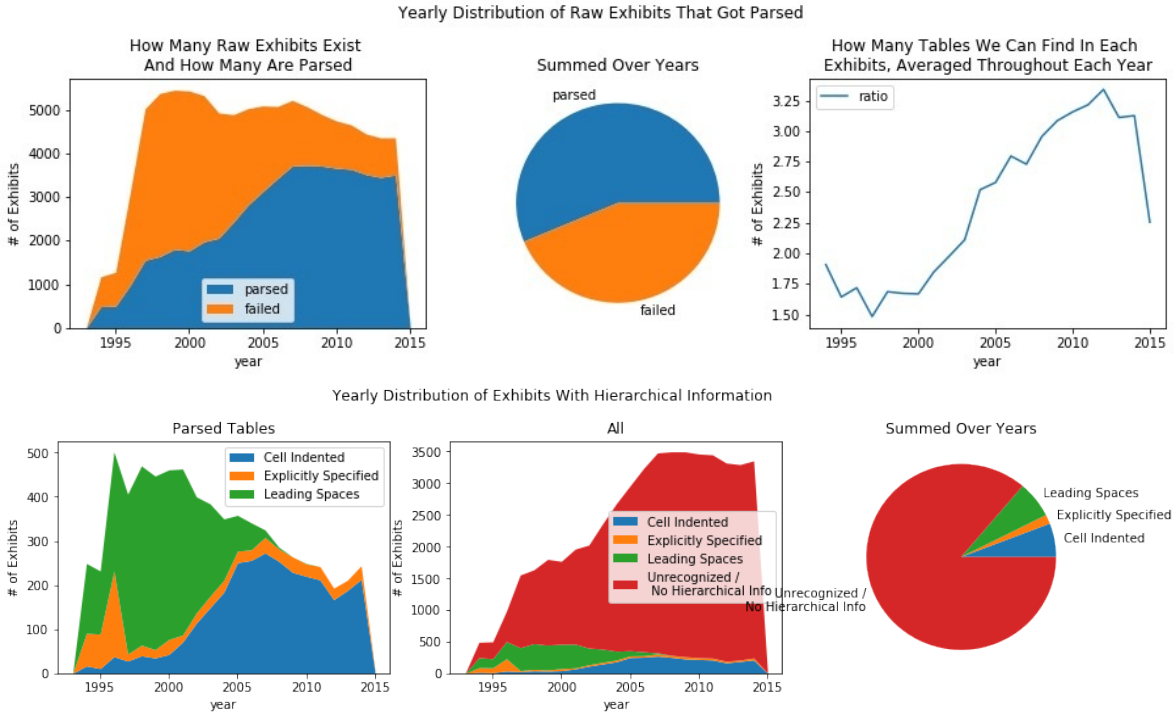

Figure 2: Yearly distribution of exhibits from which tables can be extracted, together with a yearly distribution of exhibits with hierarchical information.

As a last step, we may extract hierarchical information from the DataFrames. The hierarchy may be encoded in three different ways:

1. Explicitly specifying the direct-parental company for each subsidiary with a separate "parent" column,
2. Indenting with whitespaces in the subsidiary name, or with empty cells in the beginning of each row, resembling a tree structure, or
3. Embedding sub-tables immediately after introducing a subsidiary with a row, or simply as a separate table.

Lacking reliable method to detect nested tables, I have only implemented parsers for the first two representations of hierarchies. Shown in Figure 2, my parser only extracted a small portion of EX-21s. The chance is slight to extract adequate amount of hierichical information while maintaining a flexible program design. Hence, I decided to use CorpWatch's database dump for later analysis[5]. Written in Perl, CorpWatch's parser makes excessive use of regular expressions and SQL manipulations. This techniques ensured a high quality of their parsed data, but also limited the recall[6]. Considering that the datasets we are going to use later contain few entries preceding 2003, using CorpWatch is not very devastating option.

[5]The database dump is freely available at `http://api.corpwatch.org`.

[6]I tried to run their last public version of parser program on the EDGAR dump in WRDS servers, only to find out that the hierichical information are not extracted. CorpWatch developers explained that, due to their mixed writeup of production code together with their credential information, they are struggling to publish their latest working code without leaking their passwords.

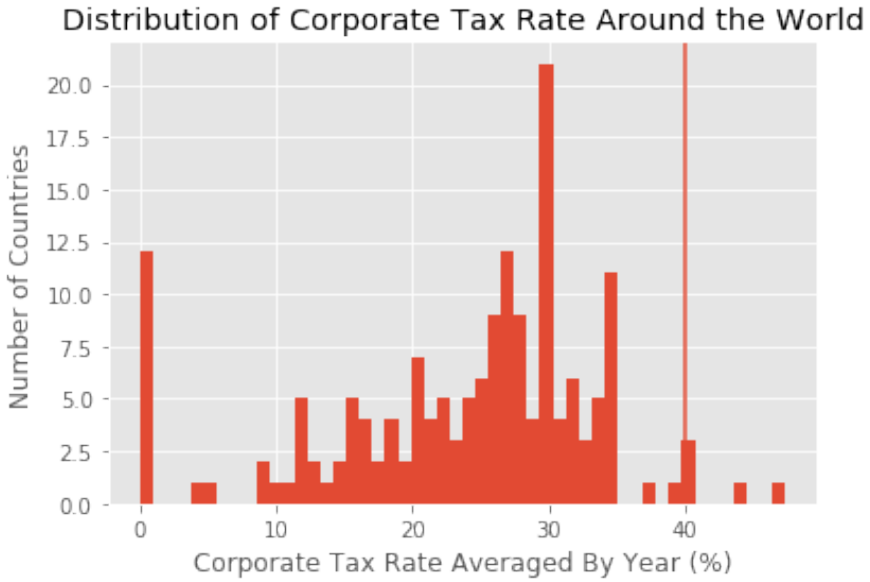

Figure 3: Histogram of corporate tax rates around the world.

2.2 List of Datasets Used

- **CorpWatch**: Hierarchical subsidiary information about US companies.[7]
  - Provides mapping SubsidOf (h, y): a list of subsidiaries of h in the year y.
  - Provides mapping Juri (c): jurisdiction of the company c.
- **TaxRates**: A mixed-source table mapping jurisdiction and year to corporate tax rate.
  - Provides mapping TaxRate (j, y): saturated corporate tax rate in the jurisdiction j in the year y. It is a real value $\in [0, 1)$.
- **CCM**: CRSP/Compustat Merged Database.
  - Provides mapping Revenue (h, y, j): revenue that the corporation led by the headquarter h generated in the jurisdiction j in the year y.
  - Provides mapping GETR (h, y): the GETR (defined later) paid by the headquarter $h$in the year y.

3 Evaluating Tax Effect On Multinational Corporations

As clarified in Definition 2, lower tax rate does not always mean a jurisdiction is a tax haven. A tax haven can provide favorable taxation via policies that one single float number can not adequately represent. Therefore, a observation-based metric may better suit our purpose. Namely, we will be using an effective tax rate.

Following Cen et al. (2017), the GAAP effect tax rate (GETR) is defined as

$$\text{GETR} = \frac{\text{Total Income Taxes}}{\text{Pretax Income}}.$$

[7] More Information: `http://api.corpwatch.org/documentation/faq.html`.

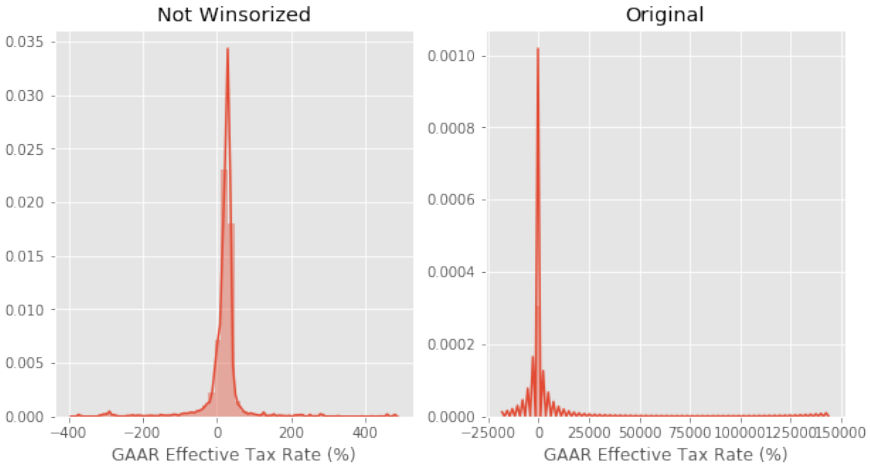

Figure 4: Comparison of the GETR data before and after 1% winsorization.

This metric, however, is prone to measure issues, such as mergers and acquisitions (M&As) and data standardization, that significantly alters the tax paid in a given year. Therefore, outlier removal is a crucial component in our data cleaning procedure. A common practice in the field of finance is called winsorization, where extreme values outside the 1% percentile is suppressed to the 1% and the 99% points. Shown in Figure 4 is two histograms of GETR before and after applying winsorization at the 1% percentile. Although extreme values have been suppressed, spurious outliers still exist outside the $(0, 100\%)$ region. Other than winsorization, one more fix we have to consider is the fiscal year window – It is the tax rates audited and imposed on year $t-1$ that corresponds to the revenue generated at year t. Therefore, we shift all years by -1 for our GETR data. With these two fixes applied, our data can finally be used for analysis.

3.1 Does Having Haven Subsidiary(-ies) Actually Correlates With Effective Tax Rate?

The most basic question in this research is to test whether having a subsidiary in a tax haven actually lowers the effective tax rate imposed on the corporation. This can be done with a t-test on the GETR. With a t-statistic of -2.255 accompanied by a p-value of 0.024, GETR reports that having a subsidiary company in a tax haven indeed reduces the effective tax rate. This can also be visualized by a two-series histogram on GETR, as shown in Figure 5.

In our dataset, we captured more corporations having haven subsidiaries than those without one, measuring at 5401 and 683, respectively. This is reflected on Figure 5: Notice that the kernel density estimation (KDE) curve of all multinational corporations (purple) almost overlaps with that of corporations with haven subsidiaries (red). Scholars may question the validity of our argument based on the bias in our samples.

A Wilcoxon Rank Sum Test is performed to overcome this problem. To prepare for Wilcoxon, we split our dataset into two sections, one holding information for corporations on those years when they did not have a haven subsidiary, and one for the years when they had one. Indices of two datasets are matched on the corporation ID to ensure pairwise comparability. That is a 738-record dataset for Wilcoxon, each entry carrying the following information:

- company identifier,
- a year when it had no haven subsidiary,
- a year when it had at least one subsidiary, and

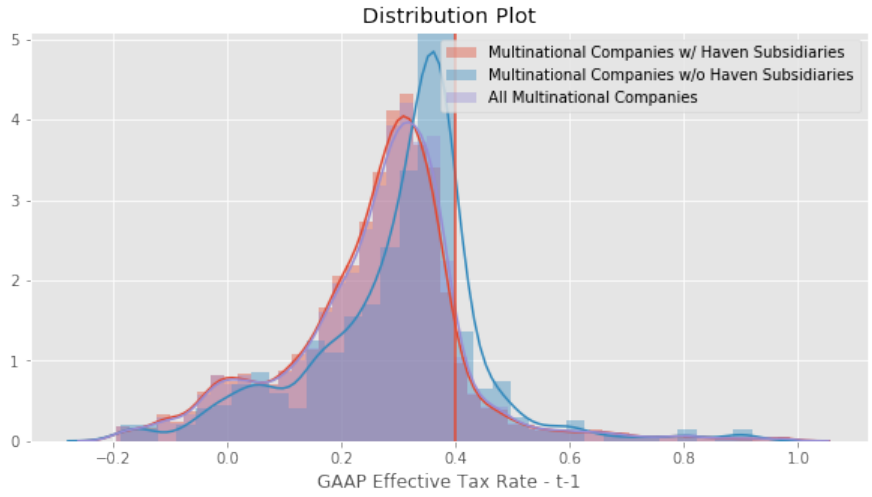

Figure 5: Histogram of GETR, plotted as two separate series with respect to whether a corporation has haven subsidiary.

- GETRs of these two years.

The Wilcoxon test is performed on the GETR values, with all other fields considered as data identifier. This non-parametric test yields a 125243.0 at a p-value of 0.055, confirming to the t-test conclusion that GETR is correlated with the dummy variable of whether haven subsidiary is incorporated or not.

3.2 Approximating Effective Tax Rates With Expected Tax Rates

Considering the outrageous outliers we encountered in GETR, we want to see if we can approximate this measured data with variables derived from knowledge *a priori*. We would call these variables "expected tax rates" (ExTRs, for short), on contrast to "effective tax rates" such as GETR.

3.2.1 Defining Expected Tax Rates

For starters, a naively-averaged ExTR is calculated. This is done by merging **CorpWatch** with **TaxRates** on year and residing jurisdiction, grouping by headquarter and year, and then calculating the average tax rates.[8]

Remark 4. **Removing purely domestic corporations.** In Figure 3, most countries demonstrated corporate tax rates lower than the US. Since we investigate with a country-level granularity without distinguishing subdivisions (i.e. states and provinces), expected tax rates for any purely domestic corporation would be 0.4, providing little information for our tax haven research. Therefore, unless otherwise stated, purely domestic corporations are dropped from our datasets.

Intuitively,

$$\text{avg\_taxRate}\left(h,y\right) \coloneqq \frac{1}{\left|\text{SubsidOf}\left(h,y\right)\right|} \sum_{i \in \text{SubsidOf}(h,y)} \text{TaxRate}\left(\text{Juri}\left(i\right),y\right)$$

where h is the headquarter identifier and y is the year.

However, one may argue that subsidiaries of distinctive scales should contribute unevenly to the overall tax rate of their corporation. To make up for this, we also invented the revenue-weighted average tax rate:

[8] Therefore, we use "expected tax rate" and "average tax rate" exchangeably.

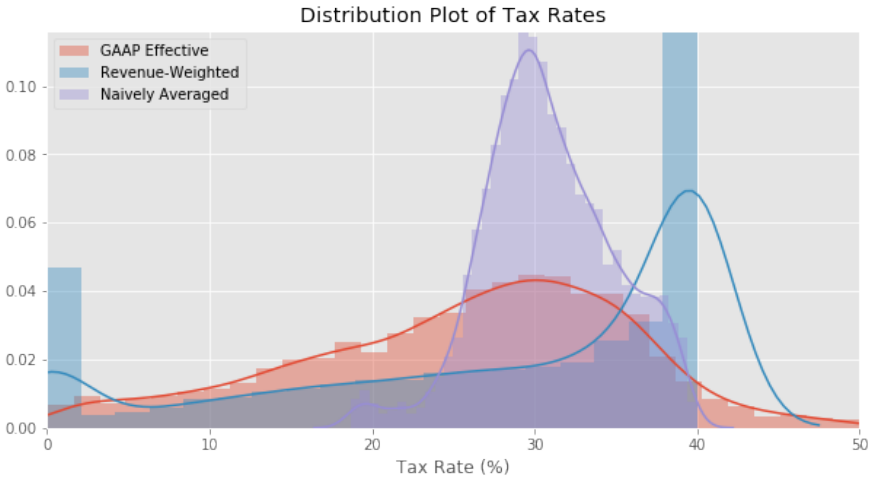

Figure 6: Histogram of three types of tax rates in the region of $(0, 40\%)$.

$$\text{wtAvgTaxRate}\left(h,y\right) := \frac{\sum_{i\in\text{SubsidOf}(h,y)} \text{Revenue}\left(h,y,\text{Juri}\left(i\right)\right)\cdot\text{TaxRate}\left(\text{Juri}\left(i\right),y\right)}{\sum_{i\in\text{SubsidOf}(h,y)} \text{Revenue}\left(h,y,\text{Juri}\left(i\right)\right)}.$$

The hypothesis accompanying this new metric is:

Proposition 5. *wtAvgTaxRate* (h, y) *represents actual tax rates better than* $avg\_taxRate\,(h, y)$.

3.2.2 Comparing Two Expected Tax Rates With The GETR

Plotted in the common region of $(0, 40\%)$ (where 40% is the corporate tax rate in the US, averaged by year), Figure 6 shows that the revenue-averaged ExTR captures near-linear increase in GETR distribution at $< 20\%$ tax rate better than the naively-averaged ExTR, while the latter captures better the peak around 30% tax rate. A combination of these two ExTRs may need to be considered to accurately model GETR, but it would be beyond the scope of this paper.

3.2.3 Which ExTR Leads To Conclusion We Drew From GETR?

Earlier, we have found that opening business in tax haven can affect a corporation's expected tax rate. Here, we want to put the two made-up ExTRs to test, and see if any of them can lead us to a similar conclusion.

As a first step, a two-series distribution plot for each variable can be a good start. A distribution can be visualized by plotting the histogram of average tax rates as two separate series, one for corporations with haven subsidiaries and one for those without one. These histograms are shown in Figure 7.

From a glance, these two average tax rates behave completely differently. While the naively-averaged ExTR suggests that corporations with a haven subsidiary tend to be lightly taxed, the revenue-averaged ExTR displays no noticeable difference. By performing t-tests on each of these two average tax rates, this discrepancy in conclusions can be confirmed (Table 1). With a p-value around 11%, the revenue-based average provides little proof that having haven subsidiary would matter, while the naively-averaged tax rate suggests the opposite. The fact that Naively-Averaged ExTR aligns better with the actual GETR suggested that revenue is a poor factor to consider when measuring haven subsidiaries' contribution to tax rate reductions. This can be explained by realizing that tax havens are usually tiny-sized countries, so that revenues should pay a little role in the correlation.

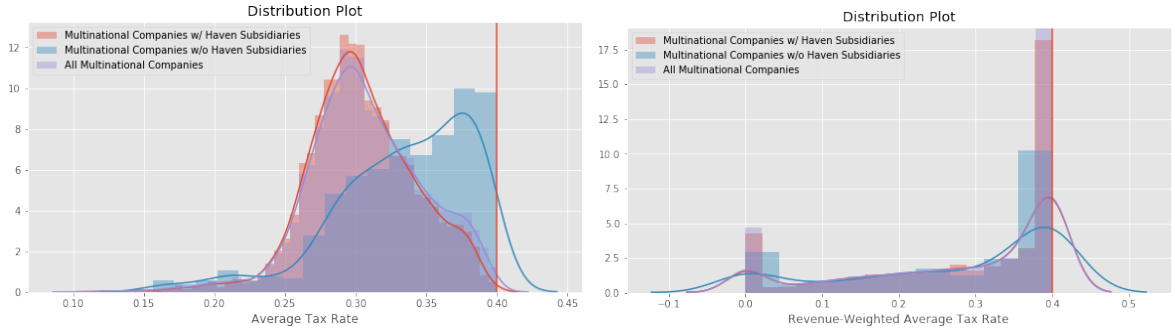

Figure 7: Histogram of average tax rates.

	t-statistic	p-value
Revenue-Weighted Average Tax Rate	1.594550	1.111880e-01
Naively-Averaged Tax Rate	-15.681695	1.722724e-48

Table 1: T-test results of average (expected) tax rates.

4 Conclusion

In this paper, we demonstrated that (1) having haven subsidiaries can actually help a multinational corporation reducing its income taxes, and that (2) revenue is a poor indicator for measuring a haven subsidiaries contribution to this reduction effect. As a disclaimer, this project is presented solely as a summary for my practicum purposes, without any implication of guaranteeing accuracy and/or applicability in a practical sense. I do, however, hope this could be of support to peers who plan to investigate further in related fields of research.

References

[1] Common EU list of third country jurisdictions for tax purposes - Taxation and Customs Union - European Commission.

[2] SEC.gov | Information About Some Companies Not Available From the SEC.

[3] Salome Chelangat. What is a Tax Haven?

[4] Adam Davidson. My Big Fat Belizean, Singaporean Bank Account. *The New York Times*, July 2012.

[5] Alexander Edwards, Todd Kravet, and Ryan Wilson. Trapped Cash and the Profitability of Foreign Acquisitions. *Contemporary Accounting Research*, 33(1):44–77, March 2016.

[6] C. Fritz Foley, Jay C. Hartzell, Sheridan Titman, and Garry Twite. Why do firms hold so much cash? A tax-based explanation. Working Paper 12649, National Bureau of Economic Research, October 2006.

[7] Mingyang Li. Detect Headers in CSV Tables Statistically, January 2018.

[8] Mingyang Li. Extracting Tables From Plain Text Files Statistically With Numpy, January 2018.