eJournal of Tax Research

Volume 9, Number 1 July 2011

CONTENTS

Atax

UNSW THE UNIVERSITY OF NEW SOUTH WALES SYDNEY • AUSTRALIA

ISSN 1448-2398

eJournal of Tax Research (2011) vol. 9, no. 1, pp. 116-137

Towards Effective and Efficient Identification of Potential Tax Agent Compliance Risk: A Stratified Random Sampling Approach

Ying Yang, Esther Ge, Ross Barns*

Abstract

We propose to use a stratified random sampling approach to identify whether a tax agent's return preparation behaviour is significantly different from its industry norm. Given a tax agent *T A*, our approach creates a statistically sufficient number of notional peers for it. These peers comprise a reference group for *T A*, and the expectation for *T A*'s tax return behaviour can be derived there from. By comparing T A's actual behaviour against its expected behaviour, one can infer whether *T A* behaves abnormally and to what degree T A incurs potential compliance risk. The novelty and advantage of our approach includes (1) effective and efficient risk identification, (2) an easy-to-understand methodology, (3) easy-to-explain results, (4) no need for any pre-defined threshold values and hence less able to be undermined by "game players" who seek to make claims just under the threshold, and (5) low cost of identification as our approach conducts unsupervised learning that does not demand a supply of labelled tax agents[1] as training data.

1. INTRODUCTION

Individual income tax is a major revenue source for the Australian government. Over 72% of individual taxpayers choose to lodge their income tax returns via tax agents (also known as tax practitioners). Therefore it is of significant importance for the Australian Taxation Office to promote voluntary compliance with tax laws, and meanwhile identify and deter non-compliance behaviours in the tax agent industry. Successfully doing so will help protect government revenue and maintain community confidence in the Tax Office's administration of Australia's Taxation system.

However, it is a nontrivial task to accomplish effective and efficient identification of tax agent compliance risk. The challenges are imposed by the large number of tax agents in operation, the large number of individual tax returns lodged by tax agents, and the fact that the tax agent client bases are immensely diversified. Currently there are over 20,000 tax agents handling about 12 million individual tax returns per year in Australia.

* The authors are, respectively, a Senior Data Miner, a Data Miner, and the National Director of Risk and Information Management Services, Micro Enterprises and Individuals, Australian Taxation Office

[1] A labelled tax agent is one that has been classified as compliant or non-compliant by a tax audit. The course of delivering such a verdict is often an expensive and time-consuming process.

A definitive solution to tax agent compliance risk identification is to check every single tax return lodged by every single tax agent and then reach a conclusive statement. However such a solution is neither practical nor sustainable due to resource constraints. As such the Australian Taxation Office business model is founded on a risk management basis, and applies a range of defensible approaches to analyse tax return preparation behaviour of taxpayers and tax agents.

Particularly in this paper we propose a novel defensible solution that is able to deliver effective and efficient identification of tax agent compliance risk. Given a tax agent T A, our approach uses stratified random sampling to create a statistically sufficient number of notional peers for T A. These peers comprise a reference group for T A and the expectation for T A's tax return behaviour can be derived therefrom. By comparing T A's actual behaviour against its expected behaviour, one can infer whether T A behaves abnormally and to what degree T A incurs potential compliance risk.

As a matter of demonstration convenience and without losing generality, this paper examines a tax agent's compliance risk in terms of rental behaviours. We assume that the tax agent behaviours are gross income and gross expense of residential rental properties lodged by tax agents on behalf of their clients. We also assume that rental gross income and expense are affected by rental location. For a rental property, gross income is the rent that landlords receive from tenants. Gross expense is the total cost that landlords incur in order to derive the rent, including bank loan interest, capital works (such as repairs and maintenance) and other expenses (such as council rates). A landlord should return rental income, and meanwhile can claim a deduction for rental expenses incurred in deriving the rental income. For the sake of simplicity, we adopt the term “a tax agent's rental properties” as a shortcut reference to the rental properties owned by this tax agent's clients and lodged in individual income tax returns by this tax agent on behalf of its clients.

It is very important to note that our paper aims at providing a generic framework for agent risk identification, and the above assumptions are made purely for illustration purpose. When reimplementing our method in their fields, researchers and practitioners should substitute their proper domain knowledge for our assumptions in order to better suit their own applications. For instance, if one deems an agent's behaviour of interest is affected by clients' jobs and incomes, the notional peers should be created with regard to client job categories and income ranges.

Nonetheless, our proposed theoretical foundation and practical methodology (such as how to obtain peers by stratified random sampling, how to calculate z-score, how to calculate risk score, how to illustrate identification results, how to explain those results and how to avoid technical pitfalls) will stay the same.

The rest of this paper is organised as follows. Section 2 introduces the definition of peers for a tax agent and proposes how to create the peers. Section 3 explains how to compare a tax agent against its peers in order to evaluate its potential compliance risk. Section 4 applies our method to the Australia national tax agent data and illustrates the risk identification results. Section 5 highlights some technical issues to help peer researchers and practitioners circumvent possible pitfalls when re-implementing our method in their fields. Section 6 presents related work. Section 7 gives concluding remarks.

2. HOW TO CREATE PEERS FOR A TAX AGENT

Given a tax agent T A, our approach creates a statistically sufficient number of peers for T A. These peers comprise a reference group (the industry norm) against which T A is compared. This section first introduces the definition of a peer and then proposes how to create peers.

2.1 Definition of a peer

For a tax agent T A, a peer needs to satisfy the following two criteria.

(a) Each peer should have the same number of rental properties as T A does. Note that only those rental properties lodged by actual tax agents are included in a peer's property base.

(b) Since rental gross income and expense are affected by rental location[2], each peer's rental properties should have the same location distribution as T A's. Particularly in this paper, we use postcodes to indicate locations.

Thus each peer is a notional (rather than an actual) tax agent, but its property base is composed of real rental properties. We choose to use notional peers rather than actual tax agents to form a reference group for T A because in reality every actual tax agent has its unique client base, and the client bases across different tax agents are immensely diversied and are often incommensurable. Thus it is often a problem to measure whether an actual tax agent is similar enough to T A to qualify for being T A's peer. Such a problem is usually no less complicated than the risk identification problem itself.

2.2 How to create a peer

We create a notional peer by stratified random sampling with replacement. The sampling is stratified because it keeps the geographic distribution of a sampled population (a peer's rental properties) equal to the distribution of the original population (the actual tax agent's rental properties).

For example, assume T A's rental properties distribute as in Table 1, which shows that T A has 33, 21, 18 and 12 rental properties in Postcode 3048, 3064, 3000 and 3029 respectively. Postcode 3048, 3064, 3000 and 3029 each in total have 509, 1475, 9734 and 2303 rental properties lodged by various tax agents.

According to Table 1, in order to create a peer for T A

(1) we randomly pick 33 rental properties from the total 509 ones in Postcode 3048, 21 rental properties from the total 1475 ones in Postcode Postcode 3064, 18 rental properties from the total 9734 ones in Postcode 3000,and 12 rental properties from the total 2303 ones in Postcode 3029;

(2) the resulting 84 (=33+21+18+12) picked properties comprise the property base of the notional peer;

[2] As explained in Section 1, such an assumption is for illustration purpose only and can be changed for different behaviours according to appropriate domain knowledge.

(3) we replace the picked properties back to their respective suburbs to get ready for the next sampling.[3]

As a result, this peer has 84 properties that follow the same geographic distribution as T A's total 84 properties.

Note that to create a peer for T A, T A's rental properties (together with other actual tax agents') are also included for the purpose of sampling.

Postcode	No. of Rental Properties Lodged by Tax Agent(s)	
	By T A	By all Tax Agents
3048	33	509
3064	21	1475
3000	18	9734
3029	12	2303

TABLE 1: THE GEOGRAPHIC DISTRIBUTION OF T A'S RENTAL PROPERTIES

2.3 How to create many peers

To create a second peer for T A, we can repeat the above three steps in Section 2.2, then repeat them again to create a third peer, and so on and so forth until we have a statistically sufficient number of peers to form a reference group. As a rule of thumb, 1000 peers is usually statistically sufficient to present the industry norm [1, 2]. However, if computing power and time allows, the more peers the better. Hence in this particular study, we use 10,000 peers.

Because of the randomness of the sampling procedure, every peer will have a different property base that might overlap somewhat with other peers' bases, and seldom will the same property base be created twice.[4] Thus, the 10,000 peers offer a spectrum to describe how diversified a tax agent can behave if lodging rental properties similar to T A's. From such a spectrum we can find out whether T A's behaviour is abnormal, and if yes, how much potential risk it possesses.

---

[3] Such a procedure is called sampling with replacement. In theory, one can do sampling without replacement as well. But we prefer the former to the latter because of the following two reasons: (1) Sampling with replacement ensures that every property has the same probability to be included into any peer's base. (2) Sampling with replacement ensures that every peer is independent of each other. This contrasts to sampling without replacement, where the current peer's property base depends on what properties have not been picked up for previous peers' bases. The detailed mathematical explanations are beyond the scope of this paper.

[4] In theory, the probability of the same property base appearing twice equals to:

$$= \frac{1}{C^{33}_{509}} \times \frac{1}{C^{21}_{1475}} \times \frac{1}{C^{18}_{9734}} \times \frac{1}{C^{12}_{2303}}$$

$$= \frac{1}{8.4 \times 10^{82}} \times \frac{1}{5.9 \times 10^{46}} \times \frac{1}{9.5 \times 10^{55}} \times \frac{1}{4.5 \times 10^{32}}$$

$$= 47 \times 10^{-187}$$

where C^r_n is the maths symbol indicating how many different combinations one can have if selecting r items from n items.

3. How to Evaluate a Tax Agent's Potential Compliance Risk

We evaluate an actual tax agent T A's potential compliance risk by comparing T A against its notional peers.

3.1 The normal distribution

Since T A's peers are created by random sampling with replacement and with stratification according to T A's rental properties' postcodes, all the peers are equal-size random samples from the same population. According to the central limit theorem [5], the mean rental gross income values of the peers will follow a normal distribution. Likewise, the mean rental gross expense values of the peers will also follow a normal distribution, as illustrated in Figure 1.

3.2 The z-score

Because of the normal distribution, it is appropriate to use the z-score to measure whether T A's rental gross income or expense is abnormal:

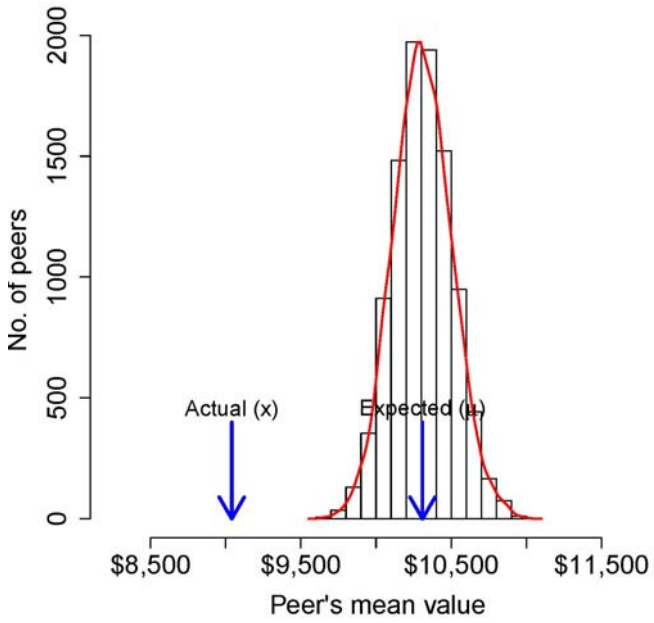
$$z = \frac{x - \mu}{\sigma}, \qquad (1)$$

where corresponding to Figure 1

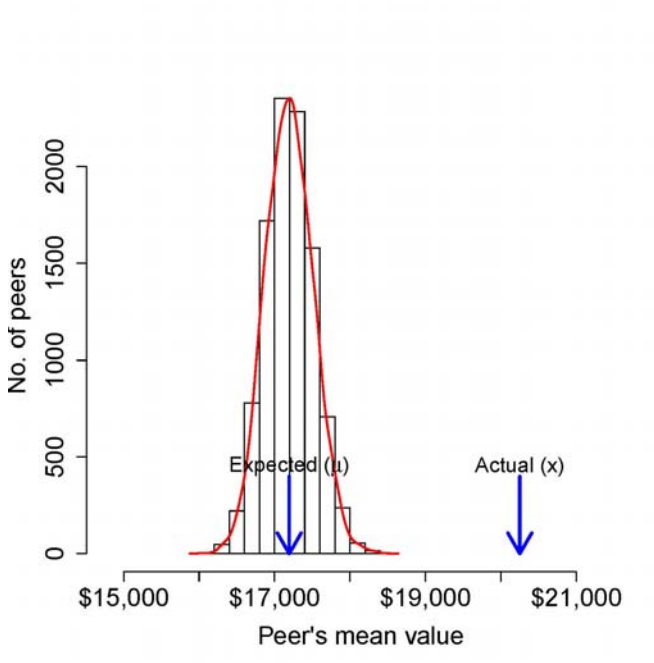
- x is T A's actual mean rental gross income (or expense);
- μ is the average value across all 10,000 peers' mean rental gross income (or expense) and hence is the expected value for T A;
- σ is the standard deviation of the 10,000 peers' mean rental gross income (or expense).

Thus a z-score is a standardised version of the raw difference between T A's actual and expected values ($x - \mu$). It tells us how many counts of standard deviations T A's actual value falls away from the average value of its peers, and in which direction [5]. T A's z-score is positive if its value is bigger than the peers' average, and negative otherwise. For example, the z-score according to Figure 1(a) will be negative because $x < \mu$, which indicates that the tax agent declares less rental gross income than expected. The z-score according to Figure 1(b) will be positive because $x < \mu$, which indicates that the tax agent claims more rental gross expense than expected.

It is important to emphasise that to measure the difference between T A's actual and expected values, the standardised difference (z-score) be more statistically sound than the raw difference ($x - \mu$). It is because the former not only takes into consideration the raw difference, but also the diversity of the peers from which the expected value (μ) and thus the raw difference are drawn. For instance, as illustrated in Figures 2(a) and 2(b), although Tax Agents A and B are equal in terms of ($x - \mu$), A's anomaly is more significant than B's due to the fact that A's peers are tightly around the expected value while B's peers are more diversified and loose.

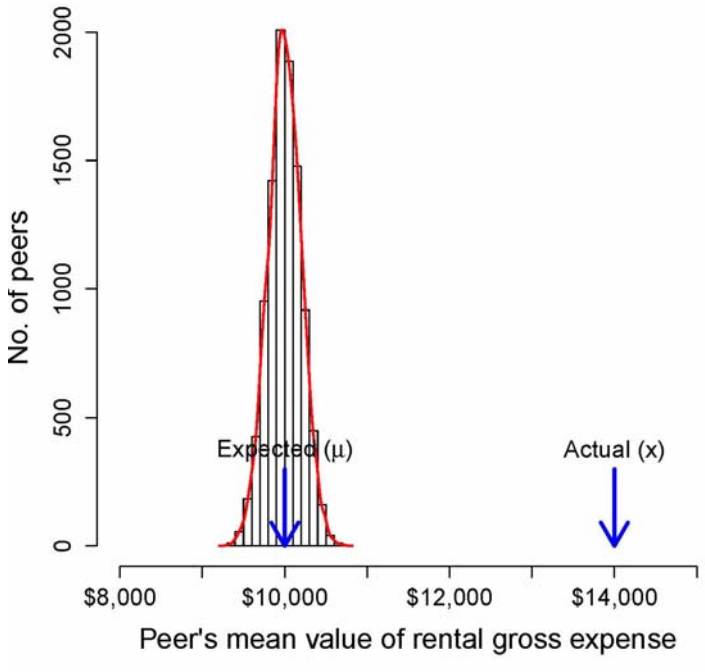


(a) Rental gross income

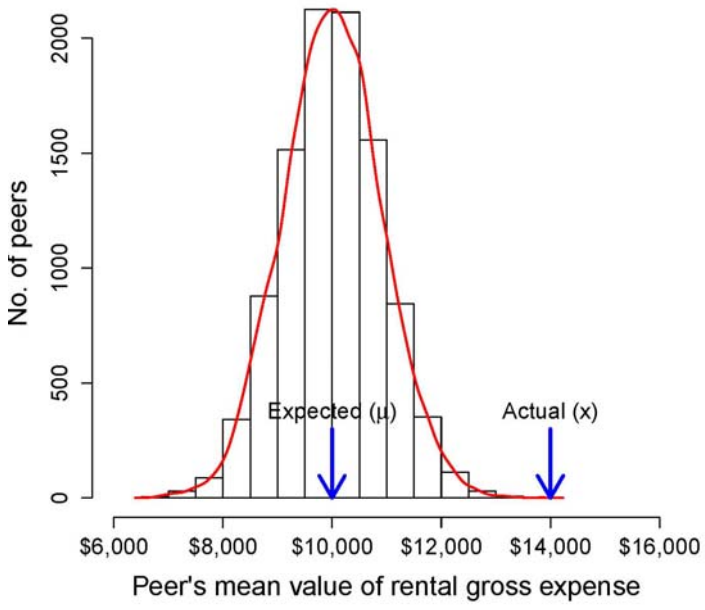


(b) Rental gross expense

FIGURE 1: Due to random sampling, the mean rental gross income or expense values of the peers will follow a normal distribution respectively.



(a) Tax Agent A



(b) Tax Agent B

FIGURE 2: To measure the difference between T A's actual and expected values, one should use the standardised difference (z-score) rather than the raw difference (x - μ)

3.3 The risk score

The risk score combines both the risk of underreporting rental gross income (z-score(income)) and the risk of overclaiming rental gross expense (z-score(expense)). Because a z-score is a standardised value that calculates how many counts of standard deviations the actual value of a tax agent falls away from the average value of its peers, z-score(income) and z-score(expense) are commensurate and hence we can apply mathematical operations on them to calculate the risk score. For T A we can calculate its z-score of rental gross income, z-score(income), as well as its z-score of rental expense, z-score(expense). The lower the value of z-score(income), the less the rental gross income declared by T A than its peers, and hence the higher the possible compliance risk T A possesses. On the contrary, the higher the value of z-score(expense), the more the rental gross expense claimed by T A than its peers, and the higher the possible compliance risk T A possesses. Accordingly, we can use Formula (2) to calculate a composite risk score that indicates T A's potential compliance risk, taking into consideration both expense and income. The higher the risk score, the higher the potential compliance risk T A incurs.

$$\text{Risk score} = \text{z-score(expense)} - \text{z-score(income)} \qquad (2)$$

4. CASE STUDY

This section applies our proposed stratified random sampling approach to over 15,000 tax agents that lodged altogether over 1.45 million residential rental properties in a tax return year. To protect privacy, we have left out the exact year information and have substituted dummy index numbers for real agent identities. If a property has multiple stakeholders associated with the same tax agent, its gross income is the sum value across all the stakeholders' shares. Likewise its gross expense. This property should be counted only once. Otherwise, if a property has multiple stakeholders associated with different tax agents, we advise to exclude this property from the input data.

4.1 Risk profiling for a single tax agent

For each actual tax agent, we demonstrate the risk identification results using one table and two figures. We take Tax Agent X as example to explain in detail what the table and figures tell us. Tax Agent X has 443 rental properties. Accordingly each of its peers has 443 rental properties as well. For the tax agent as well as every peer, we can calculate its mean value averaged across the 443 properties in terms of rental gross income and rental gross expense respectively. We report the resulting statistics in Table 2 when comparing Tax Agent X with its peers.

- No. of properties: number of rental properties owned by this tax agent's clients and lodged in individual income tax returns by this tax agent on their behalf.
- X's actual $ value per property: this tax agent's mean rental gross income or expense value.
- X's expected $ value per property: the average value across all the peers' mean rental gross income or expense values. It is the expectation for this tax agent drawn from the peers' behaviours.
- Peers' minimum $ value per property: the smallest mean rental gross income or expense value among all the peers.

- Peers' maximum $ value per property: the biggest mean rental gross income or expense value among all the peers.
- Peers' standard deviation: the standard deviation of the peers' mean rental gross income or expense values.
- z-score: the standardised difference between the tax agent's actual rental value and its expected value drawn from its peers.
- Risk score = z-score(gross expense) - z-score(gross income). It is used to rank actual tax agents in terms of compliance risk. The higher the risk score, the higher the potential compliance risk.
- Risk rank: this tax agent's rank among all actual tax agents in terms of compliance risk. The most risky tax agent is ranked as 1, the second most risky is ranked as 2, and so on and so forth.

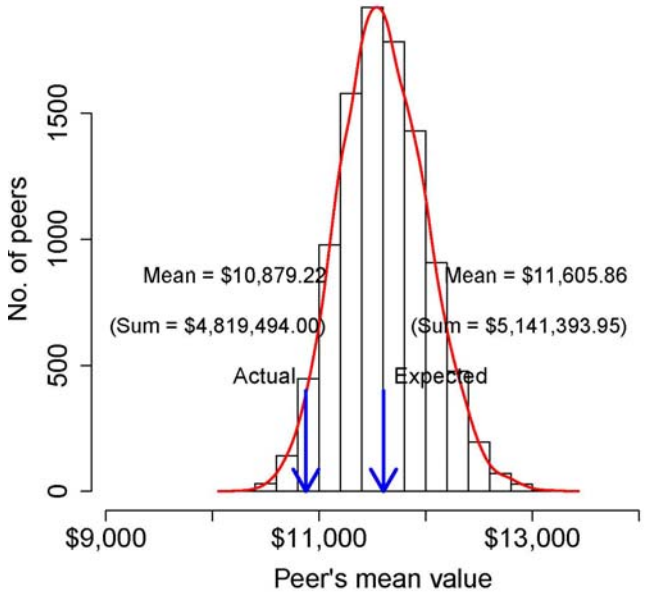
No. of properties = 443		
	Expense	**Income**
X's actual $ value per property	$28,153.76	$10,879.22
X's expected $ value per property	$15,606.33	$11,605.86
Peers' maximum $ value per property	$13,361.91	$10,232.96
Peers' maximum $ value per property	$18,465.19	$13,259.61
Peers' standard deviation	591.62	407.96
z-score	21.21	-1.78
Risk score = z-score(expense) – z-score(income) = 22.99		
Risk rank = 1		

TABLE 2: STATISTICS OF THE ACTUAL TAX AGENT AND ITS PEERS

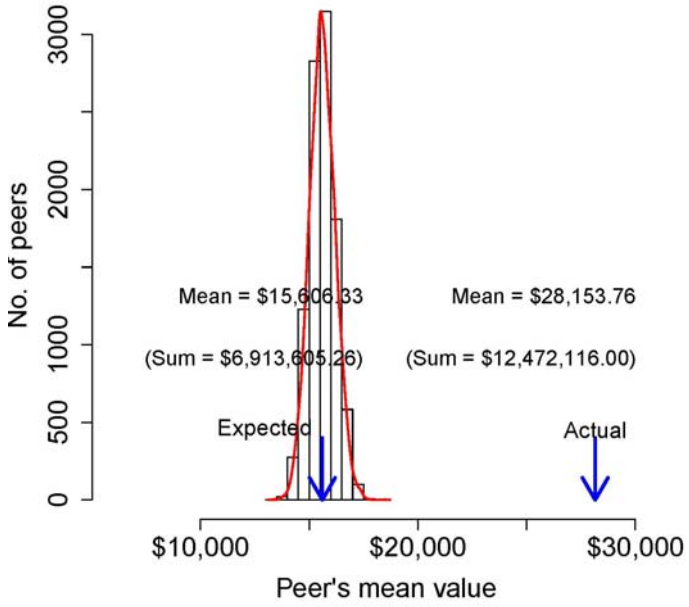
Furthermore, Figure 3 graphically portrays the information of Table 2. It shows where Tax Agent X sits in the context of its peers, when the return behaviour is the mean rental gross income (or expense) value.

Figure 3(a) shows the number of peers with specific mean rental gross income values. Most peers have their mean income values around $11,600. A few go up to $13,200, a few go down to $10,200, and the standard deviation is 407.96. The average value across all the peers' mean income values is $11,605.86, which is the expected mean value for Tax Agent X's rental gross income. In reality, Tax Agent X's mean income is $10,879.22, which is lower than the expectation and incurs a z-score of -1.78 ($= \frac{10879.22 - 11605.86}{407.96}$).

Figure 3(b) shows the number of peers with specific mean rental gross expense values. Most peers have their mean expense values around $15,600. A few go up to $18,400, a few go down to $13,300, and the standard deviation is 591.62. The average value across all the peers' mean expense values is $15,606.33, which is the expected expense value for Tax Agent X. In reality, Tax Agent X's mean expense value is $28,153.76, which is much higher than the expectation and incurs a z-score of 21.21 ($= \frac{28153.76 - 15606.33}{591.62}$).



(a) Rental gross income



(b) Rental gross income

FIGURE 3: Compare Tax Agent X's mean rental gross income and mean rental gross expense respectively against its peers'. X underreports its rental income but overclaims its rental expense.

Thus, Tax Agent X underreports its rental income but overclaims its rental expense. Overall it incurs a risk score of 22.99 (= 21.21 – (- 1.78)), which is the highest among all actual tax agents and hence is ranked number 1 in terms of potential risk.

In case that readers want to know sums in addition to mean values, we also show in Figure 3 that if all the 443 properties are considered, the total rental expense value claimed by Tax Agent X is $12,472,116.00 that is significantly more than the expectation drawn from the peers ($6,913,605.26); and the total rental income value declared by Tax Agent X is $4,819,494.00 that is less than the expectation drawn from the peers ($5,141,393.95).

4.2 Risk profiling for the tax agent industry

In addition to profile for each individual agent, our approach can also illustrate the global compliance picture of the tax agent industry. These collective results provide insight into the compliance level of the tax agent industry, and help the Australian Taxation Office promote and assist a capable and well-regulated tax and accounting profession.

Following the same line of reasoning as explained in Section 4.1, our approach can assign a risk score to each tax agent by comparing the tax agent against its own peers. Because the z-score and thus the risk score are standardised values, different tax agents' risk scores are commensurate.

Figure 4 illustrates the distribution of risk scores of over 15,000 actual tax agents operating in a tax return year. The higher the risk score a tax agent gets, the higher the compliance risk the tax agent potentially possesses. Most agents have risk scores close to 0, which indicates their behaviours are close to the expectation drawn from their peers. In contrast, a few tax agents (such as A, B and C at the right end of the spectrum) have abnormally high positive risk scores and are identified as potential high risk.

Figure 5 illustrates individual tax agents' risk scores for a tax return year, where on average one agent lodged 92 rental properties. For the sake of clarity, we only illustrate in Figure 5 those tax agents that have more rental properties than average, which results in 5075 tax agents. In this particular figure, we depict the top 5% tax agents as red triangles to represent risk agents. Blue circles represent non-risk tax agents. However, this cut-off percentage value can be tailored in practice to take into consideration available audit resources.

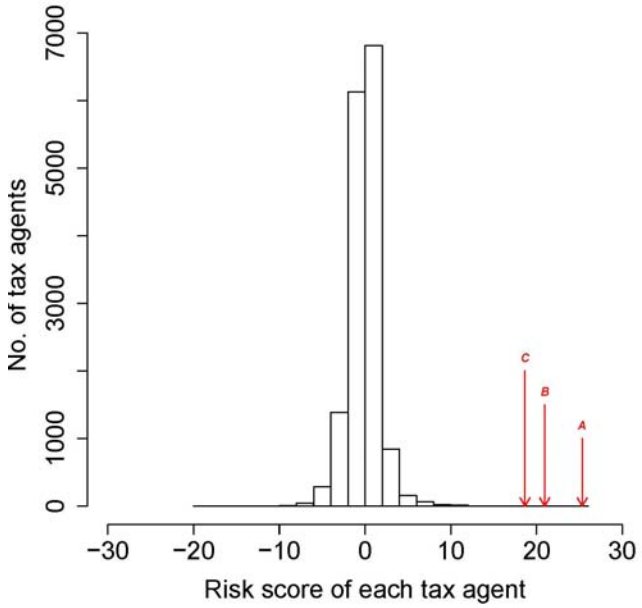


FIGURE 4: The risk score distribution of over 15,000 actual tax agents operating in a tax return year.

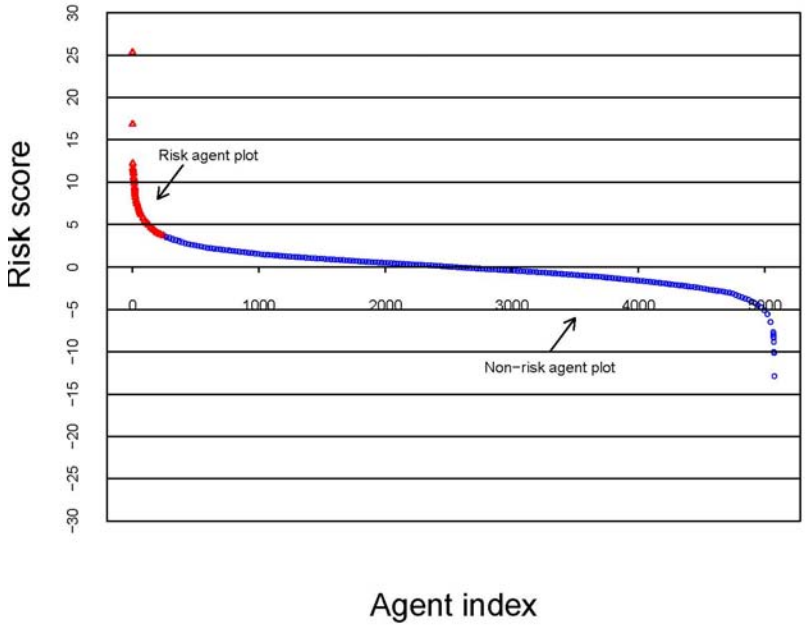


FIGURE 5: Individual tax agents' risk scores for a tax return year.

4.3 Effciency

Our proposed stratified random sampling algorithm is very efficient. Given the rental data of over 15,000 tax agents and over 1.45 million residential rental properties in a tax return year, our approach can accomplish calculating the z-score of any rental behaviour for each and every tax agent within two hours using a computer of the following configuration:

- cpu model name: Dual-Core AMD Opteron(tm) Processor 2220;
- cpu speed: 2800.469 MHz;
- cache size: 1024 KB;
- memory: 8179380 KB.

5. FURTHER DISCUSSION

We now highlight some technical issues to help peer researchers circumvent possible pitfalls when re-implementing our method in their fields.

5.1 What behaviour to evaluate

This particular paper examines a tax agent's compliance risk in terms of rental behaviours. As illustrated in Figure 6, there exist many rental behaviours that one can evaluate. We have proposed to choose rental gross income and gross expense respectively. That is not an arbitrary decision. Instead the choice is made in order to achieve an appropriate trade-off between providing enough details and providing the big picture of a tax agent.

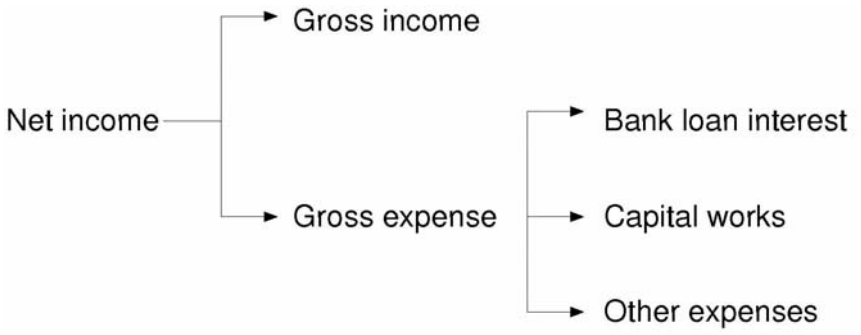


FIGURE 6: Alternative rental behaviours to evaluate

For example, rental net income is a composite quantity that reects the (possibly distinct) reporting behaviours of gross income and gross expense. It equals to gross income minus gross expense. Our stratified random sampling approach can compare Tax Agent X's mean net income value against its peers', and accordingly produce a risk profile like Figure 7. It is observed that the average rental net income reported by Tax Agent X is $-17,274.54, while the expectation1 drawn from its peers is $-3,996.04. Hence Tax Agent X declares much less net income than expected and

possesses potential compliance risk. But there can be many reasons behind such a symptom. Possibly Tax Agent X correctly reports gross income but significantly overclaims gross expense; or possibly it correctly claims gross expense but significantly underreports gross income; or possibly it both underreports gross income and overclaims gross expense. However, an analysis of net income alone would not reveal these useful details.

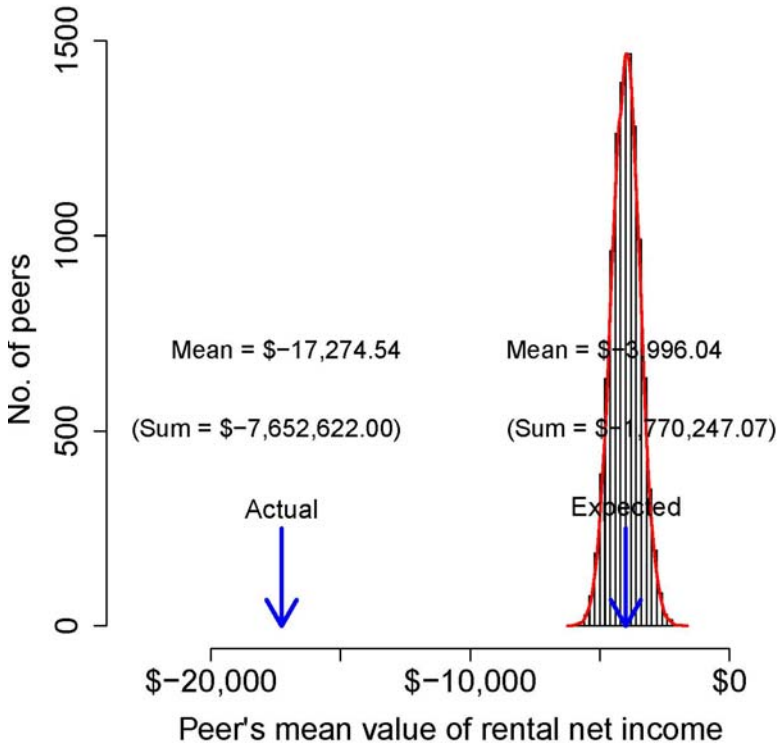


FIGURE 7: Compare Tax Agent X's mean rental net income against its peers'.

Alternatively one can use behaviours more detailed than gross income and gross expense. For instance, gross expense can be further divided into expenses of bank loan interest, capital works and other expenses. Our stratified random sampling approach can compare respectively a tax agent's mean of gross income, mean of bank loan interest, mean of capital works and mean of other expenses against its peers', and calculate the z-score for each behaviour. In such a case, the risk score should be calculated by Formula (3). However Formula (3) is sometimes over sensitive to small components and thus loses the big picture. For example, one agent does not overclaim expenses. But it puts all expense values into the item “other expenses”. Because few agents do such a thing, this tax agent incurs a significantly high z(other expenses), which dominates its z(bank loan interest), z(captial works) and z(gross income). Hence overall this tax agent incurs a high risk score. But this is an education issue rather than a compliance risk issue. Hence we suggest to avoid choosing over-detailed behaviours unless you are especially interested in some specific rental behaviour of a tax agent rather than its overall rental compliance level.

$$\text{Risk score} = \frac{z(\text{bank loan interest}) + z(\text{capital works}) + z(\text{other expenses})}{3} - z(\text{gross income}) \quad (3)$$

Note that $(gross expense) = $(bank loan interest) + $(capital works) + $(other expenses). However, z(gross expense) ≠ z(bank loan interest) + z(captial works) + z(other expenses) because a z-score is a standardised value. Instead 3xz(gross expense) ≈ z(rental interest) + z(captial works) + z(other expenses).

5.2 The central limit theorem

According to Moore [5], the central limit theorem says that the distribution of a sum or average of many small random quantities is close to normal. The theorem suggests why the normal distributions are common models for observed data. Any variable that is a sum of many small inuences will have approximately a normal distribution. How large a sample size n is needed for $\bar{x}$ to be close to normal depends on the population distribution. More observations are required if the shape of the population distribution is far from normal.

> CENTRAL LIMIT THEOREM
>
> Draw a simple random sample of size n from any population with mean μ and finite standard deviation σ. When n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately normal:
>
> $\bar{x}$ is approximately $N(\mu, \frac{\sigma}{\sqrt{n}})$.

Translated into our context, the theorem indicates that if a tax agent has too few rental properties, its peers' mean values of gross income (or gross expense) might not follow a normal distribution. An example is illustrated in Figure 8. In such a case, the z-score is no longer applicable. Hence it is compulsory to confirm that the peer values follow a normal distribution before using the z-score statistic.[5]

[5] In statistics, there exist a few methods to perform a normality test, such as the Kolmogorov-Smirnov test, the Anderson-Darling test, the Shapiro-Wilk Test and the Skewness-Kurtosis All test. There also exist graphical techniques such as the Q-Q plot to compare two probability distributions by plotting their quantiles against each other.

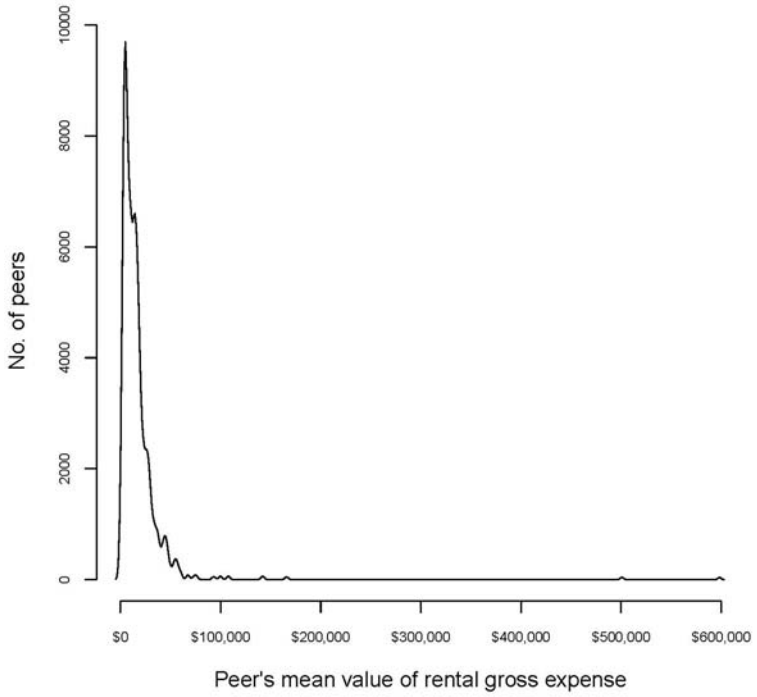
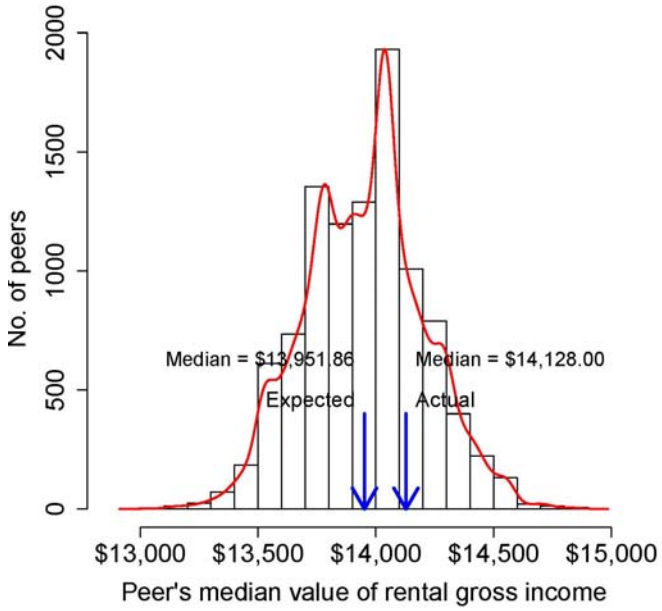


FIGURE 8: A small tax agent has only one rental property. Its peer means does not follow a normal distribution.

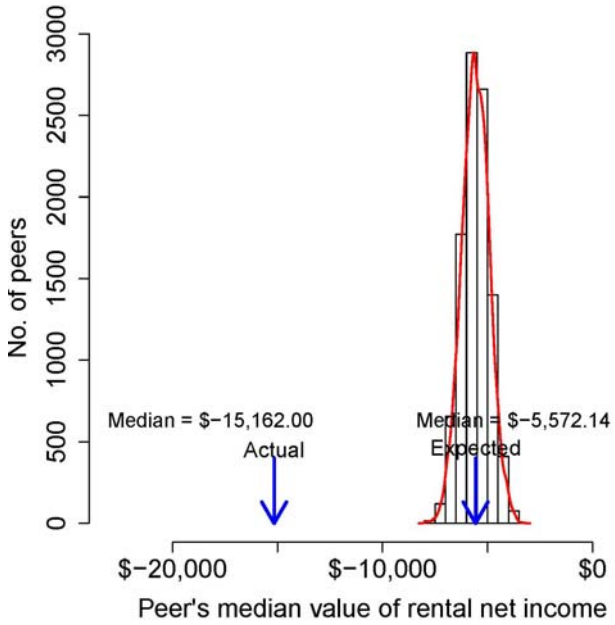
5.3 Median vs. mean

Sometimes people are interested in a tax agent's median rental value instead of its mean rental value.[6] Extra cautions are required when applying our stratified random sampling approach to compare a tax agent's median value against its peers'. Although it applies to the mean statistic, the central limit theorem does not necessarily apply to the median statistic. That is, the peers' median rental values do not necessarily follow a normal distribution. For instance, as illustrated in Figure 9(a) the median rental gross income values of Tax Agent Y's peers assume a bimodal distribution instead. As a result, a z-score is not always applicable and we cannot use Formula (2) to calculate the risk score. Nonetheless, it happens in this particular case that the median rental net income values of Tax Agent Y's peers still follow a normal distribution as depicted in Figure 9(b). Thus it is acceptable for one to calculate the z-score of Tax Agent Y's net income and evaluate its potential compliance risk therefrom. Hence same as concluded in Section 5.2, one must check peer values' distribution before using the z-score statistic.

[6] Median is often a method of choice to avoid skewness introduced by a few extremely large or small values in the population. By random sampling, our approach can successfully avoid the skewness problem and hence does not need to use the median statistic.



(a) For Tax Agent Y, the peers' median values of rental gross income follow a bimodal distribution instead of a normal distribution. Hence a z-score is not applicable.



(b) For Tax Agent Y, the peers' median values of rental net income do follow a normal distribution. Hence a z-score is applicable.

FIGURE 9: The central limit theorem does not cover the median statistic. If using median instead of mean to measure tax agent behaviour, one should always check whether peer median values follows a normal distribution before adopting the z-score to quantify a tax agent's potential compliance risk.

5.4 Ratio

In general, we discourage using ratio values as behaviour, such as $behaviour = \frac{rental\,gross\,expense}{rental\,gross\,income}$. It is because a small denominator value will blow up the ratio and distort the behaviour. The extreme is when denominator is 0 and the ratio becomes infinitely big. Even if we replace 0 with some positive value to solve the infinity problem, the distortion problem still exists. Table 3 shows a true story. Tax Agent Z has 18 rental properties, whose rental gross income and gross expense are listed in Table 3. 10 out of the 18 properties have $0 gross income. In order to calculate the ratio of gross expense divided by gross income, we replace $0 rental gross income with $1. The ratios are then calculated accordingly for each property.

	Property	Gross Income	Gross Expense	Ratio ($\frac{Gross\,Expense}{Gross\,Income}$)
Tax Agent Z	a	$14,190	$26,136	1.84186046511628
	b	$10,469	$25,867	2.47081860731684
	c	$13,543	$39,424	2.91102414531492
	d	$12,960	$39,508	3.04845679012346
	e	$7,359	$37,899	5.15002038320424
	f	$2,700	$24,492	9.07111111111111
	g	$1,603	$26,880	16.7685589519651
	h	$1,587	$28,954	18.244486452426
	i	$1	$0	0
	j	$1	$492	492
	k	$1	$2,410	2410
	l	$1	$4,852	4852
	m	$1	$6,652	6652
	n	$1	$11,278	11278
	o	$1	$14,191	14191
	p	$1	$15,927	15927
	q	$1	$16,299	16299
	r	$1	$19,885	19885

TABLE 3: RATIO IS OFTEN A DISTORTED BEHAVIOUR.

Next we use our stratified random sampling method to create all the peers. Since the mean statistic for ratios does not have a proper meaning, we calculate the median ratio value for the tax agent as well as for every peer. Assume the peers' median ratio values follow a normal distribution.

- The risk score is calculated as follows:
- Tax Agent Z's actual median ratio = 255.12
- Tax Agent Z's expected median ratio = 1.94
- Peers' minimum median ratio = 0.99
- Peers' maximum median ratio = 3.21
- Standard deviation of peers' median ratios = 0.26
- Risk score = z-score = 979.81

- Risk rank = 1.

Thus Tax Agent Z incurs a very high risk score of 979.81 and is ranked as top risk, whereas the second highest risk score among all tax agents is only 33.33. We suggest that Tax Agent Z's risk is largely exaggerated and ratio is the reason to the distortion. Hence one needs to be very cautious when using ratio.

6. RELATED WORK

Our concept of "notional peers" is inspired by Bloomquist, Albert and Edgerton's bootstrap approach to evaluating preparation accuracy of tax agents [1]. In Bloomquist etc.'s study the tax agent behaviour is the AUR discrepancy rate, which equals to the number of tax returns lodged by a tax agent with potential misreported values divided by the total number of tax returns lodged by that tax agent. The misreported errors of tax returns are identified by the Automated Underreporter (AUR) program of the US Internal Revenue Service. Assume a tax agent T A lodges 12 tax returns of Postcode 20134 and 45 tax returns of Postcode 20143. The bootstrap approach creates T A's notional peers and evaluate T A's compliance risk by the following steps.

Step 1: Randomly pick 12 and 45 tax returns from all the tax returns of Postcode 20134 and Postcode 20143 respectively. The resulting 57 (= 12 + 45) picked tax returns will contribute to create a notional peer Peer1 for T A as in Step 2.

Step 2: For each of the above 57 tax returns, a uniform random number ($0 \leq u < 1$) is generated. If the value of u is less than or equal to the AUR discrepancy rate of the tax return's corresponding Postcode, a value 1 is added into Peer1's base; otherwise, a value 0 is added into Peer1's base.

Step 3: Compute Peer1's AUR discrepancy rate as $\overline{D_1} = \frac{1}{57}\sum_{i=57}^{m} x_i$ where $x_i \subset \{0, 1\}$.

Step 4: Repeat Steps 1-3 for 1000 times, creating 1000 notional peers for T A. The expected AUR discrepancy rate for T A equals to the average value of the 1000 notional peers' AUR discrepancy rates: $\hat{\theta} = \frac{1}{1000}\sum_{j=1}^{1000}\overline{D_j}$.

Step 5: Obtain the one-tailed 95% confidence interval by sorting the 1000 peer AUR discrepancy rates in ascending order and selecting the cutoff as the 950th value.

Step 6: If T A's AUR discrepancy rate exceeds the 95% confidence interval (the 950th value), it is identified as being a potential risk.

We respectfully suggest that the bootstrap approach does not quantify tax agent compliance risk. Consequently, it does not compare risk degrees across different tax agents to offer a risk ranking among multiple tax agents. However a proper risk ranking is highly desired in tax administration organisations such as the Australian Taxation Office because it enhances the effectiveness and efficiency of tax audit under resource constraints. Hence we have instead proposed a stratified random sampling approach where we have proved via the central limit theorem that one can use the z-score to quantify potential tax agent risk regarding a behaviour. Meanwhile, since z-

scores are commensurate across different behaviours, we can apply mathematical operations on them to calculate a collective risk score for each tax agent. Multiple agents can be ranked according to their risk scores. These scores together with our proposed descriptive illustrations can provide important insight into the integrity an compliance level of a single tax agent as well as of the whole tax agent industry. Hsu etc. reported to use supervised learning to improve the audit selection procedure at the Minnesota Department of Revenue [3]. In the machine learning and data mining fields of computer science, there exist supervised learning versus unsupervised learning approaches [4, 6]. Supervised learning needs training data, that is, an unbiased and representative sample of the whole population where each of the sample returns has a known outcome (compliance or noncompliance). From the training data supervised learning infers a classifier to differentiate between compliance and non-compliance tax returns. This classifier is then used to classify other unlabelled tax returns. In their particular work, Hsu etc. had access to tax returns with auditing results and trained a naive Bayes classifier therefrom. In contrast, we lack the luxury of having good training data of agent compliance risk due to the fact that tax agent client bases are immensely diversified. Thus our proposed approach is unsupervised learning that does not demand a supply of labelled agents. As a result, our approach is of very low cost and can be easily made operational. A traditional risk identification approach in the Australian Taxation Offce is to use business expert rules. A rule system often first specifies non-compliance patterns according to domain experts' previous experience, and then sifts current data through those patterns. Tax agents that match any pre-specified behaviours will be deemed as suspicious. Thus such an approach heavily relies on historical data and previous experience, which are very valuable but will always be one step behind the current data and the newly emerging information presented by the current data. In contrast, our proposed approach is purely data driven. It explores typically large amount of data and discovers knowledge presented by the data. As a result, it can be perfectly synchronised with the current data and is very good at discovering new information that often goes beyond our existing knowledge base. Another advantage of our approach over a rule system is that our approach is robust to infiltration. A rule system often holds a few critical man-made threshold values such as w in the rule “if (rental gross expense) > w x (rental gross income) then ‘risky’”. If fraudsters find out these thresholds, it is relatively easy for them to manipulate their return values so as to make claims just under the threshold and thus to avoid being identified. On the contrary, if fraudsters intend to deceive our proposed system, they have to know the behaviours of their peers. Since the peers are randomly sampled from the whole population, the fraudsters have to know the behaviours of the whole population, which information is very difficult to obtain. Hence our approach is much more robust to malicious intrusions than a rule system.

7. CONCLUSION

In this paper we have shared our positive experience of delivering effective and efficient identification of potential tax agent compliance risk, which is traditionally a very demanding and expensive task that consumes substantial amount of auditing resource and time. Meanwhile we have shared the lessons we have learnt throughout the process. In particular, we have proposed to compare an actual tax agent T A with its notional peers in order to measure the potential risk of T A's return preparation behaviours. The notional peers are created via stratified random sampling such that they are commensurable with T A and that they offer a proper industry norm for T A. According to the central limit theorem, the peers' preparation behaviours will follow a

normal distribution. Therefore one can use the z-score to quantify the degree of T A's compliance risk potential.

We have also proposed to profile agent compliance risk through well designed illustrations. Such illustrations are easy to understand, and at the same time provide important insight into the integrity and compliance level of a single tax agent as well as of the whole tax agent industry.

We have applied our proposed method to the Australian tax agent rental data. Our preliminary results are well received and welcomed by executives and auditors in the Australian Taxation Office. Further field assessments are being undertaken on the method outcomes, which we expect to be able to help the Australian Taxation Office promote and assist a capable and well-regulated tax and accounting profession.

8. ACKNOWLEDGMENTS

We gratefully acknowledge Mr. Kim M. Bloomquist (Internal Revenue Service, US) and Mr. Graham Whyte (Australian Taxation Office) for their thoughtful and constructive comments on this paper.

REFERENCES

[1] K. M. Bloomquist, M. F. Albert, and R. L. Edgerton. Evaluating preparation accuracy of tax practitioners: A bootstrap approach. In Proceedings of the 2007 IRS Research Conference, pages 77-90, 2007.

[2] B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence errors, and other measures of statistical accuracy. Statistical Science, 1(1):54-75, 1986.

[3] K.-W. Hsu, N. Pathak, J. Srivastava, G. Tschida, and E. Bjorklund. Data mining based tax audit selection: A case study from Minnesota Department of Revenue. In Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2009.

[4] T. Mitchell. Machine Learning. McGraw Hill, 1997.

[5] D. S. Moore. The Basic Practice of Statistics, 2nd Edition. W. H. Freeman and Company, Third printing 2000.

[6] I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. Morgan Kaufmann, 2005.