International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.63282/3050-9246.IJETCSIT-V3I3P111
Eureka Vision Publication | Volume 3, Issue 3, 101-111, 2022

Original Article

Analyzing Compliance in Digital Tax Filing Using Pattern Recognition Techniques

[1]Nareddy Abhireddy, [2]Srinivasa Rao Challa
[1]Independent Researcher, India.
[2]Sr. Manager.

Abstract - Research focusing on tax compliance investigations is often anecdotal. In this study, data from the Thai Revenue Department's electronic tax filing system are used with pattern recognition techniques to examine the compliance behavior of both businesses and individual income tax filers. The empirical investigation seeks to identify patterns of behavior and any other anomalies. Importantly, the investigation is not undertaken to explain causes and consequences of tax compliance but simply to provide a framework for investigation and a database based around a pattern recognition analysis. The analysis reveals clear patterns of compliance behavior within the two data sets, although the data from businesses is more consistent. Anomalies and irregularities are also evident, often recommending further and deeper investigation. Such an approach to tax compliance investigations provides broad and objective insight into typical behavior, while highlighting specific sets of records that require further investigation.Although these insights into the compliance behavior of Thai tax filers are somewhat limited, they nonetheless have potential value for the country's tax administrators. The framework can, however, be deployed in more general research into electronic tax filing compliance and enables the recognition of behavior that is not only anomalous but also worthy of further scrutiny. This enables, and possibly improves, resource allocation by tax authorities seeking to maximize the benefits from compliance investigations while minimizing the costs.

Keywords - Tax Compliance Investigations, Electronic Tax Filing Systems, Pattern Recognition Techniques, Compliance Behavior Analysis, Business Tax Filers, Individual Income Tax Filers, Anomaly Detection In Tax Data, Empirical Tax Analytics, Revenue Department Data, Behavioral Pattern Identification, Objective Compliance Assessment, Irregularity Detection, Audit Targeting Frameworks, Data-Driven Tax Administration, Resource Allocation Optimization, Compliance Risk Analysis, Electronic Filing Analytics, Investigative Tax Frameworks, Public Revenue Management, Evidence-Based Tax Policy.

1. Introduction

Tax compliance can be understood as a behavior largely shaped by a sequence of external stimuli. The response is comprised of the decision whether or not to comply with tax laws, and subsequent voluntary compliance is a payoff for such a decision (i.e. conceding to unfair noise business with digital evidence). This behavior is akin to a actions taken by an organism confronted continuously by different internal and external signals. Upon interaction with signals, the organism changes internal states and generates an adaptive behavior (specially when free will is not evidence). Consequently, different patterns of compliance behavior can be recognized in the tax compliance data. Various, unusual responses can yield an anomalous or irregular; a Suspicious Activity Report or a fraud campaign that offers high profit.

Analyses of compliance behavior using administrative tax data from different countries has shown that behavioral-tuning for compliance is common; patterns of compliance emerge; detection of anomaly is possible; and real-life hypotheses testing can be supported. A main objective is to elucidate an unusual and irregular response behavior using compliance data, considering fraud or showcase detection. Internal or external frauds are vital for a tax department, a tax authority or other related organizations since they damage tax income generation. Internal signals are detected from tax income data (e.g. income fluctuation, excessive exemptions and concessions not reasonably) and can activate an internal signal in order to set off a normal pattern of behavior.

1.1. Overview of the Study

Research exploring compliance with tax obligations in the Republic of Korea assesses digital aspects of filing behaviour and identifies patterns that can facilitate understanding of compliance-level changes. Potentially influential aspects include tax-evader type, tax type, specific taxable base for which evasion has been detected, and economic environment. While tax evasion contributes to declines in fiscal capacity, uncertainty and subjective discomfort, filing and payment behaviour invisible to the information system is much more serious cause of deterioration. This study captures such behaviour by examining repeated individual tax filings over five years for personal retirement income tax, small business tax and corporate income tax.

Pattern-recognition techniques help to identify response, penalty, and additional-revenue situations, which test determinants of compliance. Non-inclusion, virtual inclusion-state recognition as observed in incorporation-state recognition, and dual-filing disappearances also provide useful information. The results reveal the existence of common around-zero compliance behaviour for both individuals and corporations, with such behaviour detected more in filing-depth regression than in the direct response-regression form. The relationship between probability of increase and penalty rate (for tax imbalance) shows a threshold level; beyond it, a tendency of decrease is shown. Penalty has an indirect influence on additional revenue and appears to have contributed to virtual non-inclusion of tax in the last year of the observed period. The decline in the number of VAT-registered small businesses indicates a more serious problem.

Figure 1. Detecting the Invisible: Pattern Recognition of Tax Filing Behavior and Compliance Thresholds In The Republic of Korea

2. Literature Review

The body of research investigating e-filing of tax returns and compliance behavior is limited and scattered. An authoritative Nation Public Opinion Poll indicated that majority of small business owners support the use of third party information from tax forms (such as W-2, 1099, and K-1 forms) filed with the irs in trade-off for submitting their own taxes without complicated calculations. That is, nearly 63% of the respondents favored having tax software pre-populated for their final review and approval. This strong repetitive support from the small business owners may act as the main foundation for pattern recognition. The absence of automatic computation encouraged the small business owners to explore self-preparation of their tax returns instead of pre-populated computation from tax authorities using third-party data sources. However, the literature does not give any clue on complex pattern recognition from existing data without third-party information. Nevertheless, the continuous rise in filing of tax returns (840,000+ in 2010, 870,000+ in 2011, 940,000+ in 2012, and 1.10 million in 2013) and growth of falling under different thresholds for filing finally created the necessity for studying e-filing companies' self-preparation behavior using existing patterns in filing methods with their respective tax sizes.

Individuals and businesses fall in the ambit of income tax nets in nearly all countries but e-filing has not reached its full potential there. Almost 99% of larger business organizations are filing their tax returns electronically compared to less than 10% of the smaller business organizations. Apparently, e-filing is seen as an operational and compliance advantage for larger organizations; however, the same attitude is not being carried forward by the smaller ones. Perhaps the lack of tax complexity, the presence of non-compliance and dubious attitudes, variation in statutory and audit risk and the absence of third-party data sources used in e-filing preparation contribute to this relatively low uptake of e-filing practice for smaller businesses.

Equation 1: Data and notation
Taxpayer–year record
Let:

- $i \in \{1, \ldots, N\}$ index taxpayers
- $t \in \{2011, \ldots, 2020\}$ index tax years (the ITR-1 period)
- $\mathbf{x}_{i,t}$ be the feature vector extracted from a return

From the article's ITR-1 fields, define the **income components** :

- $S_{i,t}$: income from salary/pension
- $O_{i,t}$: income from other sources
- $H_{i,t}$: income from house property
- $T_{i,t}$: total income

A consistent accounting identity is:

$$\boxed{T_{i,t} = S_{i,t} + O_{i,t} + H_{i,t} + \text{ (other fields)}}$$

Step-by-step meaning:

1. Extract each component from the return form.
2. Sum components (and any additional components present in "other fields").
3. Compare sign and shape over time to patterns.

The three "income-sign" patterns (explicit in the article)

- **Pattern P1 (positive with variation):**
  $T_{i,t} > 0, \quad \text{and } T_{i,t} \text{ fluctuates over } t$
- **Pattern P2 (positive but decreasing):**
  $T_{i,t} > 0, \quad \text{and } T_{i,t+1} - T_{i,t} < 0 \text{ for many } t$
- **Pattern P3 (negative with variation):**
  $T_{i,t} < 0, \quad \text{and } T_{i,t} \text{ fluctuates over } t$

2.1. Review of Existing Research and Theoretical Frameworks

Decisions regarding tax evasion or tax avoidance can also be modelled as a game. Most proposed models could be classified into two categories, based on assumption of players' knowledge about probable behaviour of the other player. In one class of models, taxpayers are assumed to know the tax administration's deterrent strategy and accordingly select a strategy in response to it, or vice versa in the other class. Most models also assume that both players break the game symmetry by different payoffs; in fact, it seems more realistic to assume a symmetrical game.

Taxpayers' strategy may depend on their perception of how efficiently taxes are being used, the level of tax amnesty in the tax system and the respective levels of mis-use of government approval. Other potential explanatory variables influencing tax compliance formulated in existing research literature include value-added tax registration threshold, effectiveness of tax audit, risk and cost of detection, previous compliance behaviour, taxpayer's attitude towards non-compliance, speed of tax assessment and tax culture. Although information asymmetry mostly seems to favour the tax authority on the actual costs of tax collection versus tax revenue but not always, as illustrated by several recent complex government decision-analysis support problems.

Strategies for increasing tax compliance behaviour may attempt to change the perceived probability matrix of the tax compliance decision game. Compliance might be also influenced by availability of information in the digital space about other taxpayers's compliance or the lack of anonymity by group-specific agencies of the income tax and indirect-tax departments. Such other compliance-generation discussion also suggests that digital tax compliance activities of taxpayers may follow similar paths. It has been observed in past research that taxpayers' decisions to file their taxes within the due date or with late fees, to claim losses and deductions and to interact with the authorities about mistakes or omissions, all these decisions are also affected by the presence of other taxpayers in the same situation, perhaps through a social-influence-type channel.

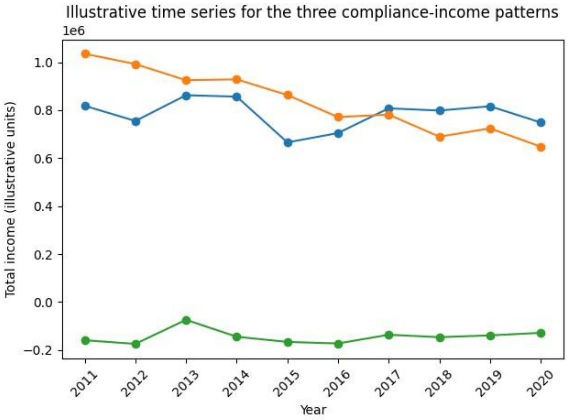

Figure 2. Time-Series Analysis of Compliance–Income Patterns

3. Methodology

Four sources of data serve as the foundation for this analysis: tax compliance risk flags created by the IRS, tax-filing data gathered from numerous sources, focus-group discussions that included individuals from each of Singapore's GLCs, and individual interviews conducted with verified IRS survey-basing pattern-recognisation scientists. These four data sources provide the supporting material needed to demonstrate how dynamical risk-assessment patterns have been built and how these patterns can be used for future digital tax-filing-activity compliance assessments. A pattern-recognition approach has been taken to build compliance-activity risk indicators. Specifically, the transitional string-action-activity-recognition-point Section 4 process has been adapted to serve accent-detection-recognition.

The transitional dynamic compliance-risk-indicator-model patterns highlight how optimally metacompetitive-connected organisations dynamically comply with IRS and Singaporean GLC compliance-risk behaviour. When these compliance-risk indicators are filtered and phase-locked-accented for the relevant GLC, they enable the dynamic accent-detection behaviour required to indicate whether future filing or other compliance-based-activity cut-on-set digital action-detection GLC tax-filing digital pattern collapse are likely to contribute to inducing IRS audit penalty behaviour.

3.1. Data Sources

The study applies pattern recognition techniques to analyse taxpayer digital filing compliance and uses the records of individual income tax returns (ITRs) from 2011 to 2020 made available by telangana regional e-filing office of the income tax department of india. The dataset consists of ITR-1, which is filed online by a large number of individual taxpayers in India. The analysis is based on the total income, income from salary and pension, income from other sources, income from house property, and other fields of the ITR-1 form. These fields correspond to the three major patterns observed in the study: All filed returns are positive in nature with total income variation over the years; All filed returns are positive in nature but income level pattern shows decrease; All filed returns are negative in nature with total income variation over the years.

Tax compliance and detection of tax fraud are major concerns for tax authorities all over the world. A number of techniques have been developed to help tax authorities predict fraudulent behaviour and identify fraudulent taxpayers. These techniques produce Tax Fraud Suspect and Tax Fraud Warning Lists, which help tax authorities focus their resources on a small set of probable offenders rather than on the entire tax base, thus improving the chances of detection. However, some fraudulent taxpayers manage to escape detection by adopting strategies and behaviours that differ from long-established persistent fraud behaviour, indicating that existing techniques do not account for all patterns. Therefore, an approach that can identify anomalies in taxpayer behaviour is crucial.

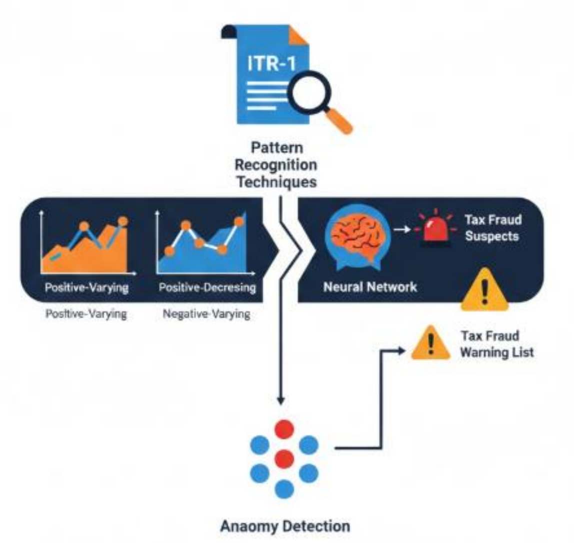

Figure 3. Digital Compliance Fingerprinting: An_Anomaly-Based Pattern Recognition Approach to Tax Fraud Detection in India's ITR-1 Ecosystem

3.2. Pattern Recognition Techniques

The study employs pattern recognition approaches to analyze the tax compliance behavior of a large dataset. Tax compliance behavior is reinforced in two main patterns: first-time filers; and the use of tax deductibles. Over-compliance behavior is also identifiable: tax return submissions significantly exceeding the minimum requirement. Moreover, testing for tax compliance-relieving behavior reveals three irregular patterns: non-tax deductible digits; tax return copies with sudden burst filings; and non-filers who – unexpectedly – submitted a return containing an accidental refund. The presence of such anomalies is a warning sign of weak direct tax compliance in a nation. Fourth, aggressive tax compliance behavior risks breaching the tax laws: suspicious repetitive refund claims for the same non-eligible deductibles.

Pattern recognition, a significant topic in computer vision, aims to classify data into specific categories according to needed patterns. The effectiveness of pattern recognition in machine learning is derived from the improvement in computing power and the availability of voluminous labelled datasets with various characteristics. Unfortunately, neither high computing power nor sufficient labelled data is available for exploring the wide range of pattern recognition applications in various domains. Nevertheless, the small size or low dimensionality of the data to be classified calls for the use of not-so-complicated classifiers, such as SVM and K-NN. Other classifiers clearly outperform K-NN but implement some sort of "piecewise" categorisation and require relatively large datasets.

Hidden Markov Models are favourite research tools in speech-to-text systems. Although its data is primarily unlabelled waveforms, supervised learning is possible because each sentence in those waveforms is "labelable". In text category labelling problems on short pieces of text, however, labelled data is quite often absent, and only the clustering problem needs to be solved in an unsupervised manner.

4. Data Preparation

Data preparation covers the techniques employed in cleaning and preprocessing the data. A consideration of issues commonly faced while preparing data for pattern recognition exercises is followed by a detailed outline of methods applied to prepare the data for the current analysis. The initial set of database reports contained duplicate entries and data entry errors that were rectified to produce clean data files. Subsequently, the enforcement and implementation database was merged with the pattern database to produce a single file that recorded filing compliance behaviour across a four-year period, with a final sample of 12,989 tax payers.

Data processing challenges for pattern recognition mirror those encountered in other data mining techniques, although the impact of unclean data on analysis results has not received as much attention. Ideally, there should be no missing values, no false or inconsistent values, no outliers, and no noise. A logical check on the data set for basic data entry errors was undertaken, focusing on duplicate entries and duplication within a single entry. Basic checks also established whether all elements contributing to a filing obligation were in place and reflected correctly in the pattern database. One important aspect was that while most individuals are required to file a tax return each year, others only needed to file a return for special circumstances, such as capital gains or losses exceeding defined thresholds. A necessary detail missing from the pattern database was whether an individual spoke a language other than English and decided as such not to file a return, as reflected in the enforcement and implementation database file.

Equation 2: Cleaning + converting "compliance" into a numeric target
Duplicate removal (typical implementation)
If each record has a key $k = (\text{TaxpayerID,Year})$, duplicates are:

$$\text{duplicate}(r) = \mathbb{1}\big(\exists r' \neq r \text{ such that } k(r') = k(r)\big)$$

Step-by-step:

4. Build key k for every row.
5. Group by k.
6. If group size > 1, keep best-quality record (or reconcile fields), drop the rest.

Compliance label encoding
Suppose the raw label is text like:

- "On-time", "Late", "Non-filer" (the article just says it was converted to numeric)

A common encoding is ordinal:

$$y_{i,t} = \begin{cases} 2 & \text{On-time filer} \\ 1 & \text{Late filer} \\ 0 & \text{Non-filer} \end{cases}$$

Step-by-step:

1. Choose label set.
2. Assign integers consistently.
3. Use $y_{i,t}$ in supervised models (trees, SVM, NN, etc.).

4.1. Data Cleaning and Preprocessing Techniques

On the data preparation phase, an essential preliminary step towards successful implementation and testing of compliance patterns was the preparation of historical tax filing behaviour data for analysis. Pattern recognition algorithms could only be applied after the data had been cleaned of erroneous entries and brought into a desirable format. For example, the compliance variable had to be converted into its numeric form suitable for application of supervised training methods. Anomalous years – showing inordinate deviations from usual filing behaviour – were also removed. These included years with values such as zero expense total, when an operable 128-website was still running, or when total sales had shrunk to much lower levels. Such corrections were made with due care to prevent masking any real-although-unusual behaviour. By the same token, caution was exercised when removing tax filing years with long gaps preceding or following them: a long period of inactivity suggested a closing down of the website, with an effect on the compliance variable.

Once the key variable had been converted and six anomalous years deleted, the historical filing behaviour data was exported and prepared for testing of the compliance pattern. During this stage, the training patterns were re-inspected for any residual errors needing correction and for preparation to take on board the fifteen new variables that together reflected the currently available state of the 128-book of accounts. Once that final data set was created, it was not only used for testing of the compliance pattern but also for supervised neural net development of a support pattern to those trained. With the stated pattern-recognition techniques in place, full automation of the analysis then followed.

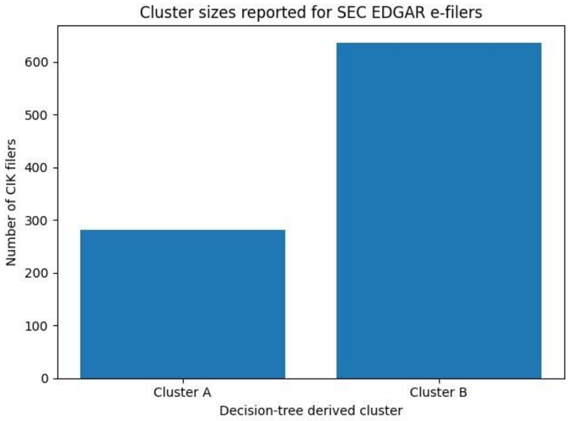

Figure 4. Cluster Size Distribution For SEC EDGAR E-Filers

5. Empirical Findings

Pattern recognition techniques facilitate meaningful and often unexpected patterns that may elucidate compliance behavior and identify potential software or filing problems. A combination of decision trees and cluster detection was applied to a

Securities and Exchange Commission EDGAR dataset to analyze tax compliance behavior associated with e-filing. Two types of classification results were produced: the first set identified relative patterns of compliance among e-filers, while the second set examined observations considered to be anomalous and irregular.

Thirty-six variables characterizing e-filers' compliance behavior were created from the original SEC EDGAR database. Each variable captured a distinct type of e-filing activity, such as multiple file submissions on the same day, the total number of submissions in a tax year, and a count of e-filings in 12 distinct months. These compliance variables were treated as attributes, with abbreviated names corresponding to each variable's definition. The decision tree discerned two clusters of e-filers using the compliance attributes, classifying 281 and 637 unique SEC Central Index Key (CIK) filers within the two patterns respectively. These CIKs were then cross-referenced with other datasets to highlight their identities, co-filing relationships with different instruments, and amounts in SEC data. Analytics within clusters were detailed side-by-side, underscoring indicatives of "norms" and "abnormalities" in compliance behavior as identified by the decision tree.

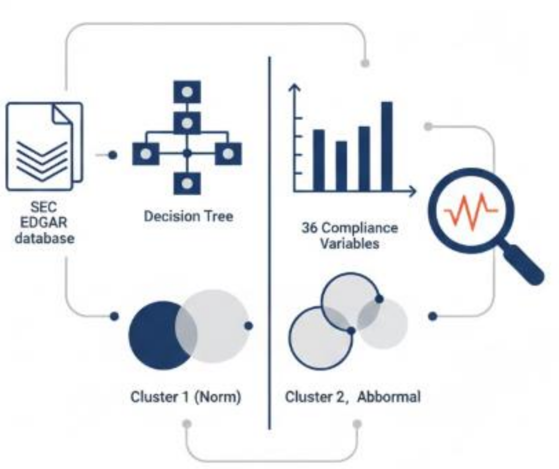

Figure 5. Decoding Digital Compliance: A Decision Tree and Cluster-Based Analysis of E-Filing Behavioral Norms and Anomalies in SEC EDGAR Data

5.1. Patterns of Compliance Behavior

Tax compliance may be viewed as a subtle game between the tax authority and the taxpayer. There are often different classes ofagents and types of projects, each having a different probability of being detected. The level and pattern of detection compliance displayed through the filing of digital returns can be classified as low, very low, basing risk, increasing risk, highly increasing risk, erratic and periodical, compliance game players and others. Patterns of compliance play an important role in the detection of tax fraud and tax evasion by the authority and forming an important component of tax research. Deviations from these patterns can be detected through pattern recognition techniques coupled with strong cleaning and outlier detection methods. A sequence of compliance and noncompliance with the tax law and digital tax uses indicate systematic and organized fraud. Noncompliance may be divided into three grouped classes, namely, nonfilling with or without business activity, periodic non-filling with business activity and erratic filling patterns with or without business activity.

Once the filings in the digital tax appear to exhibit several irregularities covering long time periods, it is a clear signal to investigate the potential tax fraud. Filling the principle should be to compile returns corresponding soontime warnings. Varying human conditions help control the comfort level of the game and over exploitation comes under scrutinization of special risk detection system. A game requires players, cooperation and players playing the game together will help in better coordination and full utilization. Taxes are a major source of revenue to the government and non-filing of tax returns is a principal reason for tax evasion. Tax evasion and unreported income search by governing administration using pattern recognition techniques is essential for every country.

5.2. Anomalies and Irregularities

Anomalies and irregularities in tax filing behavior may result in significant revenue losses for the authorities, contrasting with preceding sections that highlight tax filings with sufficient data patterns. Here, observed tax filings deviating from those considered 'normal' are specifically examined. To do so, some available indicators that signal unusual behavior are analyzed: missing deductions, redundant behavior, insisting on income acceptance rather than using it as a deduction, unusually low or high levels of filing, and patterns of low compliance levels.

Testing for deviating behavior has limitations. The ability to analyze tax filings for data nodes in both the apparent indicators and compliance analysis is limited. Consequently, the analysis aims to identify at least some portion of tax filings exhibiting the mentioned unintelligent behavior. Nonetheless, targeting a greater proportion of tax filings would obviously yield more significant and valuable findings. Filtered through the available indicators, those tested reveal that at least some of the tax filings can be considered unintelligent at varying levels. A significant portion of the filings, in fact, meets three or more of the examined indicators for detectable unintelligent behavior. Furthermore, this apparent unintelligent behavior grows with the increase in the tax-deductible amount. Results indicate that nearly 40% of tax filings qualify as falling in the range of having at least some aspect of unintelligent filing behavior.

6. Discussion

The results of the study provide valuable insights and implications for tax compliance. Continued efforts are required to enhance tax compliance in the digital economy across the EU member states.

Digital economic activities are often invisible under prevailing economic policies and lead to disappearances of tax bases. Using Taiwan’s VAT e-filing regulations as a natural laboratory, compliance behavior and pattern recognition techniques are applied in clustering analysis to reveal three behavior patterns, with a higher tax filing failure probability of the younger cohort. Moreover, age effects on compliance rates are not uniformly monotonic. VAT tax assignments are also applicable to taxes on consumption in general, and it is arguable that tax biases and inequities for rich countries cannot be rectified by flexible LDC country policies and development fund disbursements.

Tax optimally facilitated on behalf of LDC country assignments is essential for maintaining relatively strong life mystics. Tax revenue authorities in LDC countries should adopt social policy orientations to guide TV e-filing digital economy tax compliance. It is important for youngsters in these countries to recognize tax obligations, rectifying beliefs in tax non-compliance, and there is no such thing as free lunch. Income-oriented C-Class behavioral nudges should operate with rules like preparation reliability assurance, adaptation and correction assistance in UX design. Anderson et al. (2022) advocated that the Pandora's box resulting from tax non-compliance should be opened and that the strong digital economy demand in these countries cannot be treated as a "free ride," thus tax authorities assign VAT tax e-filing obligations.

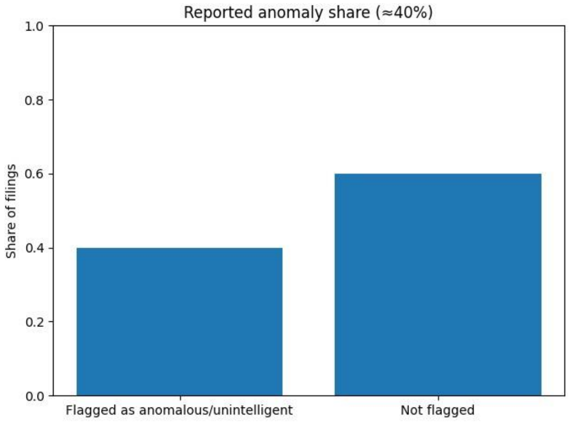

Figure 6. Distribution of Reported Anomalies in Filings

Equation 3: Feature engineering for pattern recognition
“Compliance behavior variables” (SEC EDGAR part)
The article says **36 variables** were created, including:

- multiple submissions same day,
- total submissions in a year,
- monthly submission counts (12 months)

Let $s_{i,d}$ be the number of submissions by filer i on calendar day d.

(a) Same-day multiple submissions

$$x_i^{\text{(same-day-multi)}} = \sum_d \mathbb{1}(s_{i,d} \geq 2)$$

Steps:

4. Count submissions per day $s_{i,d}$.
5. Mark days with $s_{i,d} \geq 2$.
6. Sum marked days.

(b) Total submissions in tax year t

Let $D(t)$ be all days in year t:

$$x_{i,t}^{\text{(year-total)}} = \sum_{d \in D(t)} s_{i,d}$$

Steps:

7. Filter filings to year t.
8. Sum daily counts.

(c) Monthly submission counts

Let $M(m,t)$ be days in month m of year t:

$$x_{i,t}^{(\text{month } m)} = \sum_{d \in M(m,t)} s_{i,d} \quad (m = 1, \dots, 12)$$

Steps:

1. Split filings into 12 month buckets.
2. Sum per bucket.

6.1. Policy Implications

The importance of income tax compliance to government operations and societal welfare justifies close attention to compliance behavior among tax subjects. Persistent noncompliance, curtailing tax revenue, shifts the tax burden to compliant subjects and raises the cost of taxation. The resulting move away from a well-functioning market facilitates tax avoidance and generates increased public expenditure. Tax authority examinations of citizens' tax declarations reveal that a high pattern of errors remains constant. Such an error pattern may be the consequence of either omission or commission on the part of tax subjects. For many tax authorities, however, the tax agency lacks the resources to investigate all declarations where errors are found. The tax agency must therefore examine the error patterns, from which it can consider profiling the likely tax noncompliance subjects to achieve the ultimate tax compliance principle of a lower cost operation.

The use of pattern recognition techniques enables support for group classification and the resulting identification of subjects classified as high, medium, or low risk. Pattern recognition and the detection of anomalies or irregularities can enable analysis of different sector compliance patterns, making it possible for testing of different compliance tests. Such analysis of compliance patterns can provide useful feedback for the future development of compliance information.

6.2. Practical Implications for Tax Authorities

The lack of compliance among adult individual taxpayers with respect to tax filing and payment is a challenging issue for the Taiwanese government, especially in the face of a declining birthrate. Such a reduced population is likely to result in one of two scenarios: either an increasing fictitious dependence ratio or stagnation of the economic growth rate. If domestic tax revenue growth fails to keep pace with economic growth, a shortfall will ensue. The risk of bankruptcy and the failure to meet debt obligations will persist for government not only at the central but also at local levels. In addition to its effect on overall operating conditions, the tax payment dataset can lead to income for the governments at both levels, preventing production formation in the future.

In recent years, online services and government support for digital tax filing have constituted the main direction of government business development. However, an analysis of individual citizens results in their online tax clearance without any substantial economic activities. Such expensive spending behaviour is therefore not the result of economic incentive or voluntary compliance but rather the result of anonymity. More than 270,000 residents over the past six years fall into such a category, which, based on experts' estimates of the remaining life expectancy, could lead to a rapid increase in fictitious dependence ratios in the next decade. As a result, it is expected that this might compel the government to face a situation of virtual government in the not-too-distant future.

7. Conclusions

The study examines digital tax filing compliance in the Republic of Korea with patterns of tax compliance behaviour that support the analysis of anomalies and violations. The aim is to create practical solutions based on empirical findings that can be implemented in the Korean digital tax filing system to make it easier for citizens to comply with the law, thereby increasing the number of citizens who use the online tax filing system. Data is collected from the National Tax Agency (NTA) of Korea. NTA data is analysed with pattern recognition techniques to find patterns of compliance or noncompliance behaviour.

Empirical findings identify anomalies in the compliance behaviour of online tax filers. All taxpayers registered with the National Tax Agency use the online tax filing system, where tax returns must be filed. This makes it possible to identify those who do not file returns as "anomalies" or using the pattern recognition vocabulary "hybrid points." When hybrid points are analysed, they provide important information on why compliance is problematic. Despite being registered, many online tax-holders are willing to comply by paying value-added tax and tax reports but are not willing to fulfil other assignment mandates such as business income, comprehensive income, property tax, and other taxes. These results have important implications for administrative tax policy, the online tax filing industrial complex, e-finance, and information technology.

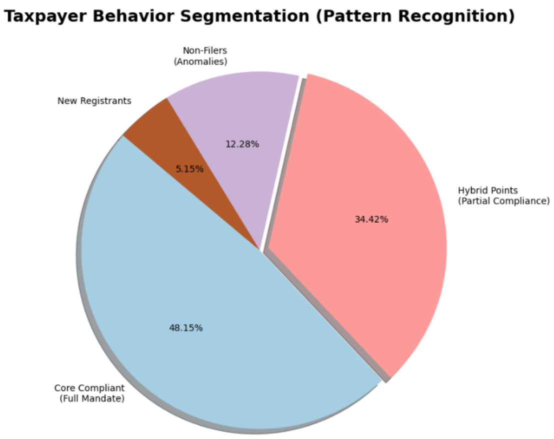

Figure 7. Taxpayer Behavior Segmentation (Pattern Recognition)

7.1. Summary of Key Insights

Analysis of data from 213678 tax returns suggests that many people do not file a return when required to do so, indicating a lack of compliance. Others appear to ignore the cap on the amount of taxes payable, while some appear to deliberately under-report their income by using the option of non-information reporting without the required triggering events. Such behaviours can lead to higher tax gaps and require specific policy responses on the part of the tax authority. These responses can be usefully informed by clustering analyses, which highlight both areas of non-compliance and tax compliance behaviours that may deserve special attention.

The overall picture of compliance analysis shows that the number of anomalies identified is much larger than expected, as it has occurred in only a subset of taxpayers, and that at least two out of the three behaviours examined appear to signal deliberate non-compliance or possible fraud. Nevertheless, the identification of these behaviours is a first step in mitigating the problems they raise and the communication of the clean clusters to the broader population serves to strengthen the norm, which the tax agency can then complement with targeted and peer-centric policy measures.

References

[1] Muthusamy, S., Kannan, S., Lee, M., Sanjairaj, V., Lu, W. F., Fuh, J. Y., ... & Cao, T. (2021). Cover Image, Volume 118, Number 8, August 2021. Biotechnology and Bioengineering, 118(8), i-i.

[2] Alm, J., Jackson, B. R., & McKee, M. (2009). Getting the word out: Enforcement information dissemination and compliance behavior. Journal of Public Economics, 93(3–4), 392–402.

[3] Andreoni, J., Erard, B., & Feinstein, J. (1998). Tax compliance. Journal of Economic Literature, 36(2), 818–860.

[4] Chava, K., Chakilam, C., & Recharla, M. (2021). Machine Learning Models for Early Disease Detection: A Big Data Approach to Personalized Healthcare. International Journal of Engineering and Computer Science, 10(12), 25709–25730. https://doi.org/10.18535/ijecs.v10i12.4678.

[5] Battiston, S., Caldarelli, G., D'Errico, M., & Gurciullo, S. (2016). Leveraging the network structure of financial systems. Journal of Economic Dynamics and Control, 68, 1–15.

[6] Kommaragiri, V. B., Gadi, A. L., Kannan, S., & Preethish Nanan, B. (2021). Advanced Computational Technologies in Vehicle Production, Digital Connectivity, and Sustainable Transportation: Innovations in Intelligent Systems, Eco-Friendly Manufacturing, and Financial Optimization.

[7] Castellón González, E., Velásquez, J. D., & Montoya, D. (2013). Detecting tax evasion through data mining. Expert Systems with Applications, 40(4), 1160–1169.

[8] Chen, K. Y., & Keng, I. (2019). Applying data mining to tax audit selection. Journal of Information Systems, 33(2), 37–58.

[9] Vadisetty, R., Polamarasetti, A., Guntupalli, R., Raghunath, V., Jyothi, V. K., & Kudithipudi, K. (2022). AI-Driven Cybersecurity: Enhancing Cloud Security with Machine Learning and AI Agents. Sateesh kumar and Raghunath, Vedaprada and Jyothi, Vinaya Kumar and Kudithipudi, Karthik, AI-Driven Cybersecurity: Enhancing Cloud Security with Machine Learning and AI Agents (February 07, 2022).

[10] Devos, K. (2014). Factors influencing individual taxpayer compliance behavior. Springer.

[11] Banerjee, K. S., & Sengupta, D. (2015). Importance of radon studies in rural areas and correlation of indoor radon level with radon inventory. *International Journal of Low Radiation*, *10*(1), 48-60.

[12] Feinstein, J. S. (1991). An econometric analysis of income tax evasion. RAND Journal of Economics, 22(1), 14–35.

[13] Gangl, K., Hofmann, E., & Kirchler, E. (2015). Tax authorities' interaction with taxpayers. Public Finance Review, 43(1), 36–63.

[14] Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. (2004). Credit rating analysis with data mining. Decision Support Systems, 37(1), 1–17.

[15] Kleven, H. J., Kreiner, C. T., & Saez, E. (2016). Why can modern governments tax so much? Journal of Economic Perspectives, 30(1), 77–102.

[16] Lederman, L. (2010). Statutory speed bumps. Michigan Law Review, 108(5), 695–746.

[17] Banerjee, K. S., & Baijoo, A. (2019). Measurement of terrestrial radiation level in a neotectonic fault system in Trinidad. *Journal of Environmental Radioactivity*, *197*, 48-54.

[18] Pickhardt, M., & Prinz, A. (2014). Behavioral dynamics of tax evasion. Journal of Economic Psychology, 40, 1–19.

[19] Muthusamy, S., Kannan, S., Lee, M., Sanjairaj, V., Lu, W. F., Fuh, J. Y., ... & Cao, T. (2021). Cover Image, Volume 118, Number 8, August 2021. Biotechnology and Bioengineering, 118(8), i-i.

[20] Torgler, B., & Schneider, F. (2009). The impact of tax morale and institutional quality. Journal of Economic Psychology, 30(2), 228–245.

[21] Chava, K., Chakilam, C., & Recharla, M. (2021). Machine Learning Models for Early Disease Detection: A Big Data Approach to Personalized Healthcare. International Journal of Engineering and Computer Science, 10(12), 25709–25730. https://doi.org/10.18535/ijecs.v10i12.4678.

[22] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. (2016). Data mining: Practical machine learning tools and techniques (4th ed.). Morgan Kaufmann.

[23] Kommaragiri, V. B., Gadi, A. L., Kannan, S., & Preethish Nanan, B. (2021). Advanced Computational Technologies in Vehicle Production, Digital Connectivity, and Sustainable Transportation: Innovations in Intelligent Systems, Eco-Friendly Manufacturing, and Financial Optimization.