TECHNICAL ASSISTANCE REPORT

SWEDEN

Corporate Income Tax Gap Estimation Based on Operational Audits

MARCH 2025

Prepared By
Soren Pedersen and Tobias Gabel Christiansen

Authoring Department
Fiscal Affairs Departments

This technical assistance (TA) was provided with financial support from the European Union.

International Monetary Fund, IMF Publications
P.O. Box 92780, Washington, DC 20090, U.S.A.
T. +(1) 202.623.7430 • F. +(1) 202.623.7201
publications@IMF.org
IMF.org/pubs

Contents

Figures

Tables

Annexes

Abbreviations and Acronyms

CD	Capacity Development
CIT	Corporate Income Tax
EC	European Commission
EU	European Union
FAD	IMF Fiscal Affairs Department
GDP	Gross Domestic Product
IMF	International Monetary Fund
LASSO	Least Absolute Shrinkage and Selection Operator
PIT	Personal Income Tax
SE	Shadow Economy
SEK	Swedish Krona (currency of Sweden)
Skatteverket	Swedish Tax Agency
SME	Small and Medium size Enterprises
VAT	Value Added Tax

Preface

In response to a request from the Swedish Tax Agency (Skatteverket), a capacity development (CD) mission team comprising of Mr. Soren Pedersen (FAD) and Mr. Tobias Gabel Christiansen (FAD short-term expert), carried out an in-country mission in Nyköping, Sweden during the period January 6 – January 17, 2025. The purpose of this mission, financed by the European Union, was to assist Skatteverket to estimate the corporate income tax (CIT) gap based on bottom-up techniques using operational audits for the income years 2016-2023.

This is a follow up to a diagnostic mission in February 2024 that assessed the current capacity of Skatteverket to measure the shadow economy (SE).[1] One of the recommendations from the previous mission was to measure the CIT gap based on operational audits. It was clear from the previous mission that Skatteverket has data on operational audits to be used in a CIT gap estimate.

Prior to this mission, Skatteverket prepared the required data consisting of CIT returns and audit results as well as background information on businesses such as sector codes and number of employees. Data was analyzed at the premises of Skatteverket. The report's findings are based on the data provided by Skatteverket.

The IMF team expresses its sincere appreciation to Skatteverket for the excellent cooperation and the excellent support provided before and during this mission. The team particularly acknowledges the excellent support provided by Mr. Hannes Jägerstedt, Statistician/Analyst, and Mr. Damián Migueles Chazarreta, PhD. The IMF team would also like to credit Skatteverket for the extremely comprehensive material they have made available for the analyses.

This report represents the draft report that was submitted to Mr. Damián Migueles Chazarreta and Mr. Hannes Jägerstedt, on February 1, 2025. It consists of an Executive Summary and the following sections: (I) Introduction; (II) Key Findings; and (III) Next Steps.

[1] This mission was part of IMF's so-called multi-pronged capacity development (see Annex III for details).

Executive Summary

This mission, financed by the European Commission, estimated the Corporate Income Tax (CIT) gap in Sweden based on a bottom-up approach using data from Skatteverket's operational audits carried out in the period 2016-2023.[2] The average yearly CIT gap across income years 2016-2023 is estimated to be 4.9 billion SEK or 2.2 percent of potential CIT liability.[3] The estimates are based on audited CIT returns. In comparison, Skatteverket estimated the CIT gap at 1.3 percent of potential CIT during this period using random audits. The difference in the results of the two methodologies is likely explained by the fact that the random audit program conducted by Skatteverket excludes businesses with turnover below 100,000 SEK. Instead Skatteverket uses the results from the random audits to estimate the tax gap for this group (by assuming their compliance is the same as those audited). The operational audit-based estimates in this report, on the other hand, include them directly in the estimations.

The CIT gap represented around 0.1 percent of the average GDP from 2016 to 2023. For the same period the CIT gap ranged between 1.8 and 2.6 percent of potential CIT liability. The yearly estimates suggests that the CIT gap has remained fairly steady over the period under consideration. In absolute terms, the CIT gap is estimated to be 6.0 billion SEK in 2023.[4] When all corrections from the audits of company records are included, the CIT gap + other corrections is estimated at 72.5 billion SEK which is equivalent to 25.8 percent of potential CIT. This estimate includes for example personal income tax (PIT) of the main shareholder of the company. An example of a PIT correction could be that the company has not correctly reported the main shareholder's use of a company car, which should be taxed as personal income and thus leads to a change in the main shareholder's personal tax. Auditing company records is the only way to uncover errors, such as those related to main shareholders' access to a company car.

As the next step, it is recommended that Skatteverket adopts the bottom-up methodology and continues to carry out the CIT gap estimation in 2025 and the coming years. Data analysts at Skatteverket who were closely involved in the mission can perform the analysis using the provided R code with detailed instructions. Alternatively, Skatteverket can request IMF capacity development (CD) for a workshop to train more analysts in Skatteverket in conducting an assisted self-assessment measure of the CIT gap based on operational audits.

The Skatteverket can also as a recommendation estimate a Social Security Gap (SSC) estimate based on the provided methods and the existing data regarding the PIT gap derived from the CIT estimates. It should be fairly straightforward for the Skatteverket to measure the SSC gap by adapting the delivered R scripts and documentation.

Skatteverket is recommended to assess the existing risk indicators by integrating them with random audits to determine if factors considered to be high-risk actually lead to high hit rates in the random audits. This could be done by grouping risk indicators into say three risk groups: high,

---

[2] This analysis measures the compliance gap, i.e., the policy gap and payment gap are excluded. For simplicity, "gap" and "compliance gap" will be used indistinguishably in this report.

[3] 2016 prices. When referring to CIT gaps in the following, these are also in 2016 prices.

[4] The CIT gap was estimated to be 5.7 billion SEK, 5.2 billion SEK, 4.7 billion SEK, 4.9 billion SEK, 4.9 billion SEK, 4.2 billion SEK, and 3.8 billion SEK in 2016-2022.

medium and low risk. Then for each group estimate hit rate and tax correction based on the random audits expecting higher hit rates and tax corrections in high-risk group compared to the two other groups.

As part of such an exercise, Skatteverket is recommended to produce a risk scoring model out of the random audit program and test it within the existing operational audit programs. This would result in an evidence-based risk model.

It is also recommended that Skatteverket revise its tax gap estimation for businesses with a turnover of less than 100,000 SEK in their published tax gap estimations. Currently, Skatteverket's random audit program, which is used to estimate the CIT gap, excludes these businesses. Compliance among businesses with less than 100,000 SEK is assumed to be the same as for those audited. This assumption does not hold because this group of businesses accounts for approximately 90 percent of the CIT tax gap identified in this mission based on operational audits.

It is recommended that Skatteverket develops a more detailed understanding of the factors contributing to the tax gap among businesses with a turnover of less than 100,000 SEK. Given that there are slightly over 166,000 businesses with turnover less than 100,000 SEK, audits alone will not be a feasible solution to address non-compliance. Since the Danish Tax Administration (DTA) faces similar challenges Skatteverket could undertake a study visit to DTA to exchange views and experiences on how to combat non-compliance for this group of businesses.

Recommendations

Estimate the CIT gap in the coming years		Due data
1	Adopt the demonstrated methodology and continue to carry out the CIT gap estimation in the coming years.	January 2026
2	Generate an SSC estimate based on the provided methods and the existing data regarding the PIT gap derived from the CIT gap estimates.	June 2025
3	Assess existing risk indicators by integrating them with random audits to determine if factors considered to be high-risk actually lead to high hit rates in the random audits.	June 2025
4	Produce a risk scoring model out of the random audit program and test it within the existing operational audit programs. This will result in an evidence-based risk model.	June 2025
3	Include a more comprehensive tax gap estimate for businesses with a turnover of less than 100,000 SEK in the published tax gap estimates, as these businesses account for approximately 90 percent of the total tax gap.	Next publication of tax gap in Sweden
4	Gain a better understanding of the factors driving the tax gap among businesses with less than 100,000 SEK in turnover in order to combat non-compliance since audits alone are not a feasible compliance tool given the large number of businesses	June 2026

I. Introduction

1. This project, financed by the European Commission (EC) under their DG REFORM Program, aims to provide advice to the Swedish Tax Agency (Skatteverket) to estimate the Corporate Income Tax (CIT) gap.[5] This estimate is based on a bottom-up approach using operational audits. Using random audits gives an unbiased estimate of the tax gap and by inference the shadow economy (SE). Operational audits on the other hand are likely to overestimate the SE since operational audits are selected based on risk criteria, leading to a focus on higher-risk businesses that are more likely to have significant non-compliance. One way for Skatteverket to move forward using operational audits is to use methods that correct for this so-called sample selection bias.

2. Skatteverket mentioned it has decided not to continue with random audits among small and medium enterprises (SMEs) and random audits will instead be more targeted going forward. This is therefore the right time for Skatteverket to use a model to correct for sample selection bias based on operational audits since the results from this model can be compared to the unbiased results from random audits.[6] This will give Skatteverket the possibility to evaluate how far a sample selection correction estimate of the SE is from the "true" estimate of SE from the random audits.

3. In this mission two different models were used to estimate the CIT gap for SME's.[7] One method to correct for sample selection bias is based on the economist (and Nobel Laurate) James J. Heckman's sample selection bias model.[8] A closely related method extends the Heckman Methodology and uses machine learning (ML) models to correct for sample selection bias.

4. The Heckman method is a two-stage procedure. In the first stage, it estimates the probability that a company is selected for audit. This is done using a probit model to predict CIT audit selection using risk indicators derived from company characteristics and information from tax returns. The second stage models the audit outcome (i.e., tax uncovered) using company characteristics (i.e., lines from the CIT return, sector, number of employees etc.) and a regressor that accounts for the selection process derived in the first step (see Annex I for more details).

5. The selection process for audits by Skatteverket's auditors involves two phases. Initially, Skatteverket uses a large number of risk indicators (500-600) to compute risk scores. High-risk companies are then isolated, and auditors select which businesses to audit based on their experience.[9]

6. In the first stage of the Heckman model, 38 risk indicators were utilized to model the selection for audits. The 38 variables were found among the 500-600 risk indicators by using Lasso (Least Absolute Shrinkage and Selection Operator) in a probit model to create a simpler and more efficient model by regularizing the coefficients and selecting the most relevant variables. The audit selection was modeled solely based on risk indicators because it was not possible to account for auditor

[5] See: https://reform-support.ec.europa.eu/index_en

[6] Based on the same population.

[7] SME's are defined as limited liability companies ("Aktiebolag" in Swedish) with up to 250 employees.

[8] James J. Heckman (1979). Sample Selection Bias as a Specification Error. *Econometrica*. vol. 47(1), pp. 153-161.

[9] This makes it difficult to predict whether a company will be selected for an audit, as modeling auditors' perceived risk of non-compliance can be challenging.

discretion and choices, as this information was unavailable.[10] In the second step a total of 48 variables consisting of tax lines and risk indicators were used to model the tax correction. The variables were also chosen using Lasso.

7. The second step of the Heckman method indicates limited unobserved sample selection bias. The coefficient of the inverse Mills ratio (IMR), which captures the degree of sample selection, is positive but insignificant. While the positive coefficient suggests that unobserved factors influencing audit selection are positively correlated with non-compliance, the insignificance implies that the unobserved selection effect is weak.

8. An alternative method to the Heckman method to estimate the CIT gap based on operational audits was also explored based on more advanced machine learning (ML) techniques.[11] Following Brewer and Carlson (2024), a control function approach similar to the Heckman Method is used to remove the effects of selection bias before training the ML model. Brewer and Carlson (2024) find that this approach reduces mean-squared prediction error in simulations.[12]

9. The ML model is based on the Random Forest algorithm.[13] The Random Forest algorithm is suitable as it effectively handles high-dimensional datasets by automatically capturing complex feature interactions and selecting the most important features in explaining non-compliance. A total of 500-600 risk indicators and around 200 tax return lines and company characteristics were used to predict the level of non-compliance (see Annex I for more details).[14]

10. Both the Heckman two-step estimator and the ML model's suitability in estimating the CIT gap depends on how audits are conducted. If audits focus narrowly on specific parts of a business, they may miss undisclosed taxes, leading to an underestimated CIT gap. Additionally, when audits target specific companies, such as a particular sector or type of firm, it becomes challenging to obtain a reliable estimate of the tax gap on the entire population. This difficulty arises from the increased extrapolation required between audited companies and the rest of the population. Finally, precautions must be taken since a fraction of noncompliance could be undetectable, even under the best efforts of auditing. This could for example be “payments under the table”. This possibility could result in an underestimation of the CIT gap.[15]

11. The estimates derived from the Heckman two-step approach and ML model consider all adjustments made to company records during audits. This includes, for instance, the personal income tax (PIT) of the company's primary shareholder. An illustration of a PIT adjustment might involve

[10] In this context, several variables were selected as important in determining whether a company was audited but not selected as important in explaining the audit correction. These variables effectively serve as instruments, but consulting auditors to validate them is recommended.

[11] Dylan Brewer and Alyssa Carlson (2024). Addressing sample selection bias for machine learning methods. *Journal of Applied Econometrics.* 2024; 39: 383–400.

[12] While we use LASSO, a linear model, for variable selection in the Heckman model, this extension incorporates a more advanced machine learning algorithm (Random Forest) that capture non-linear relationships between the level of non-compliance and explanatory variables in a data-driven manner.

[13] Breiman, L. (2001), Random Forests. *Machine Learning*, 45, 5-32.

[14] The hyperparameters were tuned using 5- fold cross-validation using gird-search to maximize R-squared on the training data.

[15] There is a tax gap that cannot be detected with audits.

the company's failure to accurately report the primary shareholder's use of a company vehicle, which ought to be taxed as personal income, resulting in a modification of the primary shareholder's personal tax. This highlights that auditing company records is essential for identifying discrepancies, including those associated with the primary shareholders' use of a company car.

12. CIT Gap estimates were obtained for eight consecutive years. These estimates cover the latest available data for the income years 2016-2023. The total number of tax returns audited is 6,040.[16] Due to the limited annual number of audits, the yearly estimates are more uncertain. One percent of audit results were trimmed in top and bottom for each year to avoid a small number of audits contributes to a significant portion of the audit yield.

[16] Limited to risk-based audits that was completed by the mission start. In 2016-2023 the original number of audits used are 519, 1,406, 950, 735, 770, 739, 556, 365 respectively.

II. Key Findings

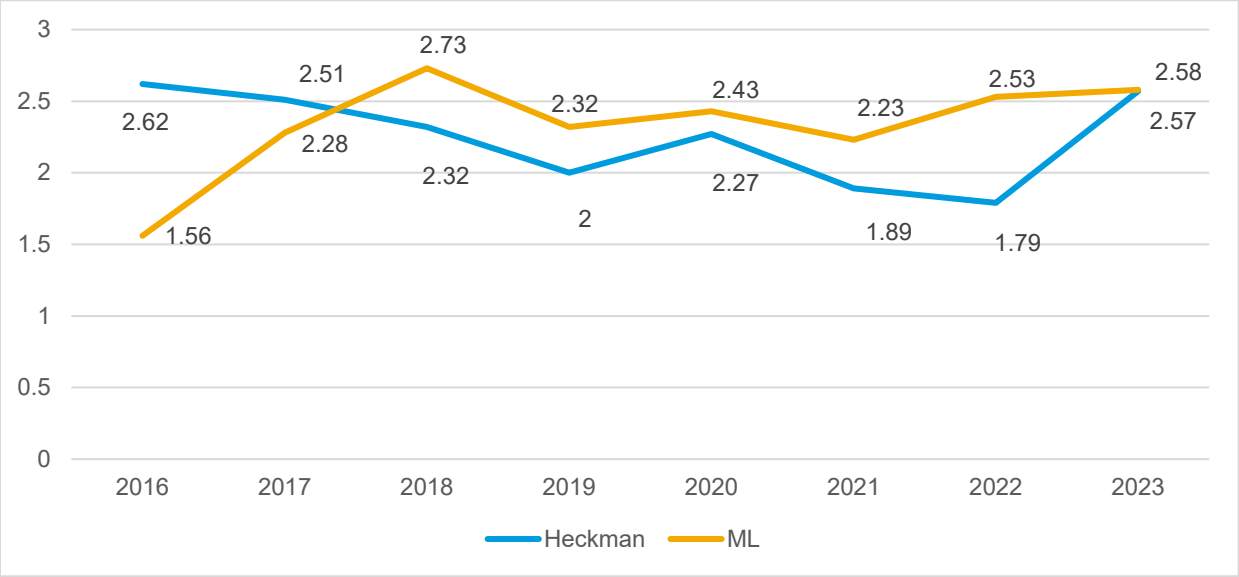
A. CIT Gap Results

13. The average yearly CIT gap across 2016-2023 is estimated to be 4.9 billion SEK[17] or 2.2 percent of potential CIT liability. The CIT gap corresponds to 0.1 percent of average GDP across 2016-2023. The CIT gap estimates are based on a Heckman model using operational audits conducted for the income years 2016-2023, i.e. covering a total of 8 years. A total of 6,040 audits, covering aspects of the company and carried out by Skatteverket over the eight-year period, were used to estimate the Heckman model. The results have been adjusted to account for the non-random selection of operational audits, addressing what is known as "sample selection bias", and one percent of audit results were trimmed in top and bottom for each year to reduce noise.

14. A machine learning model (ML) was also used to estimate the CIT gap. During the needs assessment mission in February 2024 and during preparation for the current mission Skatteverket asked for an estimate of the CIT gap based on ML.

15. The CIT gap based on ML is estimated to be 5.3 billion SEK which is very close to the 4.9 billion SEK based on the Heckman approach. The 5.3 **billion** SEK is equivalent to 2.3 percent of CIT liability. The yearly figures for the CIT gap in percent of liability from both models are shown in figure 1. The estimates are similar across years except from 2016.[18]

Figure 1. CIT gap estimated by Heckman two step approach and ML-model



Source: IMF calculations based on data from Skatteverket.

[17] Measured in 2016-prices.

[18] The difference in 2016 is likely to be driven by noise due to a low number of audits.

Table 1. Measures of CIT gap and CIT gap including other corrections

No. of employees	CIT gap alone.		CIT gap + other corrections	
	Billion SEK	Percent of CIT liability	Billion SEK	Percent of CIT liability
0	4.7	10.3	35.7	46.1
1-9	-0.4	-1.4	22.1	40.7
10-19	0.1	0.7	5.3	25.5
20-49	0.2	0.7	4.7	13.6
50-249	0.3	0.3	4.7	5.1
Total	**4.9**	**2.2**	**72.5**	**25.8**

Source: IMF calculations based on data from Skatteverket.

16. When all corrections from the audits of company records are included, the CIT gap is estimated at 72.5 billion SEK which is equivalent to 25.8 percent of potential CIT (Table 1).[19] "CIT gap alone" can be compared with Skatteverket's estimate based on random audits which is 1.3 percent of potential CIT.[20] [21] However, it should be noted that Skatteverket's CIT gap for businesses with a turnover of less than 100,000 SEK (not included in the random audits) is based on the assumption that their compliance is the same as for those included in the random audits with a turnover above 100,000 SEK. In contrast, the estimate presented here are based on all audits of SMEs with fewer than 250 employees. The smallest companies are included in these estimates because evidence from the Danish Tax Agency shows that small companies without employees account for about 60 percent of the tax gap. [22] This explains why the estimated CIT gap based on the Heckman two-step approach results in a higher CIT gap compared to Skatteverket's random audit-based CIT gap.

17. Another reason for the difference is that Skatteverket's random audits initially relied solely on the company's own accounting records, without access to control materials such as bank

[19] This estimate includes for example personal income tax (PIT) of the main shareholder of the company. An example of a PIT correction could be that the company has not correctly reported the main shareholder's use of a company car, which should be taxed as personal income and thus leads to a change in the main shareholder's personal tax. This is because auditing company records is the only way to uncover errors, such as those related to main shareholders' access to a company car. Based on Denmark's Random Audit program the Danish Tax Agency includes this type of error in the CIT gap estimate, cf. pp 66-70 in https://www.ft.dk/samling/20171/almdel/SAU/bilag/92/1839723.pdf. The denominator in the calculation of the 25.8 percent only includes CIT. Ideally relevant PIT should also be included in the denominator. However, it is very difficult to define the part of the PIT that should be included. For example, it would not be reasonable to include all withheld PIT of the employees in such a calculation.

[20] Tabell 2, p. 22 in https://www.skatteverket.se/download/18.48cfd212185efbb440b53b4/1678704377854/skattefelsrapport-2022.pdf

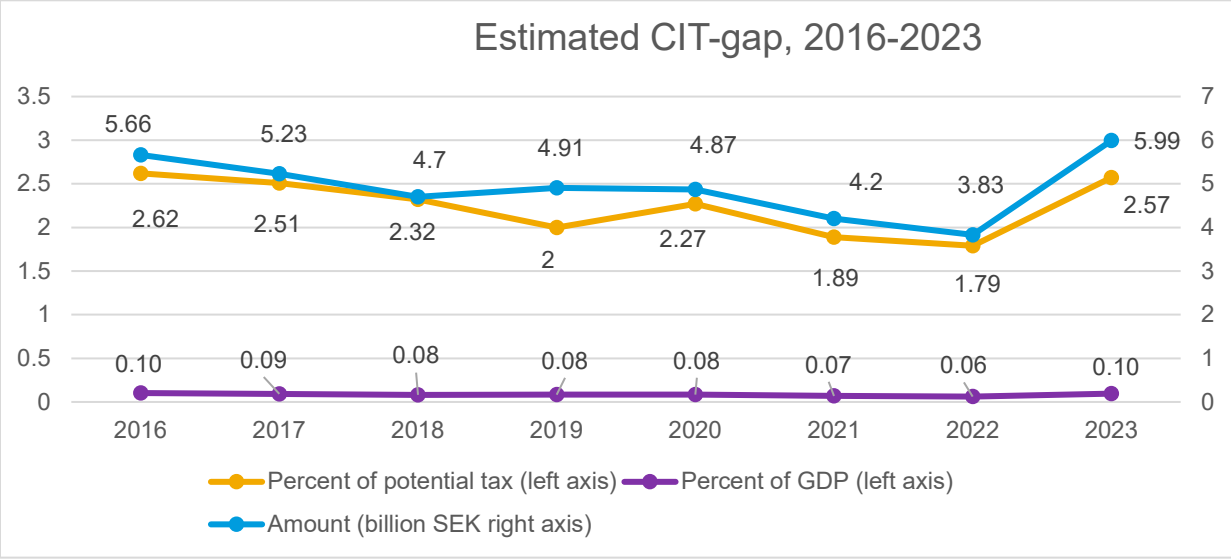
[21] The following results shown here are based on the "CIT gap alone" unless otherwise stated because this is the estimate published by Skatteverket.

[22] See p. 29 in: https://www.ft.dk/samling/20111/almdel/SAU/bilag/114/1069263.pdf

information.[23] If the initial check revealed deficiencies, deeper inspections were conducted to clarify the extent and size of the problem according to Skatteverket. With more tools, for example, if the auditors had access to information other than the company's own accounting records from the beginning, additional errors could likely have been discovered. However, such audits would also have required greater resources.[24]

18. The CIT gap in 2023 is estimated to be 6.0 billion SEK[25] or 2.6 percent of potential CIT liability (Figure 2). The CIT gap corresponds to 0.1 percent of GDP in 2023. It is important to interpret the annual fluctuations cautiously, as they primarily reflect changes in the economic structure of corporations, given that the Heckman model was developed using pooled data across years.[26]

Figure 2. Estimated CIT gap (2016-2023)



Source: IMF calculations based on data from Skatteverket.

19. The relative CIT gap is highest in the sector "Accommodation and food service" sector (Figure 3).[27] [28] As seen in figure 3, the second-highest relative CIT gap is in the "Manufacturing" sector.

[23] P. 16 in https://www.skatteverket.se/download/18.48cfd212185efbb440b53b4/1678704377854/skattefelsrapport-2022.pdf

[24] Ibid.

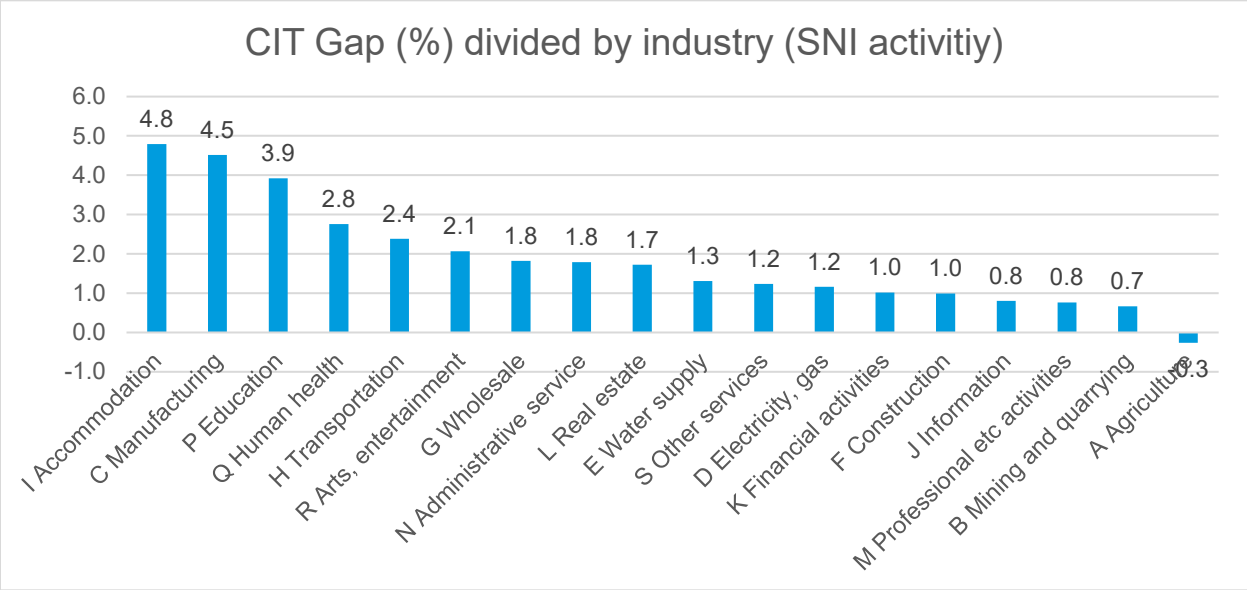
[25] Measured in 2016-prices.

[26] Notably, if companies were audited in multiple years, only a single tax assessment was recorded. This was split equally across the years that were audited.

[27] For each sector the CIT gap is shown in percent of sector CIT liability.

[28] Follows the Swedish version of Statistical Classification of Economic Activities in the European Community (NACE) called SNI (Standard för Svensk Näringsgrensindelning) in Sweden.

Figure 3. CIT gap in percent of CIT liability divided by sectors[29] (average across 2016-2023)



Source: IMF calculations based on data from Skatteverket.

20. There is a negative gap in the "Agriculture, forestry and fishing" sector. This is an average over the years 2016-2023. Skatteverket looked up a few audit cases to find an explanation. It appears that undeclared salaries are indeed present in some of these companies. The problem is that these salaries are not paid to the owner or main shareholder, and therefore, they do not appear in the PIT adjustment from the audit. Consequently, the following occurs: Deductible costs for salaries and SSC reduce CIT at the firm level. The main shareholder's tax liabilities remain unchanged. Employees' tax liabilities increase, but this is not reflected in the audit data. As a result, the total tax gap is recorded as negative, but it likely would have been positive if audit data had captured the full picture. [30]

21. The relative CIT gap is larger in corporations with few employees (Figure 4). It amounts to 10.3 percent of potential tax for corporations with 0 employees. The relative CIT gap declines as the number of employees increases and amounts to 0.3 percent for corporations with 50-249 employees.

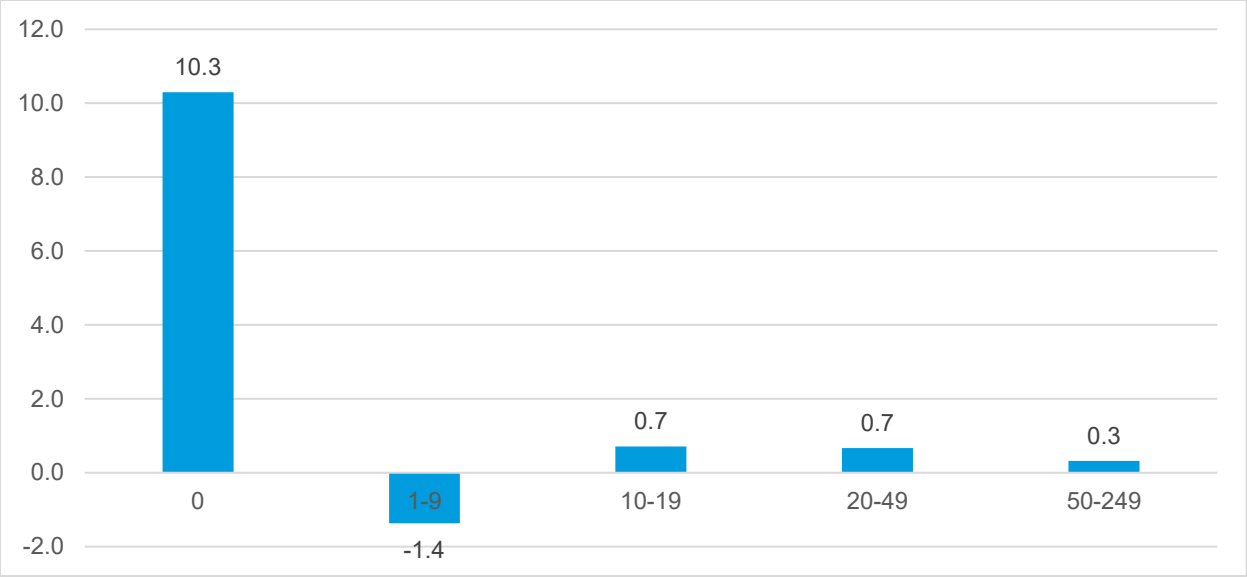
22. In absolute terms the CIT gap is 4.4 billion SEK in corporations with turnover under 100,000 SEK which is about 90 percent of the total tax gap. It is important to examine the reasons for non-compliance more closely in companies with less than 100,000 SEK in turnover. It will give valuable information for a compliance improvement plan to understand if non-compliance is due to deliberate evasion, lack of knowledge of the tax law or complicated legislation. The relative CIT gap is largest in corporations with turnover less than 100,000 SEK (Figure 5). It amounts to 15.3 percent of potential tax in corporations that have a turnover less than 100,000 SEK (approx. USD 8,950), while it is 0.5 percent in

[29] The sector shown is considered the primary sector based on turnover for a corporation if they operate in more than one sector. Sector codes are self-reported by corporations. The sectors "U Activities of extraterritorial organizations and bodies" and "O Public administration and defense; compulsory social security" are discarded due to limited number of observations.

[30] See Annex II with a concrete example from the audits that found five different errors in one audit. This shows that small (registered) businesses are participating in shadow economy activities like unreported salary payments, perhaps employing irregular persons, etc. In the Danish case there is a negative gap in 11.2 percent of businesses (corporations and self-employed), cf. p. 22 in https://www.ft.dk/samling/20111/almdel/SAU/bilag/114/1069263.pdf

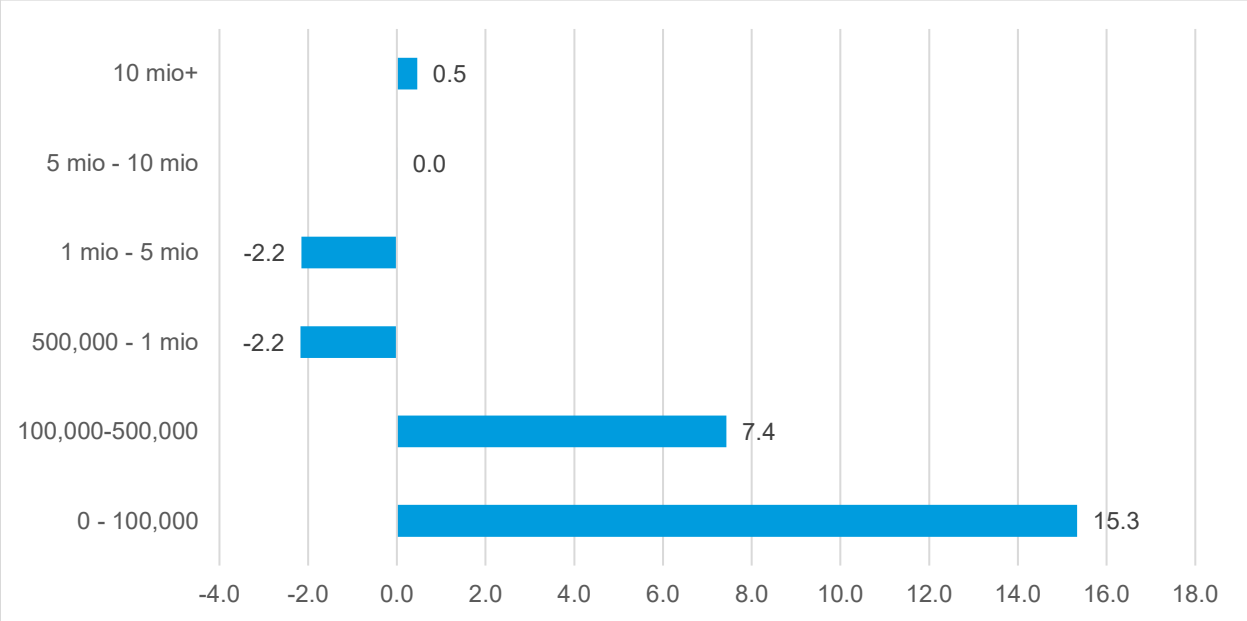
corporations with turnover above 10 million SEK (USD 895,000). It will be difficult for any tax administration to address non-compliance when the tax gap is dispersed among numerous small businesses. Audits is not an effective solution, and Skatteverket should explore alternative methods to tackle non-compliance among small companies. Gaining a better understanding of the factors driving the tax gap, will provide Skatteverket with valuable insights in this context.

Figure 4. Relative CIT gap by number of employees (average across 2016-2023)



Source: IMF calculations based on data from Skatteverket.

Figure 5. Relative CIT gap by turnover (average across 2016-2023)



Source: IMF calculations based on data from SRC.

23. Skatteverket could draw inspiration from other leading tax administrations and record whether a company's error is deliberate or unintentional. In the Danish Tax Administration, for example, when a random audit is completed, auditors' register on a rating scale from 0-6 the severity of non-compliance. This is used to divide taxpayers into unintentional and intentional non-compliers based on taxpayers' perceived willingness to comply with tax rules.[31] [32]

[31] See pp. 53-56 in: https://www.ft.dk/samling/20111/almdel/SAU/bilag/114/1069263.pdf for a description.

[32] Based on the Danish Random Audit program Christiansen, T. G., (2024), "Dynamic effects of tax audits and the role of intentions",*Journal of Public Economics*, 234, 105121 https://ideas.repec.org/a/eee/pubeco/v234y2024ics0047272724000574.html) shows how leveraging information provided by auditors on taxpayers' perceived willingness to comply can increase revenue collection in subsequent years after audits. This effect is solely driven by unintentional non-compliers who shows inattention or misunderstandings of the tax rules. This results in a revenue increase equivalent to 340 percent of initial corrections among unintentional non-compliers of the tax uncovered from the audit after 5 years. In contrast, intentional non-compliers who deliberately evade taxes and are typically targeted for operational audits do not respond to audits. He proposes using targeted and personalized guidance as a cheaper alternative to mitigate unintentional misreporting compared to expensive audits.

III. Next Steps

A. Set up a Team of Data Analysts and Apply the Developed Model

24. It is recommended that Skatteverket appoints 2-3 data analysts with responsibility to compile future tax gap estimations based on operational audits. To sustain the CIT gap models, it is recommended that at least 2-3 individuals be trained in the developed models. This will enable Skatteverket to independently conduct tax gap analysis on various taxes based on operational audits.

25. It is also recommended that the team invest a significant amount of time on working on the model. It is important to invest sufficient time to be able to perform good data analytics. Skatteverket is in a favorable position to apply the model since the analysts attached to the project are familiar with both data analytics as well as the required econometrics.

B. Other Recommendations

26. It is recommended that Skatteverket adopts the methodology and continues to carry out the CIT gap estimation in the coming years. Data analysts in Skatteverket that participated in the mission closely can perform this analysis using the provided R code with detailed instructions (see Annex IV). Alternatively, Skatteverket can seek IMF CD for a workshop to teach more analysts to be able to make an assisted self-assessment measure of the CIT gap based on operational audits.

27. Generate an estimate for Social Security Contributions (SSC) using the provided methods and the existing data on the Personal Income Tax (PIT) gap derived from the Corporate Income Tax (CIT) estimates. The Skatteverket should be able to easily assess the SSC gap by modifying the supplied R scripts and instuctions.

28. Produce a risk scoring model out of the random audit program and test it within the existing operational audit programs. Using random audits makes it possible to create a genuine evidence-based risk model.

29. Skatteverket should assess their existing risk indicators by integrating them with random audits. This approach will help determine if factors considered to be high-risk lead to high hit rates in the random audits.

30. Skatteverket should include a more comprehensive tax gap estimate for corporate businesses with a turnover of less than 100,000 SEK in their published tax gap estimates. Findings from this mission shows that these businesses account for approximately 90 percent of the total tax gap.

31. It is recommended that Skatteverket gains a more detailed understanding of the factors driving the tax gap among businesses with less than 100,000 SEK in turnover. With a bit more than 166,000 corporations with turnover less than 100,000 SEK audits will not be feasible to combat non-compliance. Since the Danish Tax Administration (DTA) faces similar challenges Skatteverket is recommend undertaking a study visit to DTA to exchange views and experiences on how to combat non-compliance for this group of businesses.

Annex I. Supplementary Material

Tax gap estimates from non-random risk-based audits are prone to sample selection bias due to the selection process being influenced by the perceived risk of non-compliance. Put differently, the audited companies are not representative of the general population of companies since they are selected as being more prone to risk of tax non-compliance based on several indicators. Hence the tax gap estimate based purely on such operational risk-based audits does not reflect that of the general population. A common approach to account for this is through the Heckman 2-step estimator.[33] The Heckman 2-step estimator corrects for sample selection bias by estimating both the selection process and the level of non-compliance (i.e., the tax uncovered from audit) in the same model. Following Wooldridge (2010)[34] the Heckman 2-step estimator is given by an outcome equation and a selection equation:

$$Y_i = X_i\beta + u_i \qquad (1)$$

$$S_i = 1[Z_i\delta + v_i > 0] \qquad (2)$$

Here equation (1) is the outcome equation, where Y_i measures the tax uncovered from audit and X_i is company characteristics (i.e., lines from the CIT return, sector, number of employees etc.)[35]. Next, equation (2) is the selection equation, where S_i is an indicator of audit, with $S_i = 1$ denoting company i was audited and $S_i = 0$ denoting company i was not, while Z_i are factors that determine whether a company is audited or not.[36] This includes 38 risk indicators used by Skatteverket to target audits.[37] Importantly, the value of Y_i is only observed if company i was selected for an audit ($S_i = 1$). Finally, u_i and v_i are independent of X_i and Z_i, with $v_i \sim N(0,1)$ and $E(u_i|v_i) = \gamma v_i$.[38]

What we are interested in estimating is $E[Y_i|X_i]$. However, since Y_i is observed only when $S_i = 1$, what we can estimate is $E[Y_i|X_i, S_i = 1]$. Using equation (1) and (2) this can be rewritten as:

$$E[Y_i|X_i, S_i = 1] = X_i\beta + E[u_i|v_i > -Z_i\delta] = X_i\beta + \gamma\lambda(Z_i\delta) \qquad (3)$$

Here $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ where $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function (pdf) and cumulative distribution function (cdf) of a standard normal distribution, respectively. The form of $\lambda(\cdot)$ follows from the assumption that $v_i \sim N(0,1)$ and it is labeled the inverse Mills ratio. Equation (3) presents a way to consistently estimate β. Following Heckman (1979) we can consistently estimate β and γ by regressing Y_i on X_i and $\lambda(Z_i\hat{\delta})$ using OLS, where $\hat{\delta}$ is obtained by estimating equation (2) using a probit model. Once an estimate of β has been obtained using the 2-step Heckman estimator, it can be used to construct an

---

[33] James J. Heckman (1979). Sample Selection Bias as a Specification Error. *Econometrica*. vol. 47(1), pp. 153-161.

[34] Jefrey M. Wooldridge (2010). Econometric Analysis of Cross Section and Panel Data. The MIT Press

[35] A total of 12 variables are used. These variables correspond to the 12 variables selected as the most important variables in a LASSO model developed to predict audit adjustments. In addition, PIT risk scores and PIT risk rankings are included.

[36] Data to estimate PIT/SSC gap stems from the CIT returns.

[37] In total, Skatteverket uses more than 500 risk indicators. To narrow down the most important ones, an L1-regularized probit model was estimated which resulted in 38 risk indicators.

[38] We only require v_i to be normally distributed. It is sufficient to assume that the conditional expectation of u_i given v_i is linear, which does not require u_i to be normally distributed.

estimator of the unconditional expectation of non-compliance (not conditioning on $S_i = 1$), given by $E[\widehat{Y_i|X_i}] = X_i\hat{\beta}$.[39] This can be applied to obtain predicted values of non-compliance for all companies in the population, and thereby the overall CIT gap. Table A1[40] presents an estimate of the selection and outcome models using all data on CIT returns and operational audits from 2016-2023. The selection model obtains an R^2 of 0.22 which indicates that the selection model fits the audit selection well.[41] A low R^2 in the selection model suggests that a substantial portion of the audit selection process remains unexplained by the included variables. This may indicate that tax authorities rely on additional criteria, discretionary judgment, or unobservable risk signals when selecting firms for audits. If the model fails to capture key determinants of audit selection, it may limit its ability to accurately reflect the selection strategy, potentially biasing the correction for selection effects in the outcome stage. Turning to the outcome model, it obtains an R^2 of 0.09. Ideally, we aim for this value to be as high as possible. A low R^2 in the outcome model indicates that a significant share of the variation in non-compliance remains unexplained. This is a common challenge in tax evasion models, as evasion behavior is influenced by firm-specific strategies, access to tax planning opportunities, and other unobserved factors. Therefore, due to the considerable diversity among companies and audit adjustments, reaching this goal is challenging.[42] Interestingly, the coefficient on the inverse Mills-ratio is insignificant (IMR in Table A2), indicating limited sample selection bias.

Two important points need to be highlighted. First, when using the Heckman 2-step estimator to predict non-compliance instead of inferring causal relationships, the accuracy of predictions depends on how well the selection model and the outcome model explain the data. Second, it is best to avoid using the same variables in both models. Doing so makes the outcome model's identification rely on the non-linearity of the inverse Mills-ratio, which can cause unstable results due to high multicollinearity. To prevent this, the selection model should include at least one variable that determines whether a company gets audited but doesn't affect non-compliance levels (known as an exclusion restriction).[43] However, finding such a variable can be tricky if audits are solely based on estimated non-compliance. In this context, a number of variables were selected as important in determining whether a company was audited but not selected as important in explaining the audit correction. These variables effectively serve as instruments, but consulting auditors to validate them is recommended.

In addition to the Heckman 2-step model, the machine learning approach by Brewer and Carlson (2024) was applied. Similar to the Heckman 2-step model, the selection model was estimated using an identical L1-regularized probit model, and the inverse-Mills ratio (IMR) was computed. The IMR were used to partial out the effect of sample selection bias from the variables in the outcome model e.g., $\widetilde{Y}_i =$

[39] Standard errors are wrong when manually estimating the 2-step Heckman estimator. Correct standard errors can be obtained using bootstrap.

[40] Due to hardware limitations, the population of companies was downsampled to 5% of the actual population. After fitting the models, the CIT gap for the full population was computed by rescaling the estimates.

[41] This is McFadden's Pseudo R^2. An alternative to McFadden's Pseudo R^2 is to calculate the percent correct predictions as a complementary goodness-of-fit index.

[42] A low R^2 indicates that the predicted $\hat{Y}_i$ can deviate quite substantially from the actual Y_i, leading to similar difference between $E[Y_i|X_i]$ and $E[\widehat{Y_i|X_i}]$ unless the errors are symmetrically distributed and therefore wash out in aggregation.

[43] In other words, exclusion restriction means that there must be at least one variable appearing with a non-zero coefficient in the selection equation but not in the equation of interest.

$Y_i - \hat{\lambda}\left(\hat{\Lambda}'\hat{\Lambda}\right)^{-1}\hat{\Lambda}'Y$ and $\tilde{X}_i = X_i - \hat{\lambda}_i\left(\hat{\Lambda}'\hat{\Lambda}\right)^{-1}\hat{\Lambda}'X$ where $\hat{\Lambda}$, Y and X are vertical stackings over i of $\hat{\lambda}_i$, Y_i and X_i, respectively. Then the partial led out variables and outcome are used in the following minimization:

$$\min_{\theta} \sum_{i=1}^{n} S_i L\left(\hat{f}\left(\tilde{X}_i, \theta\right), \tilde{Y}_i, \alpha\right)$$

where $L(\cdot)$ is the loss function which was set to Root-Mean-Squared-Error (RMSE). To model $\hat{f}(\cdot)$ the Random Forest algorithm was applied (Breiman, 2001)[44]. The Random Forest algorithm is suitable as it effectively handles high-dimensional datasets by automatically capturing complex feature interactions and selecting the most important features in explaining non-compliance. A total of 500-600 risk indicators and around 200 tax lines and company characteristics were used to predict the level of non-compliance. For computational efficiency, the Random Forest was tuned using 5-fold cross-validation using small forests consisting of 100 trees. The number of variables to consider at each split (mtry) and the minimum number of data points in a terminal leaf (min.node.size) were tuned. Once the optimal parameters were obtained, a new and large forest consisting of 1000 trees was fitted with optimal parameters. Figure A1 shows the 10 most important variables. The Random Forest obtained an R^2 of 0.18, providing a better fit of the data than the outcome model in the Heckman model. That said, the CIT gap estimated using both approaches are nearly identical, which corroborates the robustness of the results.

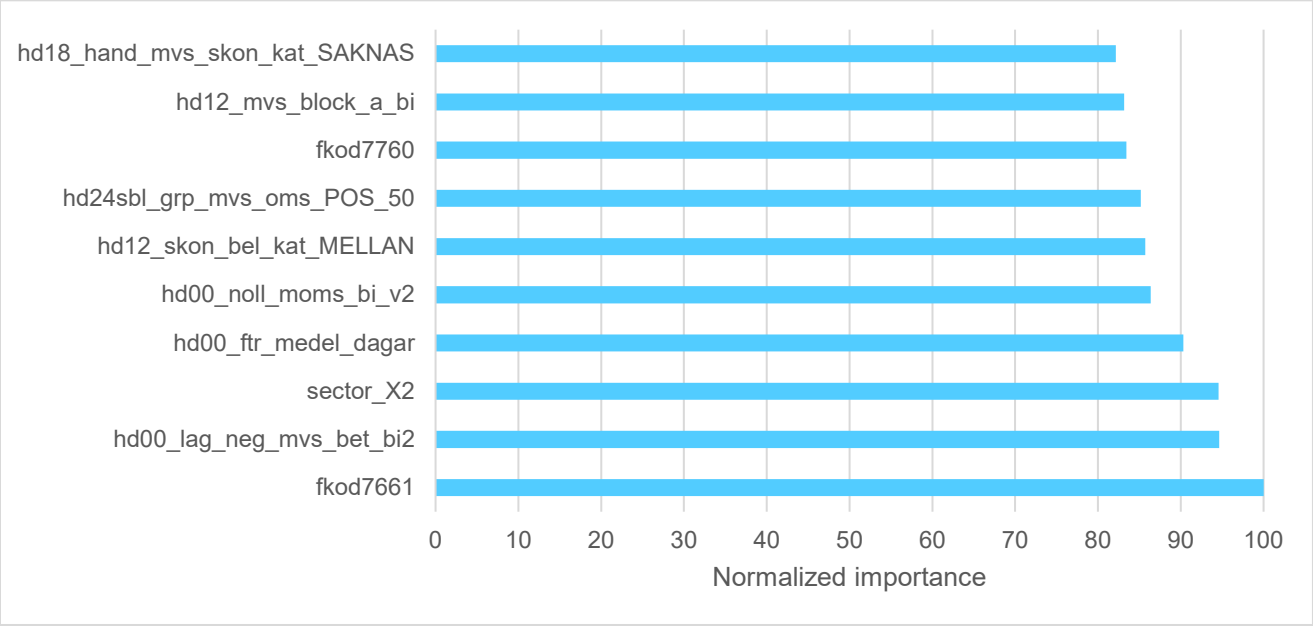
[44] Breiman, L. (2001), Random Forests. *Machine Learning*, 45, 5-32.

Table A1. Regression results

Selection Model			Outcome model		
Variables	**Coefficients**	**Std. Error**	**Variables**	**Coefficients**	**Std. Error**
(Intercept)	-2.823***	0.130	(Intercept)	4,080.374	17,901.138
score_mv	0.015***	0.001	lmr	525.923	3,157.326
score_vu	0.024***	0.009	hd00_adress_box_bi	6,500.520	4,326.936
a99_dot	-0.015***	0.003	hd12_saknade_perioder	12,217.002	15,890.195
p101	0.078***	0.011	hd24_m_saknade_perioder_andel	99,323.732	91,309.550
vu2_mima	0.442***	0.019	hd12_mvs_block_a_bi	-6,326.440*	3,422.781
hd00_bur	0.408***	0.025	hd00_alder_senaste_ag	-189.256*	102.739
hd00_urr	0.455***	0.031	hd00_noll_moms_bi_v2	40,668.365	25,457.880
hd00_gku_bruttolon	0.000	0.000	hd12_vu_upprep_0_mvs_bi	8,946.130	13,421.694
hd00_tot_lon_forman	0.000	0.000	hd00_lag_neg_mvs_bet_bi2	855,174.263	659,627.326
hd24_m_rattelse_andel	0.574***	0.065	hd00_bel_pens	0.062***	0.003
hd18sk_avg_skkonto	0.000***	0.000	hd00_sk_ku	8.826	12.590
hd18sk_stod_skkonto	0.000***	0.000	fkod7523	311.130	978.424
hd00_kk_Null	0.927***	0.098	fkod7216	0.002	0.002
hd00kk_poang_spec_bi_Nej	-0.422***	0.018	fkod7654	0.029	0.022
hd99_bas_ag_typ_kat_AG_sasong	-0.150**	0.060	fkod7752	0.220	462.932
hd99_bas_ag_typ_kat_Ej_AG_reg	-0.179***	0.020	fkod7760	0.010	0.014
hd00_gku_ant_ar0_1_kat_NULL	-0.226***	0.032	hd18_hand_mvs_skon_kat_SAKNAS	-14,764.606***	4,989.898
hd00_gku_ant_ar0_1_kat_Oforandrad	0.064	0.042	hd00_bransch_kat_scb_TILLVERKNING	17,812.873*	9,197.421
hd00_gku_ant_ar0_1_kat_Positivt	0.077*	0.040	hd99_avreg_mfa_kat_X180_365	29,273.901**	14,242.516
hd00_gku_ant_ar0_1_kat_SAKNAS	-0.367***	0.032	hd99_avreg_mfa_kat_Null	-5,832.421	5,862.363
hd00_gku_kostners_2ar_kat_J.J	0.155	0.118	hd00bas_org_sektion_X32	396,648.521	421,485.756
hd00_gku_kostners_2ar_kat_J.N	-0.289	0.183	hd09_oms_moms_3per_kat_OMS_STORRE_4MNKR	11,205.609*	5,806.008
hd00_gku_kostners_2ar_kat_N.J	-0.370*	0.201	hd00_avsl_fskatt_orsak_AVSL_FINNS	44,228.546**	20,821.876
hd00_aktiekap_kat_X1_10MN	0.037	0.051	hd00soa_ank_kat_Null	13,878.839*	8,165.558
hd00_aktiekap_kat_X10_9999MN	0.169***	0.064	hd00ink_sol_likv_grp_Likv_Sol_Noll	10,036.618*	5,170.610
hd00_aktiekap_kat_X100_500KKR	-0.169***	0.033	hd99_bas_ag_typ_kat_AG_sasong	-8,325.808	4,181.438
hd00_aktiekap_kat_X500_1000KKR	-0.217***	0.064	hd99_bas_ag_typ_kat_Ej_AG_reg	975.090	5,048.411
hd00_aktiekap_kat_NULL	0.160***	0.013	hd00_gku_ant_ar0_1_kat_NULL	-3,469.107	5,873.570
hd00_alder_rex_skuld_kat_X180_	-0.539***	0.114	hd00_gku_ant_ar0_1_kat_Oforandrad	-923.549	6,844.345
hd00_alder_rex_skuld_kat_X30_180	-0.226**	0.114	hd00_gku_ant_ar0_1_kat_Positivt	988.846	8,282.843
hd00_alder_rex_skuld_kat_SAKANS	0.179*	0.092	hd00_gku_ant_ar0_1_kat_SAKNAS	-2,735.759	6,052.358
hd36_revision_kat_REV_ej_noll	-1.001***	0.071	hd00_gku_kostners_2ar_kat_J.J	-2,358.181	11,515.353
hd36_revision_kat_REV_noll	-0.476***	0.067	hd00_gku_kostners_2ar_kat_J.N	1,975.651	15,410.809
hd36_revision_kat_REV_pag	-0.507	1.398	hd00_gku_kostners_2ar_kat_N.J	-16,445.736	11,771.033
hd06_inb_skm_kat_HOG	0.094***	0.017	hd00_aktiekap_kat_X1_10MN	6,125.528	19,881.458
hd06_inb_skm_kat_LAG	-0.197***	0.034	hd00_aktiekap_kat_X10_9999MN	-13,766.905**	6,290.934
hd06_inb_skm_kat_MEDEL	-0.096***	0.023	hd00_aktiekap_kat_X100_500KKR	2,954.601	8,873.258
hd06_inb_skm_kat_NOLL	-0.174***	0.024	hd00_aktiekap_kat_X500_1000KKR	31,977.931	29,160.852
			hd00_aktiekap_kat_NULL	6,673.077**	3,031.499
			hd00_alder_rex_skuld_kat_X180_	2,515.629	17,418.263
			hd00_alder_rex_skuld_kat_X30_180	-2,733.960	16,042.049
			hd00_alder_rex_skuld_kat_SAKANS	19,197.129	14,238.006
			hd36_revision_kat_REV_ej_noll	4,574.178	12,614.817
			hd36_revision_kat_REV_noll	-151.758	9,409.045
			hd36_revision_kat_REV_pag	15,851.300	20,559.498
			hd06_inb_skm_kat_HOG	3,462.577	2,908.037
			hd06_inb_skm_kat_LAG	1,941.783	8,570.652
			hd06_inb_skm_kat_MEDEL	623.376	4,102.759
			hd06_inb_skm_kat_NOLL	1,472.402	4,413.069
Number of observations	282,156			5,990	
R^2	0.22			0.09	

Source: IMF calculations based on data from Skatteverket. Note: 1) In 2016-prices. 1% of audit results were trimmed in top and bottom for each year to reduce noise. Standard errors are computed using bootstrap. For the selection model R^2 corresponds to McFadden's Pseudo R^2. ***, **, * denotes p < 0.01, p < 0.05, p<0.10.

Figure A1. Variable importance of the MLM



Source: IMF calculations based on data from SRC. Note: Variable importance is measured by "Gain," which measures the reduction in loss achieved by splitting on a particular variable. The variable importance has been normalized with respect to the most important variable.

Annex II. Example of an Audit from Skatteverket that Leads to a Negative Gap

Skatteverket has provided the example in the table below that explains a negative correction from an audit.

There are five different items that need corrections, shown in the first column. The next two columns show the change to the PIT and CIT bases, respectively. The last two columns show the resulting tax effect (around 42 % personal income tax rate and 22 % corporate income tax rate applied to the base columns).

To explain what is going on, the auditor revealed that the owner had deducted expenses for private insurance in company records resulting in a rise in PIT and deduction in CIT (because the expenses can be seen as wages paid to the owner).

Undeclared income (SEK 16,500 + VAT) was also revealed resulting in a rise in both PIT and CIT. The owner has also used the company car privately to the equivalent of 12,891 SEK. He (or she) is taxed +5,599 SEK in PIT. The firm now incurs an additional 2,165 SEK social security contributions, which are deductible, resulting in a 476 SEK *lower* corporate income tax.

The owner has also taken a loan in the company. This is taxed as dividends for the main shareholder/owner (as above this is seen as wages paid to the owner leading to a deduction in CIT).

Finally the audit also revealed a false invoice (”Credited invoice” below) registered in the financial statement leading to an unlawful deduction in company records.

Skatteverket enforces taxation on PIT but does not record the PIT impact of the audit in CIT. As a result, the total tax gap is recorded as negative, but it would have been positive if audit data had captured the full picture as shown in the case below.

Item	PIT base	CIT base	PIT effect	CIT effect
1. Private insurance	446	-586	194	-129
2. Undeclared income	20 625	16 500	8 959	3 630
2. SSC on undeclared income		-27 105	0	-5 963
3. Use of company car	12 891	-2 165	5 599	-476
4. Forbidden loan[45]	124 188	-163 207	53 943	-35 904
5. Credited invoice		100 000	0	21 999
	158 150	-76 563	68 695	-16 843

[45] “Forbidden loan” is forbidden in the meaning that it should be shown as dividends, not as a loan. It is not forbidden in the sense of being illegal.

Annex III. Measuring and Combatting the Shadow Economy: IMF Capacity Development Initiatives

There is no standardized definition of the "shadow economy", and other terms, such as "grey economy" or "informal economy", are used interchangeably. The OECD defines it as: "Economic activities, whether legal or illegal, which are required by law to be fully reported to the tax administration, but which are not reported, and which therefore go untaxed unlike activities which are so reported."[46]

The IMF has a multi-pronged approach to measuring and combatting the shadow economy by coordinating revenue administration, tax policy, and statistics components. Such an approach is useful in taking a holistic view of the challenges posed by the shadow economy. It allows crafting of policy and administrative responses that encourage a "whole-of-government" approach and coordination between various agencies such as tax, social security, and labor law enforcement. The operationalization of these approaches using structures such as dedicated shadow economy coordination units, behavioral economics for citizen engagement, and cross-agency working groups are being adopted by countries benefitting from IMF capacity development.

The IMF's Revenue Administration Gap Analysis Program (RA-GAP) provides capacity development for top-down and bottom-up techniques to estimate differences between potential and actual revenues. These include a well-established Value Added Tax Gap model that has an accompanying online training module and Corporate Income Tax gap estimation using econometric and machine learning techniques. These estimates can be used for tailoring Compliance Improvement Plans (CIPs) within a Compliance Risk Management (CRM) framework or for improving tax audit case selection.

The tax policy approach examines key design features of the tax system which may have a role in sustaining the shadow economy, particularly through high burdens or complex regulations. This could arise due to opaque or frequently changed tax rules, high income, consumption or payroll taxes, generous and multiple preferential tax treatments, a piecemeal approach to taxation of self-employment activities and the optional nature of various business taxation regimes.

The national accounts statistics approach envisages improving coverage to compile estimates of the non-observed economy—the part of the economy not covered by the basic data used to compile national accounts. These include the use of labor input method to estimate undeclared workers, data from administrative and non-statistical sources to estimate illegal activities, and use of tax audits to estimate under-reporting of income. This clearly loops back to the need for cooperation between revenue and statistical agencies which IMF capacity development fosters.

[46] See OECD, Shining Light on the Shadow Economy: Opportunities and threats (oecd.org), 2017.

Annex IV. Materials Left with Skatteverket

- R-code used to clean data used to estimate CIT gap.
- R-code used to estimate CIT gap using the Heckman method.
- R-code used for variable selection in selection model and outcome model
- R-code used to estimate the CIT gap using machine learning.
- PowerPoint presentation with key findings.
- R-files with tax gaps.