Tax Evasion across Industries: Bank Evidence from Greece

Nikolaos Artavanis
Virginia Polytechnic Institute
and State University

Adair Morse
University of Chicago

Margarita Tsoutsoura
University of Chicago

PRELIMINARY

March 12, 2012

Abstract

We use bank microdata on household credit to document the magnitude and incidence of tax evasion for Greece. We start from the new observation that in many developed and emerging countries, banks lend to tax-evading individuals based on the bank's perception of true income. This insight leads to a novel approach to estimate tax evasion from private-sector adaptation to a norm of semiformality. The paper has two main contributions. Replicating the bank's credit scoring of hard and soft information, we estimate a lower bound of 27 billion euros of unreported income for Greece. Our best estimate is that tax evasion is 41 billion euros. The foregone government revenues amount to 68 percent of the deficit for 2010. The second contribution is analyzing the industry distribution of tax evasion. Primary tax-evading occupations are lawyers, dentists, doctors, financial services, accountants, and other professional services. The industry distribution of tax evasion is not consistent with the theory that governments would want to subsidize the apprentice training of low-skilled workers. Instead, the industry distribution is consistent with a role for paper trails in hindering tax evasion for some industries and with politicians protecting their own occupations. We conclude by commenting on the property right of informal income and with suggestions for a business license fee to collect needed tax revenues.

1 Introduction

Even in developed countries, shadow economies are large, accounting for 14% of GDP generally in OECD countries and 25% in Greece (Schneider and Enste, 2000; Feld and Schneider, 2010). The literature on shadow economies suggests two main costs to countries with large informal sectors: lost tax revenues and inefficiencies of production due to the fact that informal businesses do not have access to formal capital markets. On the issue of lost tax revenues, the statement that tax revenues in Greece are insufficient to cover government obligations should not come as a surprise. One need not look past the front page of any recent newspaper to find the consequences. Unfortunately, Greece is not unique on this front.

On the issue of access to formal capital, however, developed and emerging countries may be different than less developed countries. Countries with generous social services often cultivate an environment of semiformality, in which individuals remain registered taxpayers, to receive the public benefits, but do not declare all of their income. But semiformality need not imply that the private sector excludes individuals from access to formal capital. Banks adapt to the culture of semiformality and provide credit to individuals based on their inference of true income. Borrowing the term from Harberger (2006), we call this bank adaptation.[1]

We motivate our study of tax evasion across industries with a table illustrating bank adaptation at work. The data are from a large Greek bank, covering tens of thousands applications by individuals for credit products.[2] Columns 1 and 2 show the monthly declared income and monthly payments on household credit products for self-employed individuals across different occupations, and column 3 presents the ratio of payments-to-income. On average, self-employed Greeks spend 80% of their monthly reported income servicing debt. To put this number in perspective, the standard practice in consumer finance (in the United States as well as Greece) is to never lend to borrowers such that loan payments are greater than 30% of income. And that is the upper limit.

The point of this table is to establish that adaptation is happening, and to motivate how

Table 1 shows evidence of adaptation in practice. Take the examples of lawyers, dentists, financial services, and accountants. In all of these occupation, the self-employed are paying over 100% of their reported income flows to debt servicing on consumer loans. This lending is no more risky; the default rate (column 4) on loans to lawyers, dentists, financial services, and accountants is no higher than on loans to people in occupations who on average are less burdened with consumer debt. The correlation between defaults and the ratio of debt payments to income is a small negative number.

Table 1 also offers a glimpse at where we are heading in the paper. Column 5 applies the rule that monthly payments can never be greater than 30% of monthly income to do a quick estimate of implied true income. From this inference, the implied tax-evaded income averages for lawyers, dentists, financial services, and accountants are all higher than €44,000 per person. In the main sections of the paper, we apply more formal models to estimate tax evasion by occupation, with the further goals of speaking to the magnitude and incidence of tax evasion and making sense of the occupational distribution.

The innovation of using bank data to estimate tax evasion is itself a contribution. We apply the general insight that because the private sector adapts to a culture of tax evasion, private sector data offer a window into the magnitude of, distribution of, and motivation for tax evasion. In general, estimating tax noncompliance is a challenging task, since tax evasion is an act that is meant to be hidden. The two main microeconomic methodologies to estimate tax evasion use audit and consumption survey data to infer true income. The first strand of the literature uses direct methods, employing audits of tax returns. The second set of methods indirectly estimate evaded income from observed expenditure data, building on Pissarides and Weber (1989), who use food expenditure survey data to estimate the underreporting of British self-employed. They find that on average the true income of self-employed is 1.55 times their reported income. Feldman and Slemrod (2007) use the relationship between reported charitable contributions and reported income, and find tax evasion among self-employed, nonfarm small-business and farm income are 1.54, 4.54 and 3.87 times reported income, respectively. [3]

[3] The consumption-based methodology has been applied in a host of settings. Andreoni, Erard and Feinstein(1998) and Slemrod and Yitzaki (2002) offer a comprehensive review of the literature. A separate literature relies on macroeconomic approaches to estimate the size of the black economy. The most common approaches are consumption methods (e.g., as in the electricity approach of Lacko (1999)) and the currency demand approach (Cagan (1958), Tanzi (1983)). These methods are best suited to estimate the size of the shadow economy, which emcompass but are not specific to income tax evasion. Sneider (2002) gives an overview of these methods, discussing their benefits and limitations and higlighting differences between the black economy estimates and income tax evasion.

One of the advantages to our approach is that it does not rely on survey data. Hurst, Li and Pugsley (2011) show that people underreport their income in surveys. Furthermore, people select whether they will participate in the survey, and tax evaders might be less willing to participate in a survey that asks questions on their income. Finally, designing a survey that is representative of the population is costly and difficult, and such surveys exist in few countries around the world. Similarly, although audit data are very detailed, they are extremely expensive to implement and frankly do not exist in most places except for the U.S. and Scandinavia (see Kleven, Knudsen, Kreiner, Pedersen and Saez (2011).

One of the ten largest banks in Greece provided us with individual-level data on the use of credit products – credit cards, term loans, mortgages, overdraft facilities and the like – for many tens of thousands of applicants. The application data include rich information on reported income, total debt outstanding, occupation, employment status (self-employed or wage earner), credit history, and demographics. We know the zip code of the borrowers, which allows us to construct soft information variables including wealth and the growth and variability of income.

When individuals apply for a bank credit or payment product, a bank officer applies a decision model to determine whether the individual qualifies. These credit scoring models utilize a host of credit history and socioeconomic characteristics as risk shifting factors, but by far the most important factor in determining credit worthiness is true income. Our methodology infers the bank's perception of true income from their decisions as to the credit capacity for a consumer or the debt-to-value allowable for a mortgage borrower. Since we know that the structure of the bank's adaptation model is occupation-specific and since we have a large dataset, we can estimate income multipliers specific to each occupation, under the assumption that wage earners are not tax evading. We subsample different types of loans and circumstances to address the selection in using application data.

Our main bank model finds 41 billion euros in evaded taxable income for 2010 just for the self-employed. GDP for 2010 was 234 billion euros, and the tax base in Greece was 86 billion euros; thus our magnitude is very meaningful. At the tax rate of 40%, the foregone tax revenues would account for 68% of the budget deficit shortfall in 2010 (or 45% for 2009). Our lower bound estimates, based on the population representative model find 27 billion euros of foregone taxable income with lost revenues which could have accounted for 44% of the deficit in 2010.

At the occupation level, we find a higher tax evasion multiple for the self-employed dentists and veterinarians, financial services, lawyers, doctors, accountants, and engineers. In terms of euros, the largest tax evaders among the self-employed are doctors, dentists and veterinarians, engineers, financial services, and lawyers. We reconcile this cross section of offenders with

a legislative bill that targeted nine select occupations. It lines up almost perfectly; the bill targets doctors, dentists, veterinarians, lawyers, architects, engineers, topographer engineers, economists, firm consultants and accountants. We also validate our finding against recent attention in the popular press about the ownership of Porsche Cayennes in Greek towns.

We then turn to making sense of the industry distribution. Although we cannot perform causal tests of theories, we can look for supporting evidence. We find no evidence that the distribution of industries reflects the government's intention to subsidize industries which provide apprentice-like training to unskilled workers. Instead, we find evidence, supporting that of Kleven, Knudsen, Kreiner, Pedersen and Saez (2011), that enforcement involves information and that when jobs involve paper trails, they are less likely to tax evade. We also find evidence of a political economy story; occupations which tax evade correlate positively with occupations represented in Parliament, even beyond lawyers.

Documenting that certain occupations tax evade does not solve the problem of collection. We suggest two solutions using our estimates. First, tax authorities could implement a *presumed income* approach, using our estimates to impose minimum taxes at the occupation-prefecture level. Second, we propose a *business license* fee (as implemented by city authorities world-wide), discuss its appealing characteristics, and applicability to specific occupations.

Our study concludes with thoughts on a property rights view of adaptation. Banks giving an entitlement to informal income provides a property right that allows individuals to use borrowing optimally to smooth lifetime consumption or overcome shocks. In this paper, we do not delve into the social welfare implications. However, because the observation of bank adaptation to semiformality is new, we conclude with thoughts on whether the haircut banks impose on hidden income in their lending should be zero, one, or somewhere in between, given a norm of tax evasion in the culture and political willpower of a country.

The remainder of the paper is as follows. Section 2 introduces our rich bank and tax authority data, and provides summary statistics by self-employed or not and by occupation. Section 3 lays out our methodologies. Section 4 reports results. Section 5 discusses validity, interprets magnitudes at the economy-level, and lays out the incidence of tax evasion. Section 6 investigate four theories to make sense of the distribution of tax evasion across industries. Section 7 is a discussion section on policy and welfare. Section 8 concludes.

2 Data

Our main data are proprietary files from one of the ten large Greek banks, which together account for eighty percent of the market share. The bank has tens of thousands of customers, with branches across the country. The dataset contains the universe of applications for consumer credit products and mortgages, both approved and rejected. Consumer credit products include term loans, credit lines, credit cards, overdraft facilities, appliance loans and cover the period January 2003 to August 2011. Mortgage applications cover the period January 2006 to August 2011.

The applications contain detailed information both about the credit requested and the applicant, including every piece of hard information that the bank observes at the time of the application. The administrative data including the date of the application, the branch office, the purpose of the loan, the requested and approved amounts and durations, and debt outstanding at all financial institutions. Thus, we know the individual's credit capacity as the combination of debt outstanding with all lenders plus the credit capacity allowed on the applied-for loan. Moreover we have detailed data on every person who was involved in the loan, as applicant, co-applicant or guarantor. We observe variables that convey information on the permanent income of the applicant, including personal income (as reported in the tax return and verified by the bank), occupation, employment type (wage worker or self-employed), education (for mortgages only), and age. Furthermore, we have information on each borrower's "worthiness", including years in job, years in address, homeownership, the status of the relationship with the bank (new customer, existing customer in good standing, existing customer in bad standing), the length of the relationship with the bank, and deposit holdings in the bank. The dataset also contains socioeconomic characteristics such as marital status and number of children.

Although we have the universe of applications for consumer loans, for our analysis, we focus on three subsamples, with dual aims in mind. The first aim is to isolate the supply side of credit by using a set of borrowers whose level of debt outstanding plus approved loans reflects being constrained in borrowing. Our estimation depends on being able to assert that the credit capacity is determined by the bank (supply) rather than customer demand. Thus, our *constrained sample* contains three sets of applicants: applicants whose requested loan amount is greater than the approved, overdraft applicants with less than 1,000 euros on deposit[4], and

[4] Overdraft facilities are issued either because the person in in distress and requests some slack or, perhaps inadvertantly, when a new customer opens a checking account or some other banking product. We filter based on individuals having 1,000 euros on deposit as a way to filter out individuals who have precautionary savings and are likely to be opening the overdraft as a part of opening or changing their banking products.

refinancing applicants.

Our other approach to focusing the analysis on constrained individuals is via a *mortgage sample*. Individuals who take a mortgage get very close to their credit capacity. The mortgage sample has another appealing characteristic, reflecting the second goal of subsampling, which is to be nationally representative. Our constrained sample suffers from the criticism that we are sampling on individuals more likely to be negative net worth. Even if we weight these people in income and location to the population, they are not fully representative. We turn to a mortgage sample because in Greece, 80% of households eventually end up owning homes. Sampling on the slice of the population that take out a mortgage does not restrict to net borrowers, since many homeowners borrow and save concurrently.

The third sample, the *population representative sample*, is defined as credit card applicants, overdraft provision applicants who do not need the facility, credit line applicants, and merchant loan applicants. In the years that we analyze there have been a lot of innovations in the use of payment systems, with credit cards in particular becoming increasingly used as means of payment. The purpose of the population representative sample is to select individuals who are independent of the need or desire to be at a bank lending office. By using a mixture of samples, we offer inferences robust across the spanning of being population representative and reflecting constrained individuals.

In addition to application data, we have detailed monthly records on the performance of each individual loan including installment amount, balance remaining, interest rate and any delays in repayment.

We supplement the bank data with detailed zip code level data from the Greek tax authority. For every zip code we have deciles of income for all tax filers as well as their classification in four employment categories: Merchants and Small Business Owners, Agriculture, Wage Earners and Self-Employed. To illuminate the detail of these data, for a population of 6 million tax filers, we have a breakdown of the number of filers and total income by 1,569 different zip codes, 10 national deciles of income and 4 professions. Each of the nearly 63,000 cells does not have many people observations in it.

We use the detailed income deciles per zip code data from the tax authorities to weight our sample to the population, aggregating to the quintile of income, four professions, and nine meta-prefecture level. For our analysis, we exclude students, pensioners and unemployed, since our goal is to focus on the active workforce.

We also use fine detail of these data to construct local per capita income growth defined at the four occupation-level crossed with income decile. We also calculate a measure of the

variability of this income growth. To construct a measure of the standard deviation of the growth of income in the cell, we have to take into account the difference in the number of people in the zip code-income decile-occupation. Thus, we use the standard error formula of the standard deviation divided by the square root of the observation count. These measures serve either as a proxy for individual income growth used by the bank as soft information or a direct measure of the soft information of local conditions. We proxy for the wealth of individuals in the zip code and occupation level in three ways. The tax authority provided us with presumed real estate values by zip code, which they used to determine property taxes, specific to the street level. We create an alternative measure of average car values by zip code and by occupation, using the bank's rich dataset of automobile loans.

Table 2 presents the mean statistics for the variables by sample and by employment status. The definitions of the variables are given in the Data Appendix. In mortgage decisions, financial institutions use loan-to-value models, therefore in the analysis of the mortgage applications the dependent variable is the ratio of the approved mortgage plus other outstanding debt relative to the value of the real estate property. (We use total debt because of the practice of splitting up mortgages in Greece.) In all three samples, the credit capacity is higher for the self-employed than wage workers. The reported income levels for the constrained and mortgage samples are if anything, lower. So even in a naive comparison of average income and credit capacity, the data show that self-employed have much higher levels of credit capacity, although they do not have higher reported incomes. Of course we are not able to derive conclusions from such a naive comparison, since, among other reasons, the distributions of income and debt outstanding might be different for self-employed and wage workers, and self-employed may have different risk profiles or growth prospects. In the next section we describe our empirical methodology that would address these challenges.

3 Methodology

We use two approaches to estimate true income from bank data, both based on a causal relationship between true income and the amount of credit a financial institution is willing to extend. The first methodology replicates the structure used by banks when they apply risk scoring models to determine individuals' credit capacity. The second methodology abstracts from what we know about bank structural models and instead uses bank data functionally by implementing a nearly saturated model of credit capacity determination.

The over-arching idea to the approaches is that we start from models of credit capacity

determination $c = f(\xi, X, S, \Theta)$ that are a function of true income ξ, risk scoring factors X, soft information S, and parameters Θ. We only observe reported income y, which we assume is equal to true income for wage earners, but possibly downward biased for the self employed. We adapt the credit capacity model such that we can estimate the cash flow relationship between true income and credit capacity off the wage earner observations, while parameterizing the impact of hard and soft information. Since one needs a certain amount of cash mechanically to service debt, the true income - to - credit capacity relationship should hold for all individuals. For the self employed who tax evade, observed reported incomes are lower than what would be needed to support the level of their debt servicing. We identify how much extra income the self employed would need to have in order to support these observed levels of credit. Thus, the essence of our methodology is the calculation of true income, conditional on all possible soft and hard information for each individual, as a function of reported income, the mechanical relationship between true income and credit capacity (estimated off wage-earners), and parameters which pick up the excess credit capacity observed for the self employed $E(\xi \mid X, S) = g\left(y, \widehat{\Theta}_{wage}, \widehat{\Theta}_{self\ employed}\right)$. The below subsections provide the details of these inferences.

3.1 Bank Structural Model

When a bank officer appraises an individual's application for a credit product, the objective is to minimize the risk of default while bearing in mind the potential for future profits. Banks first calculate the level of credit capacity supported by an individual's income and then score the person on a points system incorporating credit history, stability and socioeconomic characteristics that correlate with the bank objectives. We know that our bank adds up points across characteristics (e.g., age points plus credit history points) and has a non-cardinal scoring of points within characteristics (e.g., with age bracket dummies rather including age as a continuous variable). We know all of the hard information variables and include them nonparametrically in a "kitchen sink" approach to recreate the credit scoring.

The bank credit capacity model can be written:

$$c_i = \beta_1 \xi_i + X_i' \beta_2 + S_i' \beta_3 + \mu_{\cdot\cdot k} + \varepsilon_{ijk}. \tag{1}$$

We use three levels of indexing: i denotes the individual applying for a credit product; j denotes the occupation; and k denotes employment status, being either *wage worker* or *self employed.* Credit capacity c_i is a function of true income ξ_i, as well as (hard information) scoring factors X_i and soft information variables used by local bank branch officers S_i.[5] We write the model

[5] Credit capacity (c_i) itself is a combination of debt outstanding plus the credit capacity approved on the

as a cross section and embed time dummies in X to incorporate supply changes to the credit model.

An important risk scoring variable is self-employment, which we write down explicitly with the self employment fixed effect $\mu_{\cdot\cdot k}$. Repayment risk might be higher for the self-employed than wage workers because of higher uncertainty in the self-employed' income and because of the possible use of consumer loans to finance business activities. The bank must balance this additional repayment risk against prospects for profits from additional services, which could be larger for the self-employed. It is unclear which effect should dominate, but in any case, our objective is simply to absorb the overall implication.

True income ξ_i is the most important component of any bank's determination of credit capacity. Yet the bank observes only reported income, y_i, which is downward biased. Before we delve into that adaptation, it is worth asking whether, if we knew true income, ξ_i, we could write the debt capacity scoring model without the error term ε_{ijk}. The answer is not quite. Credit decisions allow for the use of soft information at the local level. The use of soft information S_i may just be noise, but it potentially could bias our occupation-level estimates of tax evasion. We come back to this point after we incorporate adaptation into the model.

In Greece and many other countries, banks cannot remain competitive by lending only off reported income. Instead, banks adapt by inferring true income, ξ_i from observables. We discussed this process of adaptation with a number of banks across southern Europe and learned that adaptation is a prevalent and long-established process. Banks use years of experience to fine tune their adaptation model to be a best guess of true income. We have to be careful in our use of the word *true income* in that banks might apply a haircut on the how much credit the tax-evaded portion of true income supports, to the extent that they deem tax-evaded income to have more risk.[6] Because the credit capacity decision in the bank data reflects this potential haircut taken, it is not an econometric problem for us, but it is important to note that all of our estimates of true income thus are estimates of reported income plus haircutted tax evaded income, and thus are underestimates.

The bank's estimate of haircutted true income ξ_i consists of two pieces: a corporate multi-

applied-for loan. Since the new credit approved is the marginal addition to credit capacity, we assum that all credit capacity (old loans plus new capacity) is equivalent in bank scoring. The ability of income to support debt servicing is not particular to the origin or ordering of debt.

[6] The haircut likely varies by occupation reflecting the volatilities of the sources of reported versus hidden income. It could be, for example, that certain occupations (e.g., engineering) have income is contract-based and thus very reliable and reported, but side-jobs reflect business cycle excess demand. Other occupations (e.g., retail) may systematically report a certain portion of sales.

plier function m_{jk} on reported income y_i and a local bank officer soft information adjustment for an individual i, s_i:

$$\xi_i = m_{jk} y_i + s_i, \tag{2}$$

The actual corporate adaptation model is very simple: banks apply an occupation multiplier to scale up reported income for the self employed:

$$m_{jk} = \begin{cases} 1 & for\ k = wage \\ \lambda_j & for\ k = SE \end{cases}. \tag{3}$$

The λ_j's are the occupation-specific multipliers mapping the self-employeds' reported income to true income.

The soft information component is a bit more subtle. The econometric concern is that soft information variables, particularly permanent income variables, may cause the bank to change its assessment of an individual's unseen true income in a way that is correlated with reported income, or any of the hard or other soft information variables. The most concerning of such variables are wealth and realized growth in one's local occupation. For instance, a bank officer may observe wealth via a car or an address or by knowing a family name and adjust the perception of true income. Likewise, if an individual is in a growth area fuelled by construction in the local economy, a bank officer may infer more true income.

Denoting wealth and local economy income growth by w_i and g_i respectively, we can write:

$$s_i = \gamma_w w_i + \gamma_g g_{ijk} + \nu_{ijk}, \tag{4}$$

where ν_{ijk} is soft information noise in the implementation of the adapting reporting income, under the assumption that once we condition on wealth and the local economy, other omitted soft information variables in bank officer adaptation are noise. An example of noise might be the bank officer's personal knowledge of a bank customer's clients or religious preferences.

To put these pieces together, we can go back to the big picture. When a local bank appraises a credit product application, the officer applies the corporate model (that incorporates both scoring and adaptation) and then potentially makes two adjustments based on soft information, one in adaptation and one in assessing repayment risk. Collapsing the pieces of adaptation into the baseline equation (1) leads to:

$$\begin{aligned} c_i &= \beta_1 y_{i,wage} + (\beta_1 \lambda_j)\, y_{i,,j,SE} + X_i' \beta_2 + \mu_{\cdot\cdot k} + \mu_{\cdot j \cdot} + \mu_{\cdot jk} + \\ &\quad + (\beta_1 \gamma_w + \beta_{3w})\, w_i + \left(\beta_1 \gamma_g + \beta_{3g}\right) g_{ijk} + \beta_{3,sd} sd_i + \zeta_{ijk}. \end{aligned} \tag{5}$$

Whereas use of soft information in adaptation reflects perception of true income, use of soft

information in the credit score equation reflects the local bank's assessment of repayment risk. In writing equation (5), we have broken out the soft information variables into five pieces:

$$S_i'\beta_3 = \mu_{\cdot j\cdot} + \mu_{\cdot jk} + \beta_{3w}w_i + \beta_{3g}g_i + \beta_{3,sd}sd_i \tag{6}$$

Repayment risk may vary systematically by occupation; we address this by including fixed effects for occupation $\mu_{\cdot j\cdot}$ and occupation crossed with self employment $\mu_{\cdot jk}$.[7] Repayment risk may be correlated with wealth, which we capture with $\beta_{3w}w_i$. Finally, repayment risk may be related to the bank's expected growth rate g_i or the variability (standard deviation) of past growth, sd_i, in the local sector in which the applicant is employed. We measure these latter two variables with the lagged growth rate and the time series standard deviation for the small bucket people in the zip code, income quintile, and tax authority assignment of occupation.

Re-parameterizing sets up our bank model estimating equation:

$$c_i = \beta_1 y_{i,wage} + \alpha_{1j} y_{i,j,SE} + X_i'\beta_2 + \mu_{\cdot\cdot k} + \mu_{\cdot j\cdot} + \mu_{\cdot jk} + \alpha_2 w_i + \alpha_3 g_i + \beta_{3,sd}sd_i + \zeta_{ijk}, \tag{7}$$

where the two reparameterizations are:

$$\begin{aligned}
(i) &\;:\; \alpha_{1j} = \beta_1\lambda_j \quad for\ j = 1, ..., J \\
(ii) &\;:\; \alpha_2 = \beta_1\gamma_w + \beta_{3w} \\
(iii) &\;:\; \alpha_3 = \beta_1\gamma_g + \beta_{3g} \\
(iv) &\;:\; \zeta_{ijk} = \beta_1\nu_{ijk} + \varepsilon_{ijk}.
\end{aligned}$$

Since the coefficient on the reported income of wage workers is β_1, we can identify the λ_j's using β_1 in conjunction with the coefficients on the reported income of the self-employed (the α_{1j}'s). Under the assumptions (a) that we are able to nonparametrically replicate the use of the hard information variables in the debt capacity formulation, (b) that bank officer implementation of the multiplier model is just noise beyond the effects of wealth and growth, (c) that the corporate adaptation model is a series of occupation multipliers for the self employed, and (d) that we

[7] A related story concerns the use of businesses to absorb some of personal consumption. What if, in certain occupations, proprietors can expense certain items as business use. In particular, we can think of cars. If the self employed pays for her car through the business and uses the expense to lower taxes, she might have more cash flow available to service debt for a given level of income. The occupation fixed effects interacted with self employment should solve this concern, unless the absorbing of personal consumption is correlated with income. Although it is easy to come up with a few items that proprietors can expense through the business (like lunches, office supplies, etc), it is hard to come up with substantial items that are tax expensible and correlated with income (after wealth is controlled for) other than cars. Because we are sensitive to the role of cars, we will include a measure of car value by occupation crossed with self employment in the specification.

observe constrained credit capacity, the residual term, $\zeta_{ijk} = \beta_1 \nu_{ijk} + \varepsilon_{ijk}$, will be uncorrelated with any of the independent variables, and the estimates of tax evasion will be consistent.

If no soft information is used in the adaptation model, the calculation of (haircutted) true income will just rely on the $\widehat{\lambda}_j$'s:

$$\widehat{\xi}_i = \begin{cases} \widehat{\lambda}_j y_i \ if \ k = SE, \\ y_i \ if \ k = wage \end{cases} , where \ \widehat{\lambda}_j = \frac{\widehat{\alpha}_{1j}}{\widehat{\beta}_1}. \tag{8}$$

If wealth [or local growth] affects the bank officer's assessment of true income, then we are in the situation of being able to identify α_2 [α_3] but not explicitly γ_w and β_{3w} [γ_g and β_{3g}]. However, we can identify a range for estimated true income, and this estimator should be consistent, using that structural assumption that wealth [growth] soft information can only cause a non-negative impact on the assessment of true income ($\gamma_w \geq 0$) and on credit capacity scoring ($\beta_{3w} \geq 0$). Thus, the range of true income for a self-employed in the soft information model is:

$$\begin{aligned} \xi_i^{Lower} &= \widehat{\lambda}_j + 0 \\ \xi_i^{Upper} &= \widehat{\lambda}_j + \frac{\widehat{\alpha}_2}{\widehat{\beta}_1}. \end{aligned}$$

3.2 Saturated Non-Structural Model

Rather than relying on confidence that we are replicating the bank structural model of credit scoring and adaptation, we can use the bank data in a functional econometric model. Our bank dataset, supplemented with the soft information variables we construct, contains just about everything the bank could use in credit capacity determination. We have sufficiently large number of observations such that we can implement a very saturated model of credit capacity (or loans-to-value, in the case of mortgages) with permanent income variables, credit history variables, socioeconomic variables, many of their interactions, and everything interacted with self-employment and reported income. After satisfactorily explaining as much of the variation as we can, we include occupation fixed effects and occupation crossed with self employment fixed effects. Our assumption is that any residual from our "whole kitchen sink" saturated model must surely be representative of tax evasion.

4 Results

Table 3, Panels A-C present the results from the bank structural model. Panel A analyses the constrained model. The dependent variable is credit capacity. In each estimation we include

a self-employment dummy, as well as the socioeconomic characteristics and credit worthiness variables described in Table 2. To isolate supply effect, we include time dummies. All estimations are weighted to the population using the tax authority data described in the data section. Robust standard errors are calculated. Lambdas are the ratio of the coefficient on income for each of the self-employed occupations divided by the coefficient for the wage worker.

Columns 1 and 2 report the estimates from the hard information model. In particular, column 1 presents the estimates on the reported income of the self-employed by occupation. Column 2 presents the calculated lambda. To give an example, in Table 3A column 1 self-employed accountants have a coefficient on income of 1.686 while the coefficient of income for wage workers is 0.438. This gives a lambda of 3.9 for self-employed accountants. We assume that wage workers do not tax evade so their lambda is 1.

Columns 3 and 4 presents estimates of the occupation fixed effect model. In this model we include fixed effects for occupation as well as occupation crossed with self-employment, to address concerns that repayment risk may vary systematically by occupation.

Columns 5-8 demonstrate the coefficients on income and multipliers λ of the wealth and growth information model. This model aims to address the econometric concern that soft information variables might change the bank's assessment of the applicant's unseen true income or repayment risk. The most concerning soft information variables are wealth and realized growth in one's local occupation, therefore in this specification we add our wealth proxies (car value mean zip level, car value mean by occupation, local occupation income growth based on tax authority data and standard deviation of income growth). Wealth and local growth may enter only the bank's assessment of true income, only the credit capacity scoring or both. The lambdas provide the range for estimated true income. Lambda 1 assumes that wealth and growth only enter the credit capacity scoring, lambda 2 assumes that wealth enters the adaptation and lambda 3 assumes that growth enters the adaptation. Finally the saturated model allows the wealth and growth variables to interact with occupation and occupation crossed with income to address the possibility that the use of soft information might vary systematically by occupation and income level. The last two columns aid the interpretation, converting the across-the-board significant coefficients and mean calculated λ multipliers into implied true income. The multiplier is large (around 7X reported income) for Lawyers. Dentists & Veterinarians, Doctors and Accountants are among the highest tax evaders.

In Table 3, panel B we use the sample of individuals with approved mortgages as people who have likely taken their debt capacity to the limit. Financial institutions make mortgage decisions with loan-to-value models. Thus, the dependent variable for the Table 3B estimation is

approved mortgage plus other outstanding debt relative to the value of the house. We put other debt in numerator since in Greece lenders can break up mortgages (like in the United States) into separate loans.[8] Using an approved mortgage sample is nice, in addition to providing a situation in which households are likely to be constrained, in that we are able to implement a different model (a collateral-based model) to see how robust our results are to a different model.

Table 3, panel B shows the mortgage results for the hard information model, the occupation fixed effects model and the wealth and growth soft information model. Although our sample of approved mortgages is less than five percent of the sample of consumer loan applications, we have a sufficiently large sample to identify all three models, albeit with less significant results. The largest tax evading occupations are Engineers and Agriculture. In money terms, Doctors are the largest offenders

Finally Table 3, panel C repeats the analysis in the population representative sample. The dependent variable is credit capacity. Doctors are again the largest tax evaders in money terms, followed by Dentists and Accountants. The three panels of Table 3 show that our estimates are robust to the various samples and specifications.

5 Validity and Inference

Table 4 summarizes our main results, with the first three columns reproducing the average multipliers λ across the columns for each of the panels of Table 3.[9] At the occupation level, we find the highest tax evasion multiple for lawyers, dentists and veterinarians, financial services, engineers, accountants and doctors. To assess the external validity of our results, we reconcile this cross section of offenders with a legislative bill that targeted eleven select occupations (doctors, dentists, veterinarians, lawyers, architects, engineers, topographer engineers, economists, business consultants, tax auditors and accountants). The bill recognized that these professions are the most likely to tax evade and taxpayers in these professions should be audited if they report income lower than a specified limit, set according to population criteria. The bill was rejected by the Greek Parliament before even voting. We observe from Table 4 that our estimates of the highest offenders coincide almost perfectly with the occupations targeted by the bill.

As a second test of validity, the sixth column of Table 4 presents the annual default probability, defined as the proportion of loans which go over 90 days deliquent per year from Table

[8] Note that we are careful to not double count loan applications due to mortgage dispersion.

[9] When a coefficient is insignificant, we force λ to be equal to one, indicative of no estimated tax evasion.

1. Although the individuals in tax-evading occupations have high credit outstanding relative to their declared income (from Table 1), their default rate is not higher than that of occupations with lower credit-to-income ratios.

To further validate our study, and to add perspective on incidence, we do a GIS mapping of incidence of tax evasion by zip code. Figure 2 shows that tax evasion is geographically very dispursed, which suggests that our estimates are not biased by an Athens effect and that we are able to reproduce an accepted "truth" that tax evasion is pervasive across Greece.

One interesting overlay is that in 2011, the *Financial Times* published a story about Larissa, a precinct in central Greece that is regarded to be the agricultural center of the country and heavily dependent on transfers and subsidies from the European Commission. This precinct was reported to have the highest density of Porsche Cayennes in Europe, and it overlays exactly to one of our high tax evasion districts.

The common understanding of tax evasion is that it is an upper income phenomenon. We can study the incidence of tax evasion by income level. Figure 2a shows the tax evaded income as a percentage of true income for different income buckets, while figure 2b shows the part of true income evaded and reported. As the figures show, it is indeed true that the rich hide larger fraction of their income from the tax authority and that the euro implications are large . However, the middle class are very active tax evaders as well, with average magnitudes of tax evading self-employed being in the twenties of thousands of euros per self-employed person.

6 Theories and Evidence: Making Sense of Industry Distribution

In this section, we turn to theories of tax evasion to make sense of the distribution across industries. We consider four such theories. Although we cannot offer causal evidence in any instance, our modest goal is to lay out how theory might approach explaining the distribution of industries and then to put forth evidence for consistency with the theories. Admittedly, we do not know whether the causes of the industry distribution in Greece would be the dominant ones in other countries, but this in no way hinders our being able to speak to the potential for different theories to matter.

First, if certain industries provide unskilled labor with apprentice-like training that increases the human capital of the economy, it may be that the subsidizing these industries is optimal. This theory resonates of Rosen (2005), who suggests that allowing tax evasion may be optimal if the supply of labor is more elastic in the underground economy, but our thoughts add an

educational aspect. To investigate this theory, we collect data from the United Kingdom on which occupations require apprenticeships. Table 5 shows a negative relation between the need for apprenticeship and the tax evasion multiplier for an occupation. Table 5 does show that the largest tax evaders are likely to be associated with higher education degree requirements. But since this education is already free in Greece and since Gong and van Soest (2002) show that wage differential between the formal and informal sectors increases with education, it is unlikely that the need to incentivize formal or informal education causes the distribution of tax evasion.

The second and third theories relate to enforcement. The majority of papers on tax evasion enforcement focus on the zero-one decision of business to be formal or informal rather than the degree to which all individuals tax evade, across occupations.[10] The incentives for enforcing and ability to enforce are prominent factors in decisions to evade, and it is possible these are correlated to occupational distribution.[11] One possibility for making sense of the industry distribution is that enforcement is related to geography. Maybe the tax authority is more lax in Athens (because they are more busy, or the payoff to lax enforcement is higher) or more stringent in Athens (because of career concerns or because in the countryside everybody is part of a community). If the distribution of industries in the cities is different than in the countryside, the industry pattern of tax evasion could be about the geography of enforcement. We can look again at Figure 1 to quickly cast doubt on this theory. Tax evasion happens all over Greece. We are gathering data on enforcement to investigate this theory more robustly.

The third theory builds on the ideas in Kleven, Knudsen, Kreiner, Pedersen and Saez (2011), who document (in Denmark data) that prior auditing (or the threat of future auditing) is more important than the size of the marginal tax rates in curbing tax evasion of self-reported income. Enforcement matters for extracting true information. Enforcement may be easier in occupations with nontracable information. Another way to say this is that tax evasion is simply easier in some occupations, in particular, in occupations with no paper trail. We classify industries by whether or not the main function of the occupations reflects a paper trail. Table 5 shows that indeed those industries with paper trail are much less likely to be tax evaders.

Finally, the fourth theory reflects back to the legislative bill that did not pass, which would

[10] Some notable exceptions consider the effect of tax evasion on occupation choice following Parker (1999) and Pestieau and Possen (1991).

[11] Working against the enforcement argument is Antunes and Cavalcanti (2007), who show that it is regulation costs and not the level of enforcement, that accounts for differences in tax evasion between the United States and Mediterranean Europe. We are getting more scientific data on the extent to a paper trail industries have for the next draft, but our assesment seems reasonable.

have mandated a minimum tax for certain industries. Could the distribution of tax evading industries be related to politicians looking out for their own professions? We would like to have seen the distribution of votes on that bill, but we can do a second best. We look up the parliamentarians occupational backgrounds to see whether they correlate with tax evaders. Figure 3 presents these results. Although there are some notable exceptions (journalist are very prominent in Parliament and reside in *professional other*, which is a medium-low tax evader occupation), the correlation is nevertheless positive.

7 Discussion

Our results suggest that 41 billion euros of taxable income go unreported. With a tax rate of 40% in Greece, up to 16.3 billion euros of additional tax revenue could be collected. This represents an amount equal to 68 percent of the deficit for 2010 (or 45% for 2009). However, collecting said revenue is a non-trivial task. In Greece, as in many places, the tax authorities cannot get access to individual banking records to implement a tax collection based on our direct strategy. Even if they could, the equilibrium of using the banking system would immediately change to preclude the information value to the tax authorities of recreating our estimates in real time.

One idea would be to apply a *presumed income tax* method in which the government mandates a minimum tax reporting by occupation specific to each of the prefectures using statistics from our estimation. Say, for example, the government would be willing to shift the burden of proof (of making an income lower than the presumption) onto 25% of self-employed individuals, then the tax authority could impose the presumed income based on our estimates of the 25th percentile of true income by occupation and *by prefecture*.

Implementing presumed income is burdensome and perhaps politically unsavory. We believe a better approach is an annual *business license fee*. City authorities world-wide use business licenses to collect revenues. The attractive feature of business licenses is that an implementation mechanism for the most egregious tax evaders is already established, mandatory industry associations (e.g., in the U.S. context, the legal Bar and American Medical Association). In occupation rankings in Table 5, the industries with high tax evasion (lawyers, dentists, accountants, financial services, doctors and engineers) are also industries requiring ongoing participation in the industry association. These industries also have a high pervasiveness of tax evasion, thus eliminating a lot of the burden of proving non-tax evasion.[12]

[12] The implementation of business licenses fees also may serve to reduce uncertainty regarding the taxation

For most of the other industries, associations exist, but participation is less mandatory. These are, coincidentially, industries with a greater distribution of both income and whether or not tax evasion happens. One might imagine a hybrid mechanism to collect taxes using a presumed income model which plus site occupation fees for storefront based on commercial property value and business licence fees aimed at the higher-income individuals who relying on associations for certification.

The second topic that we relegated to discussion concerns the welfare implications to the bank adaptation. Although we are limited in space and data to accomplish a study of welfare, we would like to introduce ideas since the observation of bank adaptation to semiformality is completely new, to our knowledge. In a Coasian or De Soto view of the world, the fact that banks give an entitlement to informal income provides a property right that allows individuals to use borrowing to optimally smooth lifetime consumption or overcome shocks. Two perspectives emerge from this statement. On one side, the banks' actions decrease the penalties for individuals to tax evade. We do not presume to know the elasticity of tax evasion to access to credit. However, it is likely that the societal cost to bank adaptation is offset by the benefit of credit access, given the cultural and political norm of tax evasion. In our view, it would thus be interesting to know the extent to the haircut imposed on tax evaded income and to speak to the welfare benefit accruing to borrowers. In our opinion, the idea that income is a property right with entitlement benefits like a farm deed is new and interesting. Likewise, extending the analysis of adaptation to commercial borrowing would seem worthwhile.

8 Conclusion

Using individual-level household lending data, we develop a new methodology to estimate tax evasion based on the household's ability to borrow and financial institutions' need to adapt their lending to the culture of semiformality. We quantify the extend of tax noncompliance overall in Greece; our best estimates is that unreported taxable income is 41 billion euros. Foregone tax revenues are 16.3 billion euros, or 68 percent of the deficit for 2010 (or 45% for 2009)

system. The Greek population is very aware of the inevitability of reform to the tax system. Because theories as to how reform would be accomplished (and maintained) vary in how they affect occupations and returns to self-employment, reform uncertainty may discourage entrepreneurship and investment and reduce the competitiveness of the economy. The existence of a (politically remote) fixed and ex-ante known fee that reflects the tax obligation of the self-employed may help to remove distortions.

Beyond real-time applications to Greece, the goals of the paper were twofold: to establish a new methodology to estimate tax evasion using private data and to offer insight into the distribution of industries in an economy in which tax evasion is pervasively the norm.Tax evasion happens in industries with low paper trail and in those well representation in Parliament. Although we do not develop further tests in this paper, the lining up of our results with paper trail suggest that those countries most impacted by tax evasion (often less developed countries) might focus on systems that generate paper trail to provide a source of income information of the form of Klevin et al (2011).

We cannot establish the underpinning reason why politicians lack willpower to force doctors, lawyers, and other similar professions to pay taxes. It may be that true income is hard to prove or that these industries carry political power. It may also be, however, that these occupations reflect some other underlying trait (education attainment) that the encourages the society to accept the large-scale tax evasion. Whatever the reason, clearly politicians lacked willpower in early 2011 in their failure to pass the minimum tax bill for exactly these professions. However, the dismal prognosis for the economy in 2012 suggests that the incidence of tax evasion may be a more poignant issue now than ever.

9 References

Antunes Antonio R. and Cavalcanti Tiago V. de V. "Start up Costs, Limited Enforcement, and the Hidden Economy". European Economic Review, 2007, 51 (1), pp. 203-224.

Andreoni, James; Erard, Brian and Feinstein, Jonathan. "Tax Compliance." Journal of Economic Literature, 1998, XXXVI(June 1998), pp. 818-60.

Cagan, Phillip. "The Demand for Currency Relative to the Total Money Supply." Journal of Political Economy, 1958, 66(4), pp. 303-28.

Chetty, Raj and Saez, Emmanuel. "Optimal Taxation and Social Insurance with Endogenous Private Insurance." American Economic Journal: Economic Policy, 2010, 2(2), pp. 85-114.

Feldman, Naomi E. and Slemrod, Joel. "Estimating Tax Noncompliance with Evidence from Unaudited Tax Returns." Economic Journal, 2007, 117, pp. 327-52.

Feld, Lars P. and Schneider Friedrich. â€œSurvey on the Shadow Economy and Undeclared Earnings in OECD Countries.â€İ German Economic Review, 2010, Vol. 11 (2), pp. 109-149.

Feldman, Naomi E., and Joel Slemrod. 2007. "Estimating Tax Noncompliance with Evidence from Unaudited Tax Returns." Economic Journal, 117(51), pp. 327-352.

Gong Xiadong and van Soest Arthur. "Wage Differentials and Mobility in the Urban Labour Market: a Panel data analysis for Mexico." Labour Economics, 2002, 9, pp. 513-529.

Harberger, Arnold. "Reflections on Corruption: An Economist's Perspective," Strategic Objective or Activity: 905-101/Economic Growth. U.S. Agency for International Development, 2006.

Hurst Erik, Li Geng and Pugsley Benjamin. "Are Households Surveys Like Tax Forms: Evidence from Income Underreporting of the Self Employed", Finance and Economics Discussion Series, Board of Governonrs of the Federal Reserve System, 2011.

Kleven Henrik J., Knudsen Martin B., Kreiner Claus T., Pedersen Soren and Saez Emmanuel. "Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark." Econometrica, 2011, 79, pp. 651-692.

Lacko, Maria. "Hidden Economy - an Unknown Quantity? Comparative Analysis of Hidden Economies in Transition Countries, 1989-95." Economics of Transition, 2000, 8(1), pp. 117-49.

Pestieau, Pierre and Uri M. Possen. "Tax Evasion and Occupational Choice." Journal of Public Economics, 1991, 45, pp. 307-325.

Parker, Simon C. "The Optimal Linear Taxation of Employment and Self- employment Incomes." Journal of Public Economics, 1999, 73, 107-123.

Pissarides, Christopher A. and Weber, Guglielmo. "An Expenditure-Based Estimate of Britian's Black Economy." Journal of Public Economics, 1989, 39(1), pp. 17-32.

Rosen Hurvey S. (2005)

Schneider, Friedrich and Enste, Dominik H. "Shadow Economies: Size, Causes and Consequences." Journal of Economic Literature, 2000, 38(1), pp. 77-114.

Slemrod, Joel and Yitzhaki, Shlomo. "The Costs of Taxation and the Marginal Efficiency Cost of Funds." IMF Staff Papers, 1996, 43(1), pp. 172-98.

Tanzi V. "The Underground Economy in the United States: Annual Estimates 1930- 1980". International Monetary Fund Stuff Papers, 30, pp. 283-305.

Appendix A1: Variable Definition

Credit Capacity:	Combination of debt outstanding of the applicant, plus the credit capacity allowed on the applied-for loan.
Total Loans-to-Commercial Value:	Ratio of the approved mortgage plus other outstanding debt relative to the value of the real estate property. This variable is only available for mortgages and is calculated only for approved applications.
Declaredl Income:	Income of the applicant as shown in the tax return form in the year of application.
Age:	Age of the applicant.
Real Estate Value, Mean Zip Level:	The mean "presumed" real estate value of the region where the primary applicant resides (specified by zip code). "Presumed" real estate values are periodically published by tax authorities and used to determine real estate taxes. The values here come from the most recent release published in 2007.
Car Value, Mean Zip Level:	The mean value of automobiles by zip code, as calculated from the car loan dataset of the bank.
Car Value, Mean by Occupation:	The mean value of automobiles by occupation and employment status, as calculated from the car loan dataset of the bank.
Tax Authority Income Growth:	Income growth by zipcode, income quintile and employment type level the year before the application, as calculated by the tax authority data based on the tax returns of all Greek taxpayers.
Standard Dev. of Income Growth	To construct a measure of the standard deviation of the growth of income, we take into account the difference in the number of people in the zip code-income decile-occupation. The measure is defined as the standard deviation divided by the square root of the observation count.
Years in Job:	Number of years that the applicant is working under the same employer (wage workers) or is employed in the same occupation (self-employed).
Years in Address:	Number of years that the primary applicant resides in the same address.
Status with Bank:	The variable takes the value 0 if the applicant has no existing relationship with the bank, 1 if the applicant is already a customer of the bank and 2 if he applicant is a customer of the bank but has a delayed payment in the last 6 months.
Years in Corporation with the Bank:	Number of years of the existing relationship of the applicant with the bank.
Deposits:	Total amount of deposits of the applicant in the bank. It includes amounts in any accounts the applicant holds with the bank (checking, savings, time-deposits, investment accounts).
Homeownership:	The variable takes the value 1 if the applicant is homeowner and 0 otherwise.
Marital Status:	The variable takes the value 1if the applicant is married and 0 otherwise.

Number of Children: The number of children of the applicant at the time of the application.

Appendix A2: Credit History Construction

The credit history and credit performance variables were created from a set of three products: mortgages, term loans, and credit card debt. Each product included, for each customer for each month, a customer id, a record of the current month, a record of the current delay, and the amount owed on the loan. For each product a variable was generated recording how many times a customer had been late on their loan payment within the past six months, within the past 12 months, and within the past 24 months. For each span of months, we created a variable scaling the months in delay by the amount owed over the number of months in each interval. For example, this variable for the six month interval would be calculated,

six month delay amount = (six month delay* amount owed)/six

From these individual products, an average was created for the month in delay and the month in delay amount variables. The average was calculated over all the products for each customer for each month.

Besides the month in delay variables, we generated a dummy variable noting whether a customer had been delayed for a period between one and three months, for a period between three and five months, or for a period greater than six months. Additionally, a variable was created noting if a customer had ever been late on their loan, if they had ever been over a month late on their loan, or if they had every defaulted on their loan - a customer was considered to have gone into default when they were at least three months late on their loan. We then generated a second version of this variable for the entirety of a customer's products for each month in a similar fashion to the average generated above. Finally, we created a variable noting the earliest month in which a customer was late on their loan. A second version of this variable was also generated over the entirety of each customer's accounts for each month.

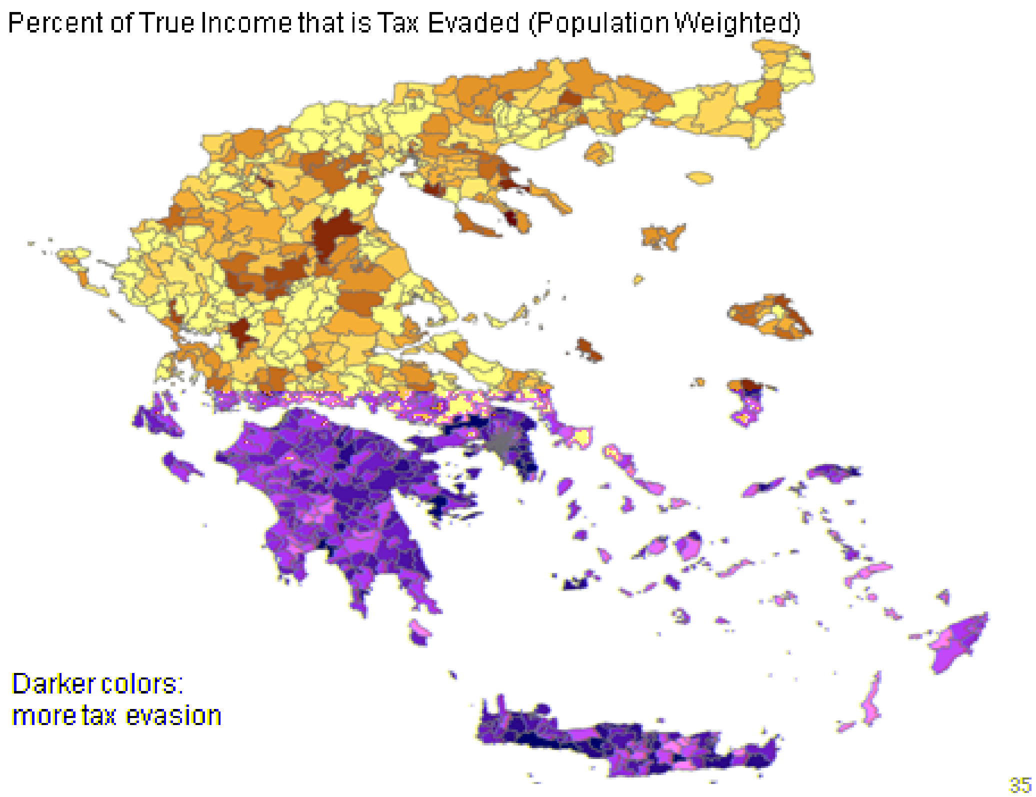

Figure 1: Geographical distribution of tax evasion

The figure presents the geographical distribution of tax evasion by zipcode.

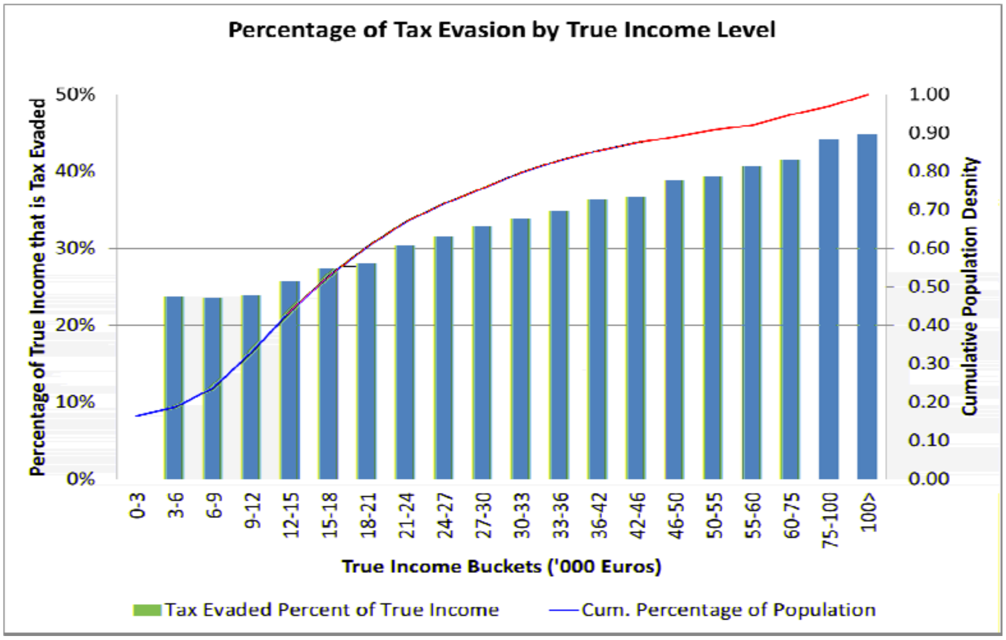

Figure 2a: Percentage of Tax Evasion by True Income Level

The figure shows the tax evaded income as percentage of the true income for buckets in true income and the cumulative population density of the true income buckets is plotted. All amounts are in thousands of euros.

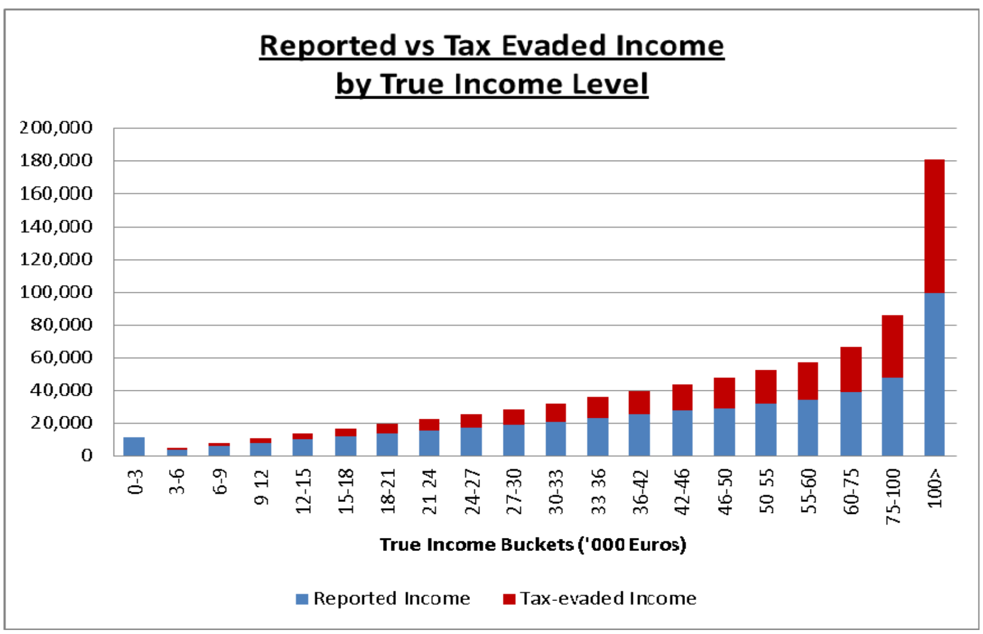

Figure 2b: Reported vs Tax Evaded Income

The figure plots the part of true income that is reported and the part that is tax evaded within true income buckets. All amounts are in thousands of euros.

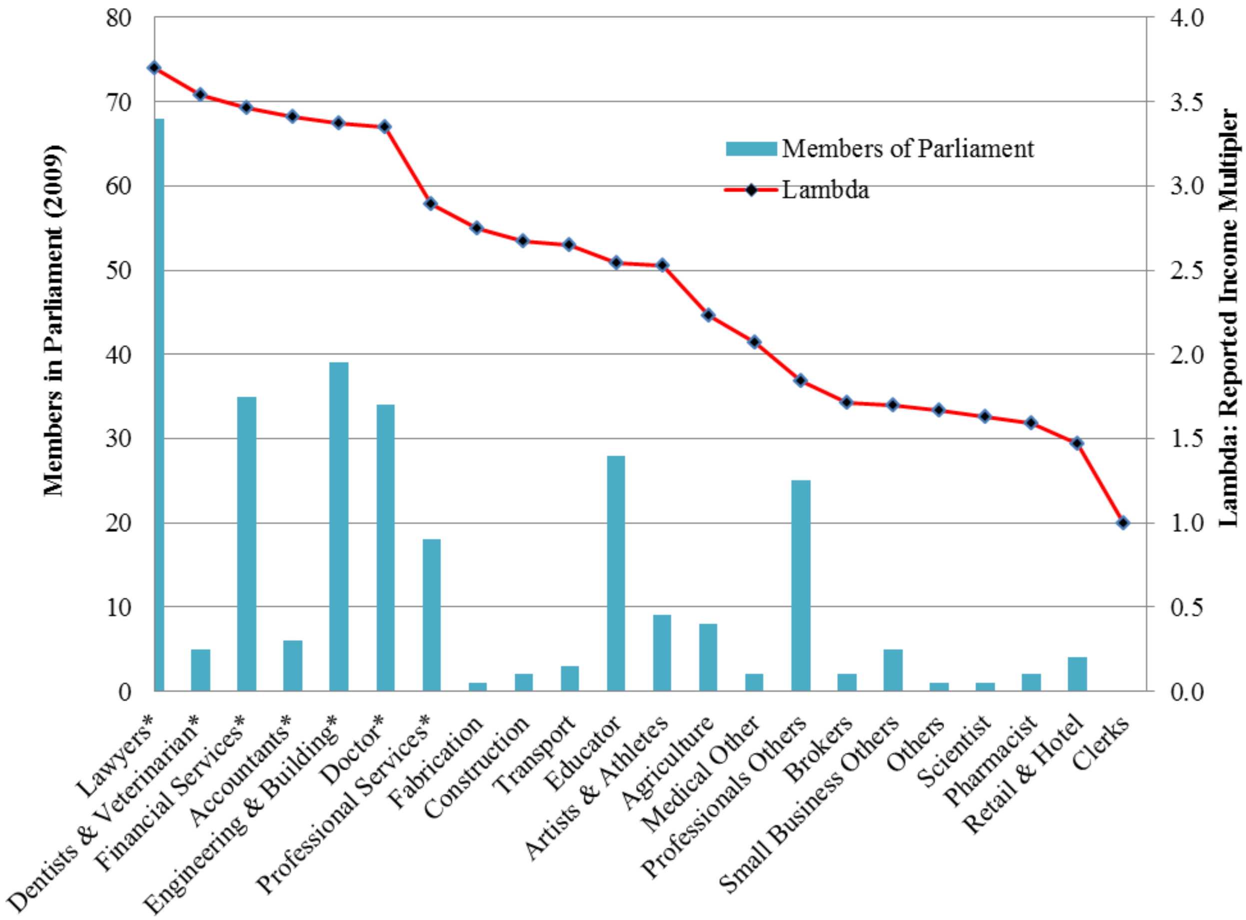

Figure 3: Tax Evasion Multipliers and Politician Distribution

The figure presents the number of members of the Greek Parliament in each occupation class (left vertical axis) along with the average lambdas of these occupations (right vertical axis). The Greek Parliaments consists of 300 members and the data refers to the current term (members elected in 2009). The average lambdas are the means of the reported income multiplier of the three samples (constrained, mortgages, population representative) across all specifications.

Table 1: Monthly Debt Servicing Motivation for Study

The bank data are from a large Greek bank, with occupation, income distribution, and geography weighted to be representative to the population of Greece at large. Data are from 2003-2011. The sample for this table are mortgage applicants and non-homeowner consumer loan applicants, which excludes homeowner consumer loan applications because we do not have a good estimate of monthly payments on other debt outstanding. Monthly declared income is the verified income reported an individual reports to the tax authority. Monthly debt payments is the servicing on the loans, with the interest rate of 10% for consumer loans and the rate applicable for mortgages (conservative averages). The annual default rate is an indicator for each loan by year as to whether the loan goes over 90 days delinquent. The monthly tax evasion implied is the amount of income individuals would have to have, given the debt payments, under the rule that debt servicing cannot be more than 30% of monthly income.

Occupation	Monthly Declared Income	Monthly Debt Payments	Ratio of Payment/ Income	Annual Default Probability	Annual Tax Evasion Implied
Accountants & Notaries	1,239	1,525	1.23	0.06	47,784
Agriculture	868	404	0.47	0.24	9,888
Artists & Athletes	1,125	805	0.72	0.14	21,924
Brokers	1,464	1,466	1.00	0.15	43,212
Construction	974	633	0.65	0.18	17,292
Dentists & Veterinarian	1,188	1,524	1.28	0.05	48,288
Doctor	2,189	2,063	0.94	0.03	60,912
Educator	983	1,088	1.11	0.15	33,804
Engineering & Building	1,320	1,173	0.89	0.07	34,632
Fabrication	1,283	1,265	0.99	0.14	37,716
Financial Services	1,449	1,768	1.22	0.15	55,416
Lawyers	1,402	1,499	1.07	0.06	44,736
Medical Other	1,255	836	0.67	0.07	21,864
Others	1,173	913	0.78	0.20	25,932
Pharmacist	1,852	2,234	1.21	0.03	71,796
Professional Others	1,225	1,081	0.88	0.12	31,392
Professional Services	1,505	1,169	0.78	0.17	32,568
Retail & Hotel	1,231	1,361	1.11	0.20	42,636
Scientist	1,422	891	0.63	0.05	24,252
Small Business Others	1,249	915	0.73	0.19	26,004
Transport	1,056	869	0.82	0.11	23,604
Total	***1,127***	***900***	***0.80***	***0.18***	***25,944***

Table 2: Summary Statistics

The table provides mean and median values for the dependent variables and mean value for hard and soft information, by employment status. The sample in (1) is the constrained sample, defined as applicants whose requested loan amount is greater than the approved, overdraft applicants with less than 1,000 euros on deposit, and refinancing applicants. The sample in (2) is mortgage applicants. The sample in (3) is the population representative sample, defined as credit card applicants, overdraft provision applicants who do not need the facility, credit line applicants, and appliance loan applicant. The motivation for the samples is described in the data section. The dependent variable for panels A and C are credit capacity. In the mortgage sample, we use total loans-to-value, as the bank decision variable rather than mortgage loan to value because of the process of dividing up home loans into multiple pieces. We use only for approved mortgages and there is no information about the years of cooperation with the bank. The definitions of the hard and soft information variables are included in Appendix A1. The samples are weighted to the population using the tax authority data.

	1:Constrained Sample			**2:Mortgage Sample**			**3:Population Representative Sample**		
Variables	**Wage Workers**	**Self employed**	**Total**	**Wage Workers**	**Self employed**	**Total**	**Wage Workers**	**Self employed**	**Total**
A. Dependent Variables									
Median Credit Capacity (€)	9,700	16,200	10,000	--	--	--	2,000	2,500	2,000
Mean Credit Capacity (€)	30,308	59,468	37,329	--	--	--	9,083	12,759	9,778
Median Total Loans-to-Commercial Value	--	--	--	0.80	0.85	0.80	--	--	--
Mean Total Loans-to-Commercial Value	--	--	--	0.94	1.16	0.97	--	--	--
B. Permanent Income									
Reported Income (€)	18,401	18,174	18,346	20,315	13,599	19,475	13,440	15,015	13,738
Age	44.74	47.52	45.41	43.44	46.58	43.84	40.82	44.67	41.55
Real Estate Value, Mean Zip Level (€)	1,167	1,102	1,151	1,175	1,090	1,164	1,147	1,048	1,128
Car Value, Mean Zip Level (€)	16,781	16,919	16,814	16,669	16,954	16,705	16,626	16,754	16,650
Car Value, Mean by Occupation (€)	16,220	19,058	16,903	16,728	19,556	17,082	16,111	19,122	16,681
Tax Authority Income Growth	0.008	0.003	0.007	0.010	0.008	0.009	0.008	0.001	0.007
Standard Dev. Of Income Growth	0.30	0.26	0.29	0.29	0.51	0.32	0.32	0.27	0.31
C. Worthiness									
Years in Job	11.20	13.48	11.75	10.28	13.67	10.71	7.81	11.20	8.45
Years in Address	14.17	17.21	14.90	10.73	14.04	11.15	13.50	15.36	13.85
Status with Bank (New customer/ existing customer/existing cust. in bad standing)	0.74/0.22/ 0.04	0.74/0.23 /0.04	0.74/0.22 /0.04	--	--	--	0.29/0.67 /0.04	0.24/0.72 /0.04	0.28/0.68 /0.04
Years of cooperation with bank	7.19	8.11	7.41	--	--	--	2.19	2.96	2.34
Deposits (€)	1,456	2,558	1,745	10,402	19,539	11,718	3,203	5,886	3,718
Homeownership	0.59	0.67	0.61	--	--	--	0.55	0.69	0.57
D. Socioeconomic Characteristics									
Married	0.62	0.69	0.64	0.74	0.73	0.74	0.48	0.60	0.50
Number of Children	0.49	0.57	0.51	0.75	0.81	0.76	0.47	0.61	0.50

Table 3: Bank Model Results

The three panels present the bank model results for the constrained, mortgage, and population representative samples. The dependent variables are credit capacity in panels (a) and (c) and total loans-to-value in the mortgage panel (b). Presented are the coefficients on income interacted with a wage worker dummy and income for the self-employed, specific for each occupation. Robust standard errors are omitted due to space. Lambdas are the ratio of the coefficient on income for each of the self-employed occupations divided by the coefficient for the wage worker. The details of the lambda1, lambda2 and lambda3 are in the text, representing the bounds analysis letting the soft information of wealth and growth affect adaptation as well as repayment adjustments made by local bank officers.] Not shown in each estimation are a self-employment dummy, age, and the socioeconomic variables and credit worthiness variables in Table 2.The occupation fixed effects models include occupation fixed effects and occupation crossed with self-employment fixed effects. [The description continues in Panel B.]

Panel A	**Hard Information Model**		**Occupation Fixed Effects**		**Wealth & Growth Soft Information Model**				**Saturated**		**Inference**	
Constrained Sample Dep Var: Credit Capacity	**Coefficient**	**Lambda**	**Coefficient**	**Lambda**	**Coefficient**	**Lambda1**	**Lambda2**	**Lambda3**	**Coefficient**	**Lambda**	**Reported Income**	**True Income**
Income, Wage Workers	0.438***		0.409***		0.383***				0.533***			
Income, Self Employed:												
Accountants & Notaries	1.686***	3.9	2.212***	5.4	2.162***	5.6	5.7	5.7	2.625***	4.9	20,854	103,389
Agriculture	0.150**	0.3	0.183**	0.4	0.181**	0.5	0.3	0.5	0.282	--	13,286	13,286
Artists & Athletes	1.171***	2.7	1.633***	4.0	1.587***	4.1	4.4	4.2	1.647***	3.1	15,315	53,230
Brokers	1.187***	2.7	0.941***	2.3	0.919***	2.4	2.5	2.4	1.508***	2.8	22,556	57,755
Construction	0.686***	1.6	0.753***	1.8	0.737***	1.9	2.0	1.9	1.151***	2.2	16,986	31,812
Dentists & Veterinarian	1.755***	4.0	1.930***	4.7	1.893**	4.9	5.2	4.9	1.721**	3.2	19,868	83,962
Doctor	1.663***	3.8	1.572***	3.8	1.505***	3.9	4.1	3.9	0.978*	1.8	33,133	111,079
Educator	1.377***	3.1	1.310*	3.2	1.253	3.3	3.3	3.3	1.246*	2.3	16,645	40,292
Engineering & Building	1.147***	2.6	1.091***	2.7	1.062***	2.8	3.0	2.8	0.686**	1.3	24,166	56,494
Fabrication	1.186***	2.7	1.382***	3.4	1.352***	3.5	3.6	3.5	1.649***	3.1	17,226	54,750
Financial Services	1.541***	3.5	1.151*	2.8	1.114*	2.9	3.0	2.9	0.641	--	22,534	57,728
Lawyers	2.556***	5.8	3.422***	8.4	3.365***	8.8	9.1	8.8	3.120***	5.8	20,442	147,442
Medical Other	2.592***	5.9	1.730	--	1.678	--	--	--	1.785	--	20,103	44,832
Others	0.829***	1.9	0.927***	2.3	0.898***	2.3	2.3	2.4	1.119***	2.1	16,713	35,954
Pharmacist	1.182***	2.7	0.330	--	0.336	--	--	--	0.138	--	50,365	71,770
Professional Others	0.893***	2.0	1.101***	2.7	1.069***	2.8	3.3	2.8	0.331	--	20,776	44,282
Professional Services	1.845***	4.2	1.728***	4.2	1.634***	4.3	4.4	4.3	1.771***	3.3	22,999	92,184
Retail & Hotel	1.010***	2.3	0.821***	2.0	0.764***	2.0	2.1	2.0	0.588***	1.1	19,336	35,843
Scientist	0.145	--	0.527	--	0.432	--	--	--	0.880	--	19,092	19,092
Small Business Others	0.961***	2.2	0.837***	2.0	0.790***	2.1	2.2	2.1	0.822***	1.5	20,121	39,479
Transport	0.428***	1.0	0.421***	1.0	0.388***	1.0	1.0	1.0	0.603**	1.1	18,638	19,347
R-squared	*0.322*		*0.326*		*0.319*				*0.324*			

Table 3, Panel B

[continued from panel A] The wealth and growth soft information model includes the occupation fixed effects as well as car value means by zip code, car value means by occupation, tax authority income growth, and the standard deviation of the income growth. The inference columns report the original reported self-employed income and the inference for the true income, based on the average lambdas for the panel, where a lambda of 1 (no inference) is applied for not significant results. ***, **, and * indicate significance at the 1%, 5%, and 10% levels respectively.

	Hard Information Model		Occupation Fixed Effects		Growth Soft Information Model			Inference	
Mortgage Sample Total Loans-to-Value	**Coefficient**	**Lambda**	**Coefficient**	**Lambda**	**Coefficient**	**Lamba1**	**Lamda2**	**Reported Income**	**True Income**
Income, Wage Workers	1.944***		1.788***		1.760***				
Income, Self Employed:									
Accountants & Notaries	1.483	--	0.447	--	0.368	--	--	16,805	16,805
Agriculture	0.321	--	11.71***	6.5	11.53***	6.5	6.6	12,687	59,617
Artists & Athletes	-1.267	--	10.81	--	10.78	--	--	13,474	13,474
Brokers	11.13	--	8.286	--	10.24	--	--	16,051	16,051
Construction	6.655***	3.4	8.760***	4.9	8.767***	5.0	5.1	10,342	45,867
Dentists & Veterinarian	1.106	--	0.752	--	0.423	--	--	12,876	12,876
Doctor	5.299***	2.7	5.241***	2.9	5.174***	2.9	3.0	28,279	81,036
Educator	2.039	--	3.300	--	3.566	--	--	7,034	7,148
Engineering & Building	7.917**	4.1	9.309**	5.2	9.268**	5.3	5.3	18,924	91,752
Fabrication	5.811***	3.0	5.445**	3.0	5.396**	3.1	3.1	13,423	40,724
Financial Services	14.23*	7.3	9.570	--	9.578	--	--	19,588	60,853
Lawyers	-0.532	--	-1.225	--	-1.351	--	--	16,058	16,058
Medical Other	7.087*	3.6	6.451	--	6.483	--	--	12,871	24,222
Others	3.687**	1.9	2.141*	1.2	2.107*	1.2	1.5	11,834	16,928
Pharmacist	4.379	--	0.289	--	0.007	--	--	14,863	14,863
Professional Others	5.365*	2.8	6.952	--	6.860	--	--	15,140	24,023
Professional Services	1.475	--	-0.149	--	-0.304	--	--	15,980	15,980
Retail & Hotel	3.089**	1.6	2.420*	1.4	2.263*	1.3	1.4	11,765	16,583
Scientist	-0.338	--	1.441	--	1.364	--	--	26,537	26,537
Small Business Others	3.120***	1.6	2.998***	1.7	2.922***	1.7	1.7	15,373	25,327
Transport	5.736**	3.0	7.833***	4.4	8.447***	4.8	4.6	11,454	46,321
R-squared	*0.086*		*0.088*		*0.089*				

Table 3, Panel C

	Hard Information Model		Occupation Fixed Effects		Wealth & Growth Soft Information Model				Inference	
Population Representative Sample Dep. Var.: Credit Capacity	**Coefficient**	**Lambda**	**Coefficient**	**Lambda**	**Coefficient**	**Lambda1**	**Lambda2**	**Lambda3**	**Reported Income**	**True Income**
Income, Wage Workers	0.158***		0.141***		0.140***					
Income, Self Employed:										
Accountants & Notaries	0.669***	4.2	0.607***	4.3	0.596***	4.3	4.4	4.3	19,520	83,225
Agriculture	-0.0618***	-0.4	0.102***	0.7	0.106***	0.8	0.6	0.8	10,580	3,814
Artists & Athletes	0.389***	2.5	0.487***	3.4	0.477***	3.4	3.5	3.4	14,726	45,718
Brokers	0.276***	1.8	0.210**	1.5	0.204**	1.5	1.6	1.5	21,123	33,013
Construction	0.203***	1.3	0.268***	1.9	0.264***	1.9	1.9	1.9	13,414	22,653
Dentists & Veterinarian	0.673***	4.3	0.849***	6.0	0.830***	5.9	6.2	5.9	15,785	85,215
Doctor	0.604***	3.8	0.543***	3.8	0.531***	3.8	4.1	3.8	34,525	131,880
Educator	0.576***	3.7	0.647**	4.6	0.614**	4.4	4.4	4.4	15,451	64,935
Engineering & Building	0.451***	2.9	0.424***	3.0	0.414***	3.0	3.2	3.0	18,991	55,798
Fabrication	0.316***	2.0	0.291***	2.1	0.283***	2.0	2.1	2.0	16,493	33,418
Financial Services	0.688***	4.4	0.697***	4.9	0.683***	4.9	5.1	4.9	19,699	93,034
Lawyers	0.503***	3.2	0.390**	2.8	0.378**	2.7	3.1	2.7	20,091	57,852
Medical Other	0.361***	2.3	0.292***	2.1	0.268***	1.9	1.9	1.9	17,356	36,252
Others	0.220***	1.4	0.209***	1.5	0.202***	1.4	1.4	1.4	15,260	21,953
Pharmacist	0.419***	2.7	0.313***	2.2	0.308***	2.2	2.4	2.2	41,078	96,728
Professional Others	0.300***	1.9	0.250***	1.8	0.242***	1.7	1.9	1.7	19,001	34,172
Professional Services	0.557***	3.5	0.535***	3.8	0.519***	3.7	3.9	3.7	18,140	66,632
Retail & Hotel	0.216***	1.4	0.147***	1.0	0.140***	1.0	1.1	0.8	17,254	19,627
Scientist	0.345**	2.2	0.463**	3.3	0.455**	3.2	3.4	3.3	14,491	42,075
Small Business Others	0.235***	1.5	0.217***	1.5	0.209***	1.5	1.6	1.5	15,564	23,424
Transport	0.352***	2.2	0.452***	3.2	0.448***	3.2	3.3	3.2	15,588	44,845
R-squared	*0.128*		*0.133*		*0.134*					

Table 4 – Average Lambdas and Validity

The table presents the average lambdas across all specifications for the constrained, mortgage and population representative samples. The fifth column shows the average lambda of the three samples, which is the criterion of the ranking. The "Legislation Bill" column indicates the occupations that were targeted by a legislation bill proposed by the Greek government in January 2011. The bill targeted 11 occupations as likely to tax-evade and provided for tax audits for the professionals in these classes that reported less income than prespecified limits, set according to population criteria. These occupations were: Doctors, Dentists, Veterinarians, Accountants, Tax Auditors, Lawyers, Architects, Engineers, Topographers, Economists and Business Consultants. The sixth column presents the annual default probability by occupation, defined as the proportion of loans which go over 90 days delinquent by year. The last column shows the mean tax evasion in euros as calculated from the average lambda of the three samples across all specifications.

Occupation	Constrained Sample	Mortgage Sample	Population Rep. Sample	Average Lambda	Legislation Bill	Annual Default Probability	Mean Tax Evasion
Lawyers	7.2	1.0	2.9	3.7	YES	0.06	54,920
Dentists & Veterinarian	4.2	1.0	5.4	3.5	YES	0.05	44,508
Financial Services	2.6	3.1	4.7	3.5	YES	0.15	49,931
Accountants & Notaries	5.0	1.0	4.3	3.4	YES	0.06	48,747
Engineering & Building	2.3	4.8	2.9	3.4	YES	0.07	47,321
Doctor	3.4	2.9	3.8	3.3	YES	0.03	76,019
Professional Services	4.0	1.0	3.7	2.9	MIXED	0.17	39,226
Fabrication	3.2	3.0	2.0	2.7		0.14	27,250
Construction	1.9	4.4	1.7	2.7		0.18	19,863
Transport	1.0	4.0	2.9	2.7		0.11	21,611
Educator	2.4	1.0	4.2	2.5		0.15	24,377
Artists & Athletes	3.5	1.0	3.1	2.5		0.14	22,969
Agriculture	1.0	4.7	1.0	2.2		0.24	15,643
Medical Other	2.2	1.9	2.1	2.1		0.07	18,325
Professional Others	2.1	1.6	1.8	1.8		0.12	15,854
Brokers	2.6	1.0	1.6	1.7		0.15	15,696
Small Business Others	2.0	1.6	1.5	1.7		0.19	12,391
Others	2.2	1.4	1.4	1.7		0.20	10,342
Scientist	1.0	1.0	2.9	1.6		0.05	9,195
Pharmacist	1.4	1.0	2.4	1.6		0.03	25,685
Retail & Hotel	1.9	1.4	1.1	1.5		0.20	7,899

Table 5 – Estimated Tax Evasion, Apprenticeship and Paper Trail

The Table presents the occupation classes ranked on their average lambdas along with information regarding their apprenticeship intensity, the source of their professional rights and the degree of paper trail. The average lambdas are the mean lambdas of the three samples (constrained, mortgages and population representative samples) across all specification. The column "Degree Requirement" indicates if a degree is required to follow the occupation. The column "Apprenticeship" shows whether there is a mandatory (YES-M) or optional (YES-O) requirement for a new entrant to work as an apprentice. The fifth column specifies the required length of apprenticeship (mandatory case) or the average length (optional case). The sixth column shows whether the job permit is issued by the union or the government. The column "Paper Trail" indicated the amount of paper trail for each profession. The last column presents the percentage of self-employed taxpayers that tax evade at least 1000 euros annually, based on their average lambda. The percentage is calculated for the self-employed that hold both mortgages and terms loans and refers to the period from 2006 to 2010.

Occupation	Average Lambda	Degree Requirement	Apprenticeship	Length of Apprenticeship	Job Permit Issued by	Paper Trail	Percentage of Tax Evaders
Lawyers	3.7	YES	YES - M	18 months	Union		83.9%
Dentists & Veterinarian	3.5	YES			Union		80.2%
Financial Services	3.5					Some	71.2%
Accountants & Notaries	3.4	YES			Union	Mixed	80.8%
Engineering & Building	3.4	YES			Union	Some	62.7%
Doctor	3.3	YES	YES - M	12 months	Union		76.3%
Professional Services	2.9	MIXED			Union/Gov.		55.7%
Fabrication	2.7					High	68.7%
Construction	2.7		YES - O	24-48 years	Government		44.6%
Transport	2.7				Government	High	68.0%
Educator	2.5	YES					70.4%
Artists & Athletes	2.5						51.3%
Agriculture	2.2		YES - O			High	35.6%
Medical Other	2.1	YES	YES - M	12 months	Government		60.3%
Professional Others	1.8	MIXED	YES - O	12-24 months		Mixed	62.3%
Brokers	1.7		YES - O	12-24 months	Government	High	74.9%
Small Business Others	1.7						51.2%
Others	1.7	MIXED	YES - O	24-48 years			54.9%
Scientist	1.6	YES					57.3%
Pharmacist	1.6	YES	YES - M	12 months		High	85.0%
Retail & Hotel	1.5		YES - O	12-24 months		Some	57.3%