Escuela de Gobierno y Transformación Pública
Tecnológico de Monterrey

A discrete choice model for labor informality in Mexico using restricted sets

Héctor Juan Villarreal Páez
hjvp@tec.mx
School of Government and Public Transformation
Tecnológico de Monterrey

Diego Vázquez-Pimentel
diego.alejo.vazquez@gmail.com
School of Government and Public Transformation
Tecnológico de Monterrey

School of Government and Public Transformation

Working Paper No. 5

Publication date: july, 2025

A discrete choice model for labor informality in Mexico using restricted sets

July 10, 2025

Diego Vázquez-Pimentel, diego.alejo.vazquez@gmail.com
Héctor J. Villarreal
Escuela de Gobierno y Transformación Pública, Tec de Monterrey

Abstract

This paper explores the dynamics of labor informality in Mexico by developing a discrete choice mixed logit model to explain the transitions between labor states—namely, not employed, formal employment, and informal employment—among individuals aged 18 to 65. The study offers critical insights into the informal sector's heterogeneity, with particular focus on voluntary versus involuntary informality, while also contributing a novel estimation strategy that combines supply- and demand-side constraints within the informal labor market. The results highlight the persistent barriers to formal employment for a significant segment of the labor force, despite policy efforts aimed at reducing informality in Mexico.

Keywords: informality, restricted choices, microsimulations, labor policies
JEL Classification: J24, J31, C54

1 Introduction

United Nations (2024) Sustainable Development Goals encourage the promotion of policies to support decent job creation and the formalization and growth of micro-, small- and medium-sized enterprises . Nonetheless, in Mexico during the 1Q2024, 32 million people- more than half of Mexican employed labor force (54.3%)-were informal workers, Instituto Nacional de Estadística y Geografía (2024b). One of the challenges faced by developing economies is the inclusion of large segments of the labor force into the modern sectors of the economy. A long-time recognized issue, for example Fields (1990), is heterogeneity in the informal sector. While there are economic activities that have low entry barriers and required low levels of skills, the informal sector also includes activities that are more flexible in their working hours and receive wages that are similar to the formal sector. Thus, informal employment can be a

subsistence strategy for some workers (avoiding unemployment), but also a convenient state by other workers given their preferences.

According to the Organization for Economic Co-operation and Development and the International Labor Office (2019), a worker is considered to have an informal job if he is outside of the regulatory framework of a country: he do not has contracts, do not pay income taxes, do not have work benefits or access to social secturity. This definition applies to workers who receive a wage, are self employed and even to those who do not receive a wage but are employed in activities that generate income. There is a strong, negative correlation between income per capita and the amount of workers with informal jobs; almost all the labor force in low income countries has informal jobs, while only a small fraction of workers in high income countries share this situation (ILO,(2018)) . According to the International Labor Office (2018), 53% of employed workers in Mexico had an informal job, similar to Colombia (60.6%), Brazil (43%) and Ecuador (59%), while the average for Western Europe and the US was 10.4%.

Despite several efforts made by the Mexican Government to reduce informal work, the issue remains pervasive. According to the latest data from the National Survey of Occupation and Employment (ENOE)(2024a), the labor informality rate has remained above 50% during the last 18 years, showing a slight decrease from 53.7% in 1Q2005 to 51.3% in 1Q2023[1]. In addition, informality is more prevalent in employed women and in least urbanized areas. Thus for the 4Q2023, while 48.2% of men worked in informal jobs, 54.6% of the women were in the same situation. In the same period, 41.9% of workers in highly urbanized areas were considered to have informal jobs, compared to 70% for rural areas.

According to the National Institute of Statistics and Geography (INEGI) in 2022 the informal economy accounts only for 24.4% of the gross added value of the economy (2024c). Considering that more than half of the workers have an informal job but only produce one quarter of the added value, there is a gap in labor productivity across formal and informal sectors. As the Economic Commission for Latin America and the Caribbean (ECLAC) suggests, there is a strong negative correlation between the labor informality rate and productivity levels at state level. On a theoretical and empirical level, Busso et al. (2012) show using data from INEGI's Economic Census that informal firms tend to be smaller, with less capital and therefore less productive, which contributes to Mexico's low Total Factor Productivity: "one (Mexican) peso of capital and labor allocated to formal firms is worth 50 percent more than in legal informal firms" (Busso et al., 2012).

This paper develops a discrete choice mixed logit model to explain the transition between labor states (not employed, formal employment, and informal employment) for individuals

[1]The stats do not include workers in agriculture and livestock sectors

between 18 and 65 years old. A utility model is proposed using disposable income, leisure hours, other socioeconomic characteristics of the household, and estimated wages for non-observed states. Formal employment is not always available to all individuals due to labor demand conditions and individual characteristics. Hence, a new version of the original model is proposed using restriction sets as proposed by Horowitz and Louviere (1995), estimating a latent class model variation that takes into account the probability of workers to be employed in the formal sector using firm data, fixed effects by localities, and other individual characteristics.

The proposed model seeks to contribute to the literature by specifying an estimation strategy that combines the insights and results of models such as Duval-Hernandez (2022) regarding the selection constrains faced by informal workers with the methodology of restricted sets to derive constraints faced by workers considering the demand for labor and the productive structure of firms. Similar to the strategy followed by Gong and van Soest (2002), the model uses wage equations to derive expected wages for individuals, expected leisure hours to obtain alternative specific coefficients, and socioeconomic characteristics of individuals and their households as case specific coefficients. Separating decisions of informal workers in a way consistent with economic theory, allows for a broader understanding of the problem and sensible public policies, versus assuming a false singleness.

The paper is organized as following. Section two gives some background on theory regarding informality, with some results for Latin America and Mexico. The model is developed in section three, including a detailed formulation of the inclusion of restriction sets in the utility specifications. Data is presented and explained in section four. Econometric results and their interpretation are given in section five. Brief conclusions are shown in the final section.

2 Previous studies

Informal employment can arise from different dynamics within the labor market. The Inter-American Development Bank (IADB) (2021) identifies three types of informal workers. 1) subsistence workers who would like to be formal but they have very low levels of productivity; 2)voluntary workers who decide to be informal due to more flexibility, or a perception that the benefits of having social security are less than the cost of paying taxes; and 3) induced informality, which consists on workers with the necessarily levels of labor productivity to be formal and with disposition to chose formal jobs, but there is not a corresponding demand for these type of workers due to regulations, taxes or low economic activity.

Several studies have characterized the causes and types of informal employment. Fields

(1990) building on previous classical dualistic models a la Harris-Todaro, establishes that urban informal workers that remain in that sector involuntary- or a result of migration or underemployment in cities- have jobs that can be characterized by low to non-entry barriers, low wages, lack of social protection, and irregular working hours. Although the author recognizes voluntary informal workers, he considers this groups as separate of the former.

According to Maloney (2004), informality in Mexico challenges theories that attribute informality to rigidities in the labor market due to red tape and unions, since there are other countries in the region with higher levels of taxes and regulation and lower levels of informality. Informal employment is heavy centered in self employed workers in firms of less than five employees, living in less urbanized areas and with lower levels of productivity. While the author recognizes the importance of voluntary workers, he also acknowledges the constraints workers face due to their education and economic context: "arguing that workers are voluntarily informal does not, of course, imply that they are not living in poverty, only that they would not obviously be better off in the formal jobs for which they are qualified. Being in the informal sector is often the optimal decision given their preferences, the constraints they face in terms of their level of human capital, and the level of formal sector labor productivity in the country."(Maloney, 2004, p. 1160)

The work of Ovando-Aldana et al. (2021) made a thorough literature revision on previous works on the informal sector and analyzes through a binary logistic model the determinants of informal employment for 2005 to 2020 to provide several stylized facts about the characterization of the sector in Mexico: 1) informality tends to be a buffer during economic crisis through underemployment; 2) the type of contract, the size of the firm of the employee, the education level, and the type of locality (urban or rural) explain the differences between formal and informal workers; 3) GDP per caita, the level of Foreign Direct Investment (FDI), the size of the market, and skills explain informality prevalence at state level; 4) the probability of a worker to be informal is higher in the agriculture sector compared to the secondary and tertiary sectors.

Fernández et al. (2017) standardized household data for six Latin American countries (Mexico, Brazil, Colombia, Peru, Uruguay and Chile) to estimate a multinomial logit with four states (formal, informal, unemployed and out of the labor force) to calculate the determinants associated to the choice made by people within the labor force. People with low education levels- i.e. without high school-, under 25 years of age, women and people in rural areas tend to be informal workers. In contrast, people with tertiary education, men, who live in cities with higher levels of productivity and higher wages tend to be employed in the formal sector. For the case of Mexico, the same authors estimate that 16% of informal workers are subsistence workers or induced into informality.

Gong and van Soest (2002) develop a dynamic multinominal logit model with panel data for Mexico using ENOE (the employment representative survey mentioned above). They develop a two dynamic wage equation to simulate the expected wage of each individual in the formal and informal sector. The authors estimate the effect of wage differentials and other socioeconomic characteristics in the transition of men and woman between three states: being informal, formal or not employed. Their definitions of formal and informal jobs are, however, different from the ILO deffinition provided at the begining of this section and the one used in this paper, since they consider the type of job and not if the worker has access to social security. Their findings suggest that 1) mobility between the formal and informal sector is larger than other OECD countries 2) wage differentials between the formal and informal are very large for highly educated individuals, while they are rather short or even negative for individuals with low levels of education, and 3) for women the number of income earners in their households decrease the probability of entering into the labor market.

Beyond the quantification of the size and type of informal employment, several studies have tried to identify to which degree is informality a self-selection process- i.e. a decision made by workers due to flexibility or less constraints- or if there is segmentation- informality is involuntary and therefore there is a share of informal workers who would like to have a formal job instead. Using data from Mexico's ENOE Alcaraz et al. (2015) propose a maximum likelihood probit model to estimate the probabilities of being formal and informal and then obtain the implicit probability of being accepted in the formal sector and as a result the share of involuntary informal workers. Their findings suggest that the proportion of involuntary informal workers is between 11% and 22% of total informal workers.

In contrast to the findings of Alcaraz et al. (2015) and Fernández et al. (2017), Duval-Hernandez (2022) develops a discrete probit choice model to explain the barriers workers face to have a formal job. Using data from ENOE along with INEGI's Module of Labor Trajectories (MOTRAL), the author can obtain data of how many informal workers would like to have benefits covered in formal employment- provided they earned a similar pre-tax wage. His results show that almost 80% of the informal workers in urban areas would prefer a job with social security even considering the taxes and contributions entailed. In addition, having more years of schooling increases the probability of being hired in the formal sector, while for women having more care work at home decreases them.

The literature review for Mexico and other LDC find evidence that the informal sector is heterogeneous in terms of the profile or workers, that there is a share of informal workers who would like to be in the formal sector for a similar wage, and that there are structural barriers for workers to enter into the formal sector related with the demand of work and the skills of workers.

3 Model

3.1 Environment and assumptions

The model employed takes as references the labor market models that simulate the decisions of workers using micro simulation techniques with discrete choice models presented in Singhal (2021) and Thoresen and Vatto (2015). These models analyze the decision-making process of people in the economically active population as of the selection of their labor options. Potential economically active population (PEAP) would be defined as people 18 years of age or older who are not studying.

The observed i.e. systematic utility of people depends on income, leisure and other labor and individual observed variables, and takes into account whether each worker is in the formal or informal sector, given the specific conditions and regulations of the current Mexican social protection system. Hence, market segmentation between formal and informal workers conditions access to social security and assumes people in the informal sector do not pay income taxes. It is also implies that the amount of time a person is not working for a wage is leisure. People with non-paid jobs will be excluded from the analysis since wages are part of the disposable income used in the utility function mentioned in the next section of the paper.

Being a discrete model, people choose from a set of finite states that consist of a "package" of predetermined conditions: working hours, tax scheme, benefits, and other variables. If a person wants to change to other job type, she must change to another state with a different package. This contrasts with continuous choice models and other labor markets where workers' wages and labor supply is set in terms of hours. Another assumption is that people can only choose one job, and labor options are mutually exclusive. People who report two jobs will be excluded from the model.

3.2 Labor states and discrete choice model with unrestricted sets

Let B and be the set of labor options. Each person from the economically active population will choose a state between a finite discrete choice of combinations of labor arrangements $b_i \in B$. There are three types of states:

- b_1 = not employed;
- b_2 = formal job, and
- b_3 = informal job

The definition of "not employed" refers to people who are not participating in the labor force (i.e. do not have paid work, regardless of whether they are actively looking for a job or not). The number of working hours is defined according to Chapter II of Mexico's Federal Labor Bill.

The social security criteria is also used to establish formal and informal jobs since in the case of Mexico, health and other social protection benefits are restricted only for workers properly registered in the Mexican Institute of Social Security (IMSS) or in the Institute for Social Security and Services for State Workers (ISSSTE). All the workers registered in these institutions- and their employers as well- have to pay monthly contributions. Since it is difficult to assess the fiscal situation of a person through household surveys, a key assumption for the model is that informal workers do not pay income taxes.

The utility function embedded in the model is similar to the ones proposed Singhal (2021) and Thoresen and Vatto (2015), first derived by McFadden (1974). It is divided in two parts: the first term is the systematic function of observable variables and the second a stochastic term that captures non-observable characteristics of the service or individual level preferences. So for any individual i from the sample who chose an option b, the utility can be expressed as

$$(1)\ U_{i,b} = V(y_{i,b}, L(H_{i,b}), X_i) + \varepsilon_{i,b} \text{ for any b}\in\text{B}$$

Where $y_{b,i}$ is the disposable income per capita of the person, $L(H_{b,i})$ is the leisure time of that person, which is a function of working hours, and X_i a vector of socioeconomic variables of each individual. Besides wages from i given an option b, this model, disposable income takes into account other income sources from the households, since consumption and overall well-being levels of a person does not depend only on their personal income given that income tends to normally be shared among members. The disposable income per capita is similar to the one proposed by Kornstad and Thoresen (2007) can be defined as

$$(2)\ y_{i,b} = \frac{w_{b,i}+A-TX(w_{b,i})}{HH}$$

Where w_b is the gross wage of the individual, A is the exogenous current income from other members of the households that is independent from option b chosen by the person, $TX(w_{b,i})$ is the income taxes and social security quotas paid by the individual i, and HH is the size of the hosehold adjusted by an equivalence scale[2].

Additionally, leisure is the number hours a week available (168) minus the hours H_b spent by the person i given their option b.

[2]An equivalence scale takes into account the number of people in the household as well as the age, since the necessary income for a given amount of consumption will have to be, for example, bigger for two adults compared to one adult and one child

$$(3)\ L(H_{b,i}) = 168 - H_{b,i}$$

As noted by Train (2009), the election problem can be posed as the person i will choose an option b_j over b_k if and only if it maximizes her utility, therefore

$$(4)\ U_{i,b_j} > U_{i,b_k} \forall b_j \neq b_j.$$

Considering the utility function in (1), the problem can be managed as the probability of a person i choosing alternative b_j as

$$\begin{aligned} P_i(b=j) &= P\left(U_{i,(b_j)} > U_{i,(b_k)} \forall b_j \neq b_k\right) \\ P_i(b=j) &= P\left(V_{i,b_j} + \varepsilon_{i,b_j} > V_{i,b_k} + \varepsilon_{i,b_j} \forall b_j \neq b_k\right) \\ (5)\ P_i(b=j) &= P\left(\varepsilon_{i,b_k} - \varepsilon_{i,b_j} < V_{i,b_j} - V_{i,b_k} \forall b_j \neq b_k\right) \end{aligned}$$

The probability of person i choosing option b_j will depend not only on the systematic part of the utility function but on the unobserved part. This can be expressed as a cumulative distribution of the errors $\varepsilon_{i,b}$. Therefore some assumptions on the errors should be made in order to derive a statistical model. In this case, following Kornstad and Thoresen (2007), the model assumes that the errors are i.i.d. and distributed as the standard type I extreme value distribution.

$$(6)\ F(\varepsilon_{i,b_k} - \varepsilon_{i,b_j}) = F(\varepsilon_{i,k,j}) = \frac{\exp\left(\varepsilon_{i,k,j}\right)}{1+\exp\left(\varepsilon_{i,k,j}\right)}$$

As a result, assuming independence of irrelevant alternatives following Train (2009) and Marschak (1960), the model with three labor states b can be expressed as a multinomial logit- which is consistent with random utility models [3]

$$(8)\ P_i(b=j) = \frac{\exp\left(V_{i,b_j}\right)}{\sum_{b=1}^{3} \exp\left(V_{i,b_k}\right)}$$

The proposed model is consistent with individuals choosing the option amongst the three different states from the set B that maximize their utility assuming a specific distribution of the unobserved part of utility so long that there is independence of irrelevant alternatives. A representative lineal utility function is proposed.

Considering equation (1), while the vector X_i is case-specific (i.e. they do not change over different states since it measures individual characteristics), disposable income and leisure are alternative-specific. Since the purpose of the model is to generate counterfactuals

[3]The mixed logit models presented in the Results section were re-estimated with the formal and informal state being dropped one at a time to see if the alternative specific coefficients changed in comparison with the option of not being employed. All of them remain statistically significant, with the same sign and very little variation in their value.

to run policy simulations, following Cameron and Trivedi (2005) a mixed logit form will be estimated, combining a multinomial specification with alternative conditional regressors. This will allow the calculation for any individual of the probability of being in each state under different fiscal policy scenarios.

As a result, for any option, $b \in B$ will be determined by (9)

$$(9)\ P_i(b=j) = \frac{exp\left(\alpha_1 y_{i,b_j} + \alpha_2 L(H_{b,j}) + \sum_{n=1}^{m} \beta_{n,b} X_{i,n}\right)}{1 + \sum_{k=2}^{3} exp\left(\alpha_1 y_{i,b_k} + \alpha_2 L(H_{i,b_k}) + \sum_{n=1}^{m} \beta_{n,b} X_{i,n}\right)}$$

Where coefficients α_1, α_2 estimate the effect of the variables within the utility function across options, while the case-specific vector β (i.e. socio-demographic variables) changes across states, which means that the likelihood of an individual to choose a certain state is conditioned by certain characteristics of the household. The term α_3 reflects the non-linear effect of leisure on the utility function. The model is estimated without any case-specific intercepts to allow variability due to changes in alternative specific regressors in the policy simulations.

So far, the model makes the assumption that every person in the labor force can choose the employment he wants, the one that maximizes his utility. However, as pointed by Alcaraz et al. (2015) and Duval-Hernandez (2022), some of the workers that choose to be employed in the informal sector are there involuntary, since they would have prefer a job in the formal sector. The following section will develop a discrete choice model variation to account for these restrictions.

3.3 Choice problem under uncertainty: consideration sets

One downside of estimating only equation (9) is that the model proposed assumed that all the options are available for and individual willing to enter into the labor market. However, some options might not be available in the decision set of an individual. In the case of informal workers, some of them might find a job in the formal sector as the best alternative according to their preferences; yet, those jobs might be unattainable given their levels of education or the characteristics of the labor demand in their place of residence.

Following the model proposed by Horowitz and Louviere (1995), consider B the set of all alternatives $B = (b_1, b_2, b_3)$ and B_c the known consideration set that denotes the attainable alternatives for the person. If all the options are available and considered by individuals, then $B_c = 3$. If there is a subset in the population of workers in the informal sector that could not choose a formal job, then only b_1 and b_3 are available therefore $B_c = 2$, so the alternatives are delimited by the consideration set C_{b_1,b_3}. For all individuals with that consideration set, the probability of choosing alternative b_2 will be zero. Therefore, probabilities for the other options can be stated as

$$(10) P(b_1, C_{b_1,b_3}|B_c = 2) = \left[\frac{\exp\left(V_{b_1}\right)}{\sum_{l=1}^{3}\exp\left(V_{b_l}\right)}\right] \times \left[\frac{\exp\left(V_{b_3}\right)}{\sum_{l=2}^{3}\exp\left(V_{b_l}\right)}\right]$$

$$(11) P(b_3, C_{b_1,b_2}|B_c = 2) = \left[\frac{\exp\left(V_{b_3}\right)}{\sum_{l=1}^{3}\exp\left(V_{b_l}\right)}\right] \times \left[\frac{\exp\left(V_{b_1}\right)}{\sum_{l=1}^{2}\exp\left(V_{b_l}\right)}\right]$$

According to the empirical findings of Horowitz and Louviere (1995), the model with consideration sets provides more efficient parameters compared to other models,including the multinomial logit without restrictions, taking into account that consideration sets use more information about consumer preferences. However, one of the main assumptions of the proposed model is that the consideration set B_c is known for individuals.

If the researcher knows the consideration set of individuals, the probability of choosing any available option if if $B_c = 3$ is just equation (9) . If $B_c = 2$, then the probability of selecting a formal job b_2 will be zero for all the individuals with this consideration sets, while the probabilities to select not to work (b_1) or to work in the informal sector (b_2) can be derived by combining (9) with (10) and (11)

$$(12)\ P(b_1, C_{b_1,b_3}|B_c = 2) = \frac{exp\left(\alpha_1 y_{i,b_1} + \alpha_2 L(H_{i,b_1}) + \alpha_3 (L(H_{i,b_1}))^2 + \sum_{n=1}^{m}\beta_{n,b1}X_{i,n}\right)}{1+\sum_{k=1}^{3} exp\left(\alpha_1 y_{i,b_k} + \alpha_2 L(H_{i,b_k}) + \alpha_3 (L(H_{i,b_k}))^2 + \sum_{n=1}^{m}\beta_{n,b}X_{i,n}\right)} \times \frac{exp\left(\alpha_1 y_{i,b_2} + \alpha_2 L(H_{i,b_2}) + \alpha_3 (L(H_{i,b_2}))^2 + \sum_{n=1}^{m}\beta_{n,b2}X_{i,n}\right)}{1+\sum_{k=2}^{3} exp\left(\alpha_1 y_{i,b_3} + \alpha_2 L(H_{i,b_3}) + \alpha_3 (L(H_{i,b_3}))^2 + \sum_{n=1}^{m}\beta_{n,b3}X_{i,n}\right)}$$

$$(13) P(b_3, C_{b_1,b_3}|B_c = 2) = \frac{exp\left(\alpha_1 y_{i,b_3} + \alpha_2 L(H_{i,b_3}) + \alpha_3 (L(H_{i,b_3}))^2 + \sum_{n=1}^{m}\beta_{n,b3}X_{i,n}\right)}{1+\sum_{k=1}^{3} exp\left(\alpha_1 y_{i,b_k} + \alpha_2 L(H_{i,b_k}) + \alpha_3 (L(H_{i,b_k}))^2 + \sum_{n=1}^{m}\beta_{n,b}X_{i,n}\right)} \times \frac{exp\left(\alpha_1 y_{i,b_1} + \alpha_2 L(H_{i,b_1}) + \alpha_3 (L(H_{i,b_1}))^2 + \sum_{n=1}^{m}\beta_{n,b1}X_{i,n}\right)}{1+\sum_{k=1}^{2} exp\left(\alpha_1 y_{i,b_3} + \alpha_2 L(H_{i,b_3}) + \alpha_3 (L(H_{i,b_3}))^2 + \sum_{n=1}^{m}\beta_{n,b3}X_{i,n}\right)}$$

Note that, for both (12) and (13), the left side of each equation is just the probability of that option estimated through equation (9) which is estimated along with the right side of each equation through the maximum likelihood method.

However, it is difficult to observe directly how many people who apply for jobs would be most likely to be excluded from formal employment. In order to estimate this, a probit model similar to the one made by Alcaraz et al. (2015) can be estimated to calculate the probability of a person to have a formal job, not only taking into account the personal characteristics of the individual i, but also the conditions of the production units and the demand of labor in the zone g in which she lives.

$$(14) P_{i,g}(formal) = \phi\left(X_i\gamma + M_i\omega + Z_g\theta\right)$$

Where ϕ is the cumulative density of normal distribution, X_i is a vector of individual characteristics, M_i a vector of the characteristics of the firms for people employed in their observed state, and Z_g a vector of state-level aggregated economic indicators that reflects

the productive structure of the zone where the individual lives. With the estimates of this model, probabilities of having a formal job can be calculated. For individuals who are not employed- and therefore data of the demand for labor cannot be observed- the model is estimated without $M_i\omega$.

Equation (14) estimates the probabilities for each individual to have a formal job, and therefore having a consideration set $B_c = 3$. Therefore, the sample can be divided into two groups: individuals who have in their choice set all the three alternatives available, and individuals who are likely to be involuntary informal workers, who can only choose between working in the informal sector or not being employed.

Therefore, a latent variable model can be estimated as a linear combination for each individual i of choosing an option j without any restriction or choosing an option j contained in restriction set $B_c = 2$

$$(15)\ P_i(b = j) = P(b_j, C_{b_1,b2,b3}|B_c = 3)P_{i,g}(formal) + P(b_j, C_{b_1,b3}|B_c = 2)(1 - P_{i,g}(formal))$$

Equation (15) can be calculated as a latent class mixed logit as shown by Train (2009), obtaining the vector parameters α, β, λ, ω and θ through a maxim likelihood estimation. It is worth noting the first term in the right side of the equation is the unrestricted mixed logit derived from equation (9) for any option j. The results section presents the results of the estimation using the unrestricted model and then the latent class model from equation (12) with estimations taking into account the consideration sets.

4 Data

The data used for the estimations comes from the 2022 National Survey of Household Income and Expenditure (ENIGH)(2022). Its sample is representative at national and state level and has information on employment characteristics, sources of income and socioeconomic variables. The model takes into account only individuals between 18 and 65 years of age, 140,898 individuals are considered.

For the selection of states b, the model assumes that someone has a job if the individual has worked at least an hour in the last week, has a subordinate employment relationship and received a wage[4]. Since the model is discrete, self employed individuals are excluded since they have more control of the hours spent working and also might face a different tax regime that is not associated with formality status, which is key in the development of this paper. Since the model also considers the demand for labor, there are issues modeling the demand

[4]ENIGH also reports workers who do not receive a payment- for example, members of a family who runs a business- or people who get paid with a non cash remuneration

and supply of work for self employed workers, given they somehow are simultaneously their own supply and demand.

Leisure is calculated using the working hours per week reported ad converting them as the percentage of total hours per month (729). The model defines leisure as the individual's available total time subtracting the hours worked.

To determine the formality status, the model adheres to the definitions of both ENOE and ILO to consider a worker informal if she does not have access to social security. Using ENIGH, an individual is considered to have access to social security if he reports having access to health insurance in his job through social protection institutions (such as IMSS, ISSSTE, PEMEX, etc). Household surveys in Mexico do not ask if people have paid income taxes, hence a sensible assumption is that informal workers do not pay income taxes.

For people defined as "not employed", the model considers two types of individuals: 1) those who are unemployed- i.e. actively looking for a job, and 2) people who are not looking for a job but could potentially take a job, mostly without a paid job doing care or domestic work. The model excludes people who do not have a job because they receive a pension, they are studying or they have a permanent disability that prevents them from working.

Disposable income per capita is calculated by the sum of 1) the labor income of the individual (wages plus tips, bonuses, Christmas bonuses, etc), plus 2) sources of income from other members of the household such as additional wages, income from pensions, rents and businesses plus transfers (both private and from Government programs), minus 3) income taxes in the case of formal workers, 4) divided by the size of household adjusted with the adult scale.

Net wages are considered in disposable income, as reported in the survey. Other sources of income are independent of taxes and i) the model assumes them as exogenous and therefore will not be considered in the policy simulations, ii) there is no detailed data at individual level to calculate the effective tax rate of other sources. In order to derive the tax rate from disposable income reported in ENIGH, the model uses the formula used by Absalón and Urzúa (2012) to recover the gross rate from ISR for the individuals of the model. These calculations will be used later for the policy simulations.

The equivalence scale used comes from the official methodology to calculate poverty measures by the National Council for the Evaluation of Social Policy CONEVAL(2023), which adjust the size of the household by weighting the age of each member of it.

Since we can only see the individual's selected state, for the case specific variables of the model (wages and leisure) estimations are made. To obtain alternative-specific wages, a OLS regression is used with a specification similar to Kornstad and Thoresen (2007), using as dependent variables the years of education, years of experience, the square of years

of experience, a dummy variable to account for people who have complete high school, a dummy variable to account for people who have complete a professional degree, a dummy variable for the formal and informal status and a categorical variable for six regions of the country from INEGI (2024e). These estimates are used to create a disposable income function that is alternative-specific. To estimate the hours for each state, the average working hours are imputed to each individual taking into account the state in which the individual lives, whether the job is formal or informal and whetheher the individual lives in a rural or urban locality.

The case-specific regressors of equation (9) are sex of the person, the type of locality in which he lives (rural or urban), the self declared ethnic group, whether his native language is indigenous, the type of household [5], a dummy variable to indicate if there is a senior citizen in the household, the number of income recipients in the household, a dummy to indicate availability of basic housing services[6], and a dummy to indicate a place with basic housing quality materials[7].

Finally, the variables used to estimate the probit model of equation (12) are derived from the characteristics of the firms declared by workers in ENIGH: size of the firm, type of employer [8], and economic sector using the North American Classification System (SCIAN). In addition to individual level reported characteristics, other variables at state level are added from ENOE 3Q2022, the Minisry of Economy(2024) and Mexico's national accounts(2024d) at state level to account for geographical fixed effects: the percentage of unemployed people who have been searching for a job more than six months, foreign direct investment per worker, GDP per worker, the labor critical conditions rate[9], and the economic complexity index score[10].

Table 1 and Table 2 show the descriptive statistics used in the mixed logit model using the sample design of the survey that has a sample of 140,898 observations. Table 8 shows the descriptive statistics used to estimate equation (12); since the available information is only reported for employed individuals and not all workers report all this variables, the sample used for this estimation is reduced to 89,567 observations; the state level variables are reported with their means at individual level.

[5] nuclear, single person, extended, compounded, co-habitation

[6] running water, sewerage, electricity, gas

[7] proper walls, floors, roof, and sufficient number of rooms for the number of people living in the household

[8] family business, corporation or limited liability firm, Government, NGO

[9] An indicator calculated by INEGI using ENOE. It measures the percentage of workers with precarious or inadequate labor conditions such as workers who work less than 35 hours a week for reasons external to their decisions, workers who work more than 35 hours a week with monthly earnings less than the minimum wage, and workers who work more than 48 hours a week with earnings less than two minimum wages

[10] An indicator calculated by the Ministry of Economics that measures the level of specialization, abilities and technical capabilities of a region that determines the technology, human capital and infrastructure

Table 1: Geographic distribution of the sample

	Mean	Standard Deviation
% living in Northwest region	0.1400	0.0013
% living in Northeast region	0.1330	0.0016
% living in West and Bajio region	0.2176	0.0023
% living in Mexico City	0.0780	0.0016
% living in Center-South and East region	0.2995	0.0028
% living in South region	0.1316	0.0015
Source: made by the author using ENIGH 2022		

Table 2: Descriptive statistics of the variables used in the model

	Mean	Standard Deviation
Employment status (Set B)		
% employed in formal sector	0.3861	0.0024
% employed in informal sector	0.3383	0.0023
% not employed	0.2754	0.0017
Independent variables		
Monthly labor income (Mexican pesos)*	6,820.31	9,446.94
Monthly household disposable income per capita (Mexican pesos) *	7,353.81	9,450.38
% Men	0.4649	0.0014
% Women	0.5350	0.0014
Age	38.56	12.80
Hours worked per week*	33.69	24.83
Years of education	13.54	4.43
Years of experience	22.02	14.4
% living in rural areas	0.2046	0.0020
% living in urban areas	0.7953	0.0020
% of self declared indigenous	0.2503	0.4332
% indigenous native language	0.0483	0.2145
% type of household single person	0.0381	.0008
% type of household nuclear	0.6038	0.0031
% type of household extended	0.3442	0.0032
% type of household compounded	0.0100	0.0006
% type of household co-habitation	0.0036	0.0003
% having a senior citizen in household	0.1657	0.3718
Number of income recipients	1.81	0.3862
% living with basic housing services	0.8563	0.3507
% living with basic housing quality	0.9238	0.2653
*The labor income, disposable income and working hours calculations include individuals who are not employed.		
Source: made by the author using ENIGH 2022		

Table 3: Descriptive statistics for probit model

	Mean	Standard Deviation
Size of the firm		
% working in micro	0.4788	0.0032
% working in small	0.2593	0.0025
% working in medium	0.1312	0.0020
% working in large	0.1305	0.0022
Type of employers		
% working in family business	0.3582	0.0032
% working in corporations/limited liability	0.4886	0.0033
% working in Government	0.1395	0.0022
% working in NGOs	0.0136	0.0006
Economic sector		
% working in agriculture and livestock	0.0786	0.0018
% working in mining, energy, water	0.0107	0.0005
% working in construction	0.0846	0.0016
% working in manufacturing	0.1991	0.0026
% working in retail and transportation	0.2245	0.0023
% working in professional services	0.1101	0.0018
% working in health, education, sports, culture	0.1151	0.0018
% working in other miscellaneous services	0.0460	0.0011
% working in Government agencies	0.0557	0.0013
% working in other non-classified sectors	0.0002	0.00007
State level variables		
Foreign Direct Investment per worker (USD)	$587.56	572.86
GDP per worker (MXN)	$434,759	182,088
% of unemployed in the state looking for a job more than 6 months	0.0616	0.0300
% of worker in the state in critical conditions	0.2923	0.0777
Economic complexity index	0.3164	0.9632
Source: made by the author using ENIGH and INEGI 2022		

5 Results

A Mincer equation as the one proposed by Heckman et al. (2003) using OLS is estimated to obtain the wages for the formal and informal state using a sample of 101,400 workers who have reported positive wages. The wage for the state of not employed is zero.

Table 4: Monthly Labor earnings estimation

	(I) Levels	(II) Levels	(III) Logs	(IV) Logs
Constant	-4,323.70***	-1,187.44***	7.22***	7.58***
	[262.75]	[271.68]	[0.0232]	[0.0221]
Education years	579.48***	372.36***	0.0697***	0.0385***
	[16.07]	[15.33]	[0.0015]	[0.0014]
Finished high school	2,886.85***	2,886.88***	0.1980***	0.1988***
	[132.88]	[128.17]	[0.0126]	[0.0116]
Finished professional studies	10,066.91***	10,280.95***	0.3458***	0.3774***
	[949.03]	[929.03]	[0.0241]	[0.0229]
Experience	373.87***	328.15***	0.0465***	0.0393***
	[8.62]	[8.09]	[0.0009]	[0.0009]
Experience squared	-5.38***	-4.82***	-0.0007***	-0.0006***
	[0.1598]	[0.1514]	[0.0002]	[0.0009]
Formal status		3,698.40***		0.6004***
		[67.53]		[0.0071]
Region (Northeast)		-650.32***		-0.0553***
		[142.18]		[0.0093]
Region (West and Bajio)		-1,632.48***		-0.1083***
		[114.29]		[0.0093]
Region (CDMX)		-219.73		-0.0564***
		[330.54]		[0.0147]
Region (Center-South and East)		-2,671.97***		-0.2579***
		[128.90]		[0.0097]
Region (South)		-2,751.59***		-0.3254***
		[110.23]		[0.0096]
Adj R^2	0.1979	0.2453	0.1998	0.3208
RMSE	8,900.3	8,633.5	7,945.63	7,679.08
N	101,400	101,400	101,299	101,299
Notes: Robust standard errors in brackets. ***p<0.01, **p<0.05, *p<0.1				

The estimation takes into account the education years, a premium for finishing high school and professional studies, years of experience and its squared, the a dummy for formal workers and a series of dummies for regional variation.

All the models in Table 4 show that education, experience, experience squared and being in the formal sector are statistically significant and show the signs normally observed in the literature. For all the models, the premium of finishing a professional career is large and statistical significant. In order to select the best model for estimated wages, adjusted R-squared and the root mean squared error were calculated. For the models using the log of wages, the root mean squared error is calculated after transforming into levels following (Wooldridge, 2002, p. 202) approach.

The selected model to estimate the expected wages for different non observed states was (IV) using as dependent variable the log of wages and regional dummies to account for geographical variation since it has the lowest adjusted R-squared and root mean squared error. Wages for informal and formal states are estimated for non observed states, while declared wages for observed states are kept to use as much as observed data as possible.

5.1 Results for unrestricted mixed logit model

The first model is the mixed logit with all the states available for all individuals from the sample assuming that the choice set of all workers consist in three alternatives. The baseline the scenario is individuals not employed to compare the differences in coefficients between choosing a formal and an informal job. Table 5 show the estimates for alternative-specific coefficient and Table 6 the case-specific coefficients.

Table 5: Mixed logit estimation Part(1/2)

Alternative specific coefficients α	
Disposable income	0.000115*** [0.0000012]
Leisure	0.01317*** [0.00037]
N	140,892
McFadden pseudo-R^2	0.3156
Notes: Robust standard errors in brackets. ***p<0.01, **p<0.05, *p<0.1	
The baseline state is "not employed"	

Table 6: Mixed logit estimation Part(2/2)

	Formal	Informal
Case specific coefficients β		
Woman	-2.87***	-2.84***
	[0.0010]	[0.0011]
Urban	0.6491***	-0.0153***
	[0.0010]	[0.0010]
Indigenous language	-0.6114***	-0.2072***
	[0.0021]	[0.0017]
Indigenous self ident.	-0.2986***	0.1006***
	[0.0009]	[0.0009]
Household (single person)	0.7797***	1.73***
	[0.0027]	[0.0029]
Household (extended)	-0.1724***	-0.1263***
	[0.0009]	[0.0009]
Household (compounded)	-0.0555***	0.0065
	[0.0037]	[0.0041]
Household (co-habitation)	1.26***	0.4669***
	[0.0097]	[0.0105]
Have senior citizen	-1.25***	-1.12***
	[0.0019]	[0.0019]
Women with senior citizen	1.26***	0.9320***
	[0.0022]	[0.0022]
Income recipients	1.21***	1.77***
	[0.0008]	[0.0008]
Basic housing services	0.3876***	-0.2563***
	[0.0013]	[0.0012]
Basic housing quality	-0.2652***	-0.4360***
	[0.0014]	[0.0013]
N	140,892	
McFadden pseudo-R^2	0.3156	
Notes: Robust standard errors in brackets. ***p<0.01, **p<0.05, *p<0.1		
The baseline state is "not employed". The baseline state for the variable "Household" is a nuclear family		

All the alternative-specific coefficients are statistically significant at the 99% confidence

level and show signs consistent with the literature and theory of labor supply. Both more disposable income and more hours of leisure per week in any given state increases the probability of selecting that alternative.

Some general insights can be drawn from the case-specific coefficient in terms of individual and households characteristics associated with the probabilities of entering into the labor market. The probability of being employed decreases for women for both formal and informal states. People living in urban areas are more likely to entering into the formal labor market, while it decreases the probability of working in the informal sector.

The composition of the household is also associated with the probability of entering into the labor market. Since the base state is people living in a nuclear family (i.e. a head of the household and other members of the family such as partners and/or siblings) the probability of choosing to work in either formal or informal jobs increases for individuals living in households with a single person or co-habitation, since it is less likely for a person to choosing to not work and have to share income from other individuals. In contrast, individuals living in extended households have a lower probability to transition into the labor market, since is more likely to share income with other members of the household, who tend to be relatives.

In addition, individuals who are women and live in a household with at least one senior citizen have a higher probability of entering into the labor market, and could be possibly explained by the fact that those senior citizen could be relatives who could help with the care work for children, as noted by Conelly (1992). Finally, individuals living in homes with basic housing services and quality decreases the probability to enter into the labor market-regardless of the sector- being these variables non income proxies for the socioeconomic status of the household and the type of assets and services within the house.

Table 7 compares the proportion of observed state with the mean of the simulated probabilities and their 95% level intervals. Taking into account that the model was estimated without a constant, the model shows very similar proportions compared to observed data for all the states, slightly overestimating the proportion of formal workers.

Table 7: Observed vs predicted probabilities of the model

	Observed			Predicted		
State	Mean	SD	95%CI	Mean	SD	95%CI
not employed	0.2754	0.0017	(0.2720,0.2789)	0.2700	0.2215	(0.2684,0.2716)
formal employment	0.3861	0.0024	(0.3813,0.3910)	0.3934	0.1689	(0.3917,0.3950)
informal employment	0.3383	0.0023	(0.3338,0.3429)	0.3365	0.1516	(0.3350,0.3380)
Note: Both observed data and model estimates presented use ENIGH 2022 sampling weights.						

5.2 Probabilities to select potential involuntary informal workers

Drawing from previous literature for Mexico, equation (14) is estimated under two different specifications (I) and (II) shown in Table 8. The first one takes into account 1) at individual level: the years of education and two dummy variables to account for having finished high school and a professional career; 2) the size, type of employer and economic sector of the firms of workers; and 3) a set of economic variables discussed in Section 5 related with formal labor market at state level. The first model is estimated over the sample of employed workers, while the other model is estimated the whole sample to include individuals not employed. The second one only estimates the first and third types of variables, since the demand for labor at firm level cannot be observed for individuals who are outside of the market- only the general conditions at state level.

Both models are used to calculate a subset of individuals who can be considered to be potential involuntary informal workers- a group where the choice set is restricted to the informal market given their skill level and the general conditions of the labor market. The first specification is restricted to wage workers and will be used to classify individuals currently working who are less likely to move to the formal market. The second one is estimated to the whole sample and will be used to classify individuals currently not employed but also less likely to find a formal job.

As showed in 8, while education increases the probability of being employed in the formal sector, only having a high school diploma decreases the probability of formality for already employed workers, while having finished professional studies increases it for both sets of the sample. Working in a corporation or limited liability private company or in government also increases the probability of being formal compared to being in a family business. All economic sectors that are associated with a larger probability of being employed compared to agriculture- which is consistent with previous literature (see Ovando-Aldana et al. (2021)); the sector with higher coefficients are manufacturing, mining and utilities, and professional and technical services. Finally, some of the insights from dualistic classical models mentioned in Fields (1990) and Maloney (2004) seems to hold in terms of informality as an option to buffer or avoid extended unemployment: being in states where there is a larger share of the labor force looking for jobs more than six months decreases the probability of being formal, while individuals living in states with higher economic complexity increases their chances of being employed in the formal sector.

Table 8: Probability of being employed in the formal sector

	(I) Levels	(II) Levels
Constant	-2.08*** [0.0021]	-1.72*** [0.0014]
Education years	0.0627*** [0.0001]	0.1119*** [0.00007]
Finished high school	-0.0898*** [0.0009]	0.0069*** [0.0006]
Finished professional studies	0.0194*** [0.0019]	0.3438*** [0.0014]
Private sector employer	1.17*** [0.0006]	
Public sector employer	1.51*** [0.0013]	
NGO employer	1.06*** [0.0023]	
Activity: mining and utilities	0.7884*** [0.0031]	
Activity: construction	0.3484*** [0.0014]	
Activity: manufacturing	0.7553*** [0.0012]	
Activity: trade and transportation	0.6321*** [0.0012]	
Activity: professional and technical services	0.7361*** [0.0013]	
Activity: health, education and recreation	0.5069*** [0.0015]	
Activity: accommodation and food	0.3617*** [0.0014]	
Activity: other services	0.3939*** [0.0016]	
Activity: government	0.3283*** [0.0019]	
FDI per worker	-0.00001*** [0.0007]	-0.00001*** [0.00005]
GDP per worker	0.00005*** [0.00003]	0.00003*** [0.00002]
Unemployed more than six months	-0.0153*** [0.0001]	-0.0125*** [0.0001]
Labor critical conditions rate	-0.0147*** [0.00004]	-0.0090*** [0.0002]
Economic complexity index	0.0938*** [0.0003]	0.1626*** [0.0002]
Pseudo R^2	0.4285	0.1252
N	89,567	140,892
Notes: Robust standard errors in brackets. ***p<0.01, **p<0.05, *p<0.1		

To select the sub set of the sample who could be potentially involuntary workers, estimated probabilities are calculated for both models and then transformed into dummy variables using different thresholds μ. Individuals below threshold μ are assumed to be in-

voluntary informal, using the first model (I) to classify informal workers and (II) to classify non employed workers as involuntary informal. Confusion matrices are then calculated to assess the quality in the classification of estimated probabilities against observed states.

Table 9 shows the results of the fitness of the models. As expected, model (I) has better prediction, since the sample for the second model takes into account the unemployed individuals who are by definition not working in the formal sector and have few information about the demand for work. As mentioned in Cameron and Trivedi (2022), accuracy predicts the percentage of observations correctly identified by the model, while precision calculates the percentage of observations predicted by the model as "formal" that are correctly identified as a share of the total "formal" predictions. Specificity calculates the percentage of observations predicted as "not formal" from the total of observations observed as "nor formal". Finally, the statistic F1 weights the precision of the model with other indicator, "recall", which measures the percentage of correctly estimated "formal" observations as a share of total observed "formal" observations.

Table 9: Indicators of confusion matrix of models (I) and (II)

	(I)				(II)			
μ	Accuracy	Precision	F1	Specificity	Accuracy	Precision	F1	Specificity
0.7	0.7947	0.8804	0.8014	0.8711	0.6768	0.6539	0.4139	0.9030
0.6	0.8201	0.8553	0.8369	0.8212	0.6724	0.6839	0.3601	0.9331
0.5	0.8277	0.8304	0.8508	0.7703	0.6857	0.6200	0.5072	0.8408
0.4	0.8264	0.8105	0.8542	0.7279	0.6704	0.5602	0.5724	0.7221
0.3	0.8183	0.7880	0.8518	0.6786	0.6025	0.4840	0.6087	0.4709

The selected models are the ones using as threshold $\mu = 0.4$ for (I) and $\mu = 0.3$ for (II), since they have a relatively high accuracy and F1. For working individuals, a potential involuntary worker will be all those individuals who have an observed state of "informal workers" and have a probability lower than 0.4 from model (I). For non-working individuals, a potential involuntary worker will be all those individuals who have an observed state of "not employed" and have a probability lower than 0.3 from model (II).

Taking these thresholds for (I) and (II), the model selected 46,397 individuals of the sample as potential involuntary workers, which represents 45% of the sample individuals not employed and 58% of the sample individuals working in the informal sector. Using sampling weights, the percentage of potential involuntary workers is 54% of the population of individuals not employed and 59% of the total population of informal workers.

Comparing the general socioeconomic characteristics of both groups, the results suggest

there is consistency between the statistical procedure to obtain the potential potential involuntary workers and the characterization made in the literature for Mexico for this group. People who could potentially be involuntary workers (i.e. they don't have formal jobs in their choice set) tend to be individuals living in low income households close to subsistence[11], they tend to live more in rural areas than the other comparison group, they mostly work in micro firms with less than five workers, and the majority of them live in areas of the country with high levels of informality[12]. Table 10 shows the descriptive stastitics for these variables.

Table 10: Descriptive statistics between groups

	Unrestricted choice set		Potential involuntary workers	
	Mean	Standard deviation	Mean	Standard deviation
Monthly Disposable income (Mexican pesos)*	8,713.33	10,766.29	4,540.42	4,636.17
Monthly labor income (Mexican pesos)*	8,505.52	10,716.63	3,332.94	4,143.54
% living in rural areas	0.1417	0.0020	0.3348	0.0043
% living in urban areas	0.8582	0.0020	0.6652	0.0043
% of workers who work in micro firms	0.3629	0.0032	0.9204	0.0029
% living in Center-South, East and South Region	0.3480	0.0047	0.6033	0.0085
*The labor income and disposable income calculations include individuals who are not employed.				

5.3 Results for mixed logit accounting for potential involuntary workers

With the two groups, the model from equation(9) is run simultaneously for both groups, with the group of potential involuntary workers having only the choice set reduced to not bein employed or working in the informal sector. Tables 11 and 12 show the results both alternative and case specific coefficients.

[11]According to the National Council for the Evaluation of Social Policy, the per capita poverty line for July 2022 was $4,105 Mexican pesos for urban areas and $2,928 for rural areas

[12]According to the data for 2Q2022 from ENOE, the states in those regions have the vast majority of workers as informal compared to the national average (55.7%): Oaxaca(81%), Guerrero(79%), Chiapas(77%), México(56%), Hidalgo(73%), Morelos(64%), Puebla(73%), Tlaxcala(72%), Veracruz(68%), Campeche(63%), Quintana Roo (47%), Tabasco(67%), and Yucatán(62%)

Table 11: Mixed logit estimation for two groups Part(1/2)

Alternative specific coefficients α		
	Unrestricted choice set	Potential involuntary workers
Disposable income	0.000090*** [0.0000012]	0.000162*** [0.0000068]
Leisure	0.0169*** [0.000025]	0.0184*** [0.000043]
N	94,496	46,396
McFadden pseudo-R^2	0.1376	0.3361
Notes: Robust standard errors in brackets. ***p<0.01, **p<0.05, *p<0.1		
The baseline state is "not employed"		

Both models have similar alternative-specific coefficients. One insight is that the disposable income elasticity might be higher for the potential voluntary workers. As noted by (Cameron and Trivedi, 2005, p. 502) the elasticity of choosing an option over alternative specific variables is a function of the alternative specific coefficient for any given probability of occurrence for an individual. For the case-specific coefficients, the differences between the two groups show that it is harder for women in the potential involuntary workers group to be employed compared to the unrestricted choice set group, even comparing between the same state (informal employment). The variable of changing from a rural to a urban area now has different effects across alternatives and groups. Now, living in an urban area is associated with a higher probability of working for potential involuntary workers but the effect on informal market is negative for the unrestricted choice set, which is consistent with the characterization of the two distinct informal workers made by Maloney (2004). The magnitude and signs for the composition of the household remain very similar across models.

Table 12: Mixed logit estimation for two groups Part(2/2)

Case specific coefficients β			
	Unrestricted choice set		Potential involuntary workers
	Formal	Informal	Informal
Woman	-2.57***	-2.11***	-3.39***
	[0.0013]	[0.0014]	[0.0018]
Urban	0.4061***	-0.1469***	0.3204***
	[0.0013]	[0.0015]	[0.0014]
Indigenous language	0.1228***	0.4083***	-0.1962***
	[0.0033]	[0.0033]	[0.0024]
Indigenous self ident.	0.0431***	0.3175***	-0.1131***
	[0.0012]	[0.0013]	[0.0014]
Household (single person)	1.68***	2.20***	1.21***
	[0.0035]	[0.0037]	[0.0044]
Household (extended)	-0.0961***	-0.0617***	-0.3345***
	[0.0011]	[0.0012]	[0.0014]
Household (compounded)	-0.2565***	-0.0624***	0.3157***
	[0.0044]	[0.0049]	[0.0060]
Household (co-habitation)	1.13***	0.5823***	-0.1087***
	[0.0103]	[0.0117]	[0.0221]
Have senior citizen	-1.21***	-1.05***	-0.9437***
	[0.0023]	[0.0026]	[0.0030]
Women with senior citizen	1.21***	0.9789***	0.6497***
	[0.0026]	[0.0030]	[0.0034]
Income recipients	1.77***	1.85***	1.74***
	[0.0010]	[0.0012]	[0.0014]
Basic housing services	-0.0407***	-0.6940***	0.1771***
	[0.0018]	[0.0019]	[0.0016]
Basic housing quality	-0.3031***	-0.7093***	0.0244***
	[0.0020]	[0.0021]	[0.0018]
N	94,496		46,396
McFadden pseudo-R^2	0.1376		0.3361
Notes: Robust standard errors in brackets. ***p<0.01, **p<0.05, *p<0.1			
The baseline state is "not employed". The baseline state for the variable "Household" is a nuclear family			

Similarly to section 6.1, the two models can be used to obtain the estimated probabilities. 13 compares the proportion of observed state with the mean of the simulated probabilities and their 95% level intervals for the two groups and their respective options. Both models slightly overestimate the share of workers in the informal sector, and for the one with unrestricted choice set slightly underestimates formal worker participation.

Table 13: Observed vs predicted probabilities of the model for the two groups

	Observed			Predicted		
State	Mean	SD	95%CI	Mean	SD	95%CI
Unrestricted choice set						
not employed	0.2207	0.0020	(0.2168,0.2246)	0.2230	0.1928	(0.2214,0.2246)
formal employment	0.5727	0.0026	(0.5675,0.5779)	0.5532	0.1874	(0.5516,0.5547)
informal employment	0.2065	0.0023	(0.2018,0.2112)	0.2237	0.1124	(0.2222,0.22519)
Potential involuntary workers						
not employed	0.3887	0.0034	(0.3819,0.3956)	0.3522	0.2894	(0.3502,0.3542)
informal employment	0.6112	0.0034	(0.6043,0.6180)	0.6477	0.2894	(0.6457,0.6497)
Note: Both observed data and model estimates presented use ENIGH 2022 sampling weights.						

6 Conclusions and discussion

The general results of this paper underscore the complexity of labor informality in Mexico. An econometric specification consistent with economic, a mixed logit model with restricted choice sets for the group of potential involuntary workers, proves valuable. The paper demonstrates that informal employment depends on preferences but is also significantly influenced by labor demand conditions and individual characteristics related with the skill level of individuals. Hence results show that a considerable portion of informal workers are in this segment involuntarily, facing structural barriers to entering formal employment, such as low education levels and weak labor demand in certain regions.

These insights suggest that policy interventions aimed at reducing informality should go beyond supply-side measures, focusing on improving access to formal jobs through targeted regional economic development and investment in human capital. A very promising line of research, is using the model's parameters to generate fiscal simulations that explore more realistically the effectiveness policies. Also, the model could aim the evaluation of interventions that foster transitions from informal to formal employment and long-term impacts on welfare and productivity in Mexico.

References

Alcaraz, C., D. Chiquiar, and A. Salcedo (2015). Informality and segmentation in the mexican labor market. *Banco de México. Documentos de investigación Working papers No. 2015-25*.

Busso, M., M. Fazio, and S. Levy (2012). (in)formal and (un)productive : the productivity costs of excessive informality in mexico. *IDB Working Papers Series IDB-WP-341*.

Cameron, C. and P. Trivedi (2005). *Microeconometrics Methods and Applications*. Cambridge University Press.

Cameron, C. and P. Trivedi (2022). *Microeconometrics Using Stata 2nd Ed.* Stata Press.

Cárdenas, M., C. Fernández, A. Rasteletti, and D. Zamora (2021). *Consideraciones para el diseno de políticas fiscales para reducir la informalidad en América Latina*. Banco Interamericano de Desarrollo.

Conelly, R. (1992). The effect of child care costs on married women's labour force participation. *The Review of Economics and Statistics 74*, 83–990.

Consejo Nacional de Evaluación de la Política de Desarrollo Social (2023). Medición de la pobreza. programas de cálculo 2016, 2018, 2020, 2022. https://www.coneval.org.mx/Medicion/MP/Paginas/Programas$_B$D$_2$022.*aspx*. *Accessed* : $2023 - 12 - 20$.

Duval-Hernandez, R. (2022). Choices and constraints: The nature of informal employment in urban mexico. *The Journal of Development Studies Vol 58 No. 27*, 1349–1362.

Fernández, C., L. Villar, and N. Gómez (2017). Taxonomía de la informalidad en américa latina. *Coyuntura Económica: Investigación Económica y Social XLVII*, 137–167.

Fields, G. (1990). *The informal sector revisited*, Chapter Labor market modelling and the urban informal sector: theory and evidence, pp. 46 – 69. Organization for Economic Co-operation and Development.

Gong, X. and A. van Soest (2002). Wage differentials and mobility in the urban labour market: a panel data analysis for mexico. *Labour Economics 9 Issue 4*, 513–529.

Heckman, J., L. Lochner, and P. Todd (2003). Fifty years of mincer earnings regressions. *IZA Discussion papers 775*.

Horowitz, J. and J. Louviere (1995). What is the role of consideration sets in choice modeling? *International Journal of Research in Marketing 12*, 39–54.

Instituto Nacional de Estadística y Geografía (2022). Encuesta nacional de ingreso y gasto de los hogares 2022. https://www.inegi.org.mx/programas/enigh/nc/2022/microdatos. Accessed: 2024-06-14.

Instituto Nacional de Estadística y Geografía (2024a). Encuesta nacional de ocupación y empleo. infolaboral. https://www.inegi.org.mx/sistemas/Infoenoe/Default$_1$5$mas_e n.aspx$. *Accessed* : $2024 - 04 - 25$.

Instituto Nacional de Estadística y Geografía (2024b). Encuesta nacional de ocupación y empleo. primer trimestre del 2024. https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2024/. Accessed: 2024-08-25.

Instituto Nacional de Estadística y Geografía (2024c). Medición de la economía informal. base 2018. https://www.inegi.org.mx/temas/pibmed/informacion$_g eneral$. *Accessed* : $2024 - 04 - 25$.

Instituto Nacional de Estadística y Geografía (2024d). Producto interno bruto por entidad federativa (pibe). ao base 2018. https://www.inegi.org.mx/programas/pibent/2018/tabulados. Accessed: 2024-06-14.

Instituto Nacional de Estadística y Geografía (2024e). Red nacional de metadatos. https://www.inegi.org.mx/rnm/index.php/catalog/434/variable/F12/V1179?name=REGION. Accessed: 2024-06-14.

International Labor Office (2018). *Women and Men in the informal economy: a statiscal picture (third edition)*. ILO.

Kornstad, T. and T. Thoresen (2007). A discrete choice model for labor surplus and child care. *Journal of Population Economics 20*(4), 781–803.

Maloney, W. (2004). Informality revisited. *World Development 32*, 1159–1178.

Marschak, J. (1960). Binary-choice constraints and random utility indicators. *Economic Information, Decision, and Prediction 7*, 218–239.

McFadden, D. (1974). *Frontiers in Econometrics*, Chapter Conditional logit analysis of qualitative choice behavior, pp. 105–142. Academic Press.

Organization for Economic Co-operation and Development (2019). *Tackling Vulnerability in the Informal Economy (AnnexA)*. OECD/ILO.

Ovando-Aldana, W., C. Rivera-Rojo, and M. Salgado-Vega (2021). Características del empleo informal en méxico, 2005 y 2020. *Papeles de Población 27*, 147–184.

Secretaría de Economía (2024). Complejidad económica a nivel regional. https://www.economia.gob.mx/datamexico/es/profile/. Accessed: 2024-06-14.

Singhal, N. (2021). Discrete choice models for estimating labor supply. *Congressional Budget Office Working papers Series No. 2021-4*.

Thoresen, T. and T. Vatto (2015). Validation of the discrete choice labor supply model by methods of the new tax responsiveness literature. *Labour Economics Vol 37*.

Train, K. (2009). *Discrete Choice Methods with Simulation, Second Edition*. Cambridge University Press.

United Nations (2024). Goal 8: Promote inclusive and sustainable economic growth, employment and decent work for all. https://unric.org/en/sdg-8/. Accessed: 2024-08-14.

Urzúa, C. and C. Absalón (2012). *Fiscal Inclusive Development: Microsimulation models for Latin America*, Chapter Distributive effects of the 2010 tax reform in Mexico: a microsimulation analysis, pp. 101 – 120. Instituto Tecnológico y de Estudios Superiores de Monterrey.

Wooldridge, J. (2002). *Introductory Econometrics: A Modern Approach 2nd Ed.* South-Western College Pub.