INTERNATIONAL MONETARY FUND

Unveiling the Informal Economy

An Augmented Factor Model Approach

Jiaxiong Yao

WP/24/110

2024
MAY

WORKING PAPER

 WP/24/110

IMF Working Paper
European Department

Unveiling the Informal Economy: An Augmented Factor Model Approach

Prepared by Jiaxiong Yao*

Authorized for distribution by James Walsh
May 2024

ABSTRACT: This paper develops a new approach to estimating the degree of informality in an economy. It combines direct yet infrequent measures of the informal economy in micro data with an augmented factor model that links macro indicators of the informal economy to its causes. We show that the prevailing model used in the literature, the multiple indicators multiple causes model, is a special case of the augmented factor model and depicts an incomplete picture of the informal economy. Using the augmented factor model approach, we show that the dynamics of the informal economy is shaped by the strength of overall economic activity as well as the interplay between the formal and informal economies. Contrary to previous work that typically finds declining informality for most countries, we find that the degree of informality has increased for low-income countries for the past two decades.

RECOMMENDED CITATION:

JEL Classification Numbers:	E26, C38, O11
Keywords:	Informality; augmented factor model; MIMIC model; survey data
Author's E-Mail Address:	JYao@imf.org

* The author thanks Shu Yu, Yuan Liao, James Walsh, and seminar participants at the IMF for helpful discussions. The views expressed here are those of the author and do not necessarily represent the views of the IMF, its Executive Board, or IMF management.

Contents Page

I. Introduction 4

II. Literature Review 7

III. Measuring the Informal Economy: An Augmented Factor Model Approach . . 8

IV. Data 12
A. Enterprise Survey Data 12
B. Other Survey Data 13
C. Causes and Indicators 14

V. Unveiling the Informal Economy 16
A. Projected Principal Component Analysis 16
B. Estimates of the Degree of Informality: Country Examples 20
C. Informal Economy: Causes and Patterns 24

VI. Conclusion 26

References 28

Appendices

A. The MIMIC Model 31

B. Estimation of the Augmented Factor Model 32

C. Relationship between WBES and Other Survey Data 33

D. Adding Nighttime Lights 33

E. Country Groups 36

List of Tables

1. Country and Year Coverage of the WBES 14
2. Causes and Indicators of the Informal Economy 16
3. Projection of Indicators on Causes 18
4. Relationship between Estimated Factors and Indicators 18
5. Survey Data and Estimated Factors 20
6. Estimated Weights on Projected Indicators 20
7. Estimated Informal Economy and Causes 25

8. Survey Data and Estimated Factors with Nighttime Lights 35
9. List of Countries with Causes, Indicators, and WBES Data 36
10. List of Countries in Each Income Group 37

List of Figures

1. Estimated Degree of Informality from the WBES 15
2. Projected Principal Component Analysis 19
3. Degree of Informality: Selected Countries 21
4. Decomposition of the Degree of Informality by Contribution of Factors: Selected Countries 22
5. Decomposition of the Degree of Informality by Contribution of Indicators: Selected Countries 23
6. Median Degree of Informality 26
7. Degree of Informality: 2002 vs. 2021 26
8. Degree of Informality in Survey Data: WBES vs. Labor Surveys 34
9. Projected Principal Component Analysis with Nighttime Light 35
10. Estimated Degree of Informality with and without Nighttime Light 35

I. Introduction

The informal economy is as mysterious as it is important: mysterious, because its size is often unfathomable—even its definition varies widely across different studies; important, because should it be formalized, it would potentially raise economic growth, boost government revenue, and improve social welfare.

Existing studies that estimate the size of the informal economy have broadly followed two approaches. The first is an empirical approach that attempts to obtain estimates directly from micro data. The obvious downside of this approach is that micro data are only available infrequently, making it difficult to assess the dynamics of the informal economy. The second is a statistical modeling approach that treats the informal economy as a latent variable and tries to estimate it from macro data of various causes and indicators. Such an approach is also not without limitations. A key problem is that the latent variable is subject to many different interpretations, rendering its exact meaning abstruse. In addition, the estimates of this approach are sensitive to model specifications, benchmark calibration, and sample coverage.

In this paper, we develop a new approach to estimating the degree of informality in an economy, marrying direct measures from surveys and insights from a statistical model and combining micro and macro data. We call this approach the augmented factor model approach, building on recent results on augmented factor models from Fan, Ke, and Liao (2021). It consists of two parts: one part is a factor model of indicators of the informal economy augmented by its observable causes; the other is a function that maps the estimated factors to direct measures in survey data. The augmented factor model summarizes the channels through which the causes of the informal economy affect its indicators, while the mapping into survey data weighs these different channels in terms of their relevance to the informal economy as defined in the survey. The estimated degree of informality is directly comparable to survey results, and therefore has direct interpretability.

The definition of the informal economy varies from study to study and it depends highly on the context.[1] Sometimes, concepts such as informal economy, unofficial economy, hidden economy, shadow economy, underground economy, illegal economy, among others, are used interchangeably. In other context, some of these concepts are mutually exclusive or a proper subset of another. For example, OECD (2002) classifies non-observed economy into the underground economy, the informal economy, and the illegal economy. In this paper, because our results are directly comparable to the survey data in use, the

[1] Schneider and Enste (2000),Chen (2012), Ohnsorge and Yu (2022), Dell'Anno (2022) provide excellent reviews of the definitions of the informal economy.

definition of the informal economy naturally follows that implied in the relevant survey questions. Specifically, we focus on the World Bank Enterprise Surveys (WBES), which ask questions about competition against unregistered or informal firms. It follows that the definition of the informal economy in this paper is most closely related to the definition by OECD (2002): productive activities conducted by unincorporated enterprises in the household sector that are unregistered and/or are less than a specified size in terms of employment, and that have some market production.

Since at least Frey and Weck-Hanneman (1984), the leading statistical model used to estimate the size of the informal economy has been the multiple indicators multiple causes (MIMIC) model. It links the causes and indicators of the informal economy through a latent variable. As its name alludes, the MIMIC model recognizes and leverages the multifaceted nature of the informal economy, making it superior to earlier single indicator approaches, such as the currency-demand approach and the electricity-consumption approach.[2] The limitations of the MIMIC model are also well known and widely discussed in the literature.[3]

A frequently-mentioned yet perhaps less appreciated drawback of the MIMIC model is the meaning of the latent variable. The latent variable is often interpreted as the informal economy. However, many causes of the informal economy, such as high tax, also affect the formal economy. Similarly, many indicators of the informal economy, such as labor participation rate, reflect formal economic activities as well. To the extent that the informal economy and the formal economy are correlated, the latent variable may well represent the formal economy in the MIMIC model. By contrast, the augmented factor model approach, developed in this paper, estimates more than one factors that link the causes and indicators of the informal economy. We show that the MIMIC model is a special case of the augmented factor model. Its latent variable is the first principal component of the augmented factor model under strong assumptions about the indicators. Furthermore, instead of treating the estimated factors directly as estimates of the size of the informal economy, we take a step further and use them as predictors of the degree of informality. We show that empirically, the first principal component, which is interpreted as the index of the informal economy in the MIMIC model, is often not a useful predictor of the degree of informality in survey data.

The augmented factor model has a few more appealing benefits. The projected principal component analysis helps us understand the channels through which the indicators of the

[2] See Cagan (1958), Tanzi (1983) for early examples of the currency-demand approach and Del Boca and Forte (1982), Kaufmann and Kaliberda (1996) for the electricity-consumption approach.

[3] Feige (2016),Schneider and Buehn (2017) provide more recent reviews of the criticisms of the MIMIC model.

informal economy are related to its causes. Despite the factors and loadings of the agumented factor model identifiable up to a rotation matrix, the mapping into survey data obviates the need for normalization and calibration, which in the MIMIC model can be sensitive to the choice of normalizing variable and the calibration year. The augmented factor model allows for decomposition of the estimated degree of informality into contributions by factors and by projected indicators, making it transparent the role of each factor or indicator plays in shaping the dynamics of the informal economy.

To obtain direct survey estimates of the degree of informality, we use the WBES, which covers 146 countries at infrequent intervals between 2006-2022. In principle, the augmented factor model approach does not limit us to any particular type of surveys. We focus on the WBES mainly because of its wide country coverage and long time span. With a simple metric of the prevalence of unregistered firms, we can infer the degree of informality using a consistent methodology across countries and over time. We then establish a predictive relationship between the estimated factors and survey estimates. The predicted degree of informality is then the estimates of the augmented factor model approach. As a robustness check, we also consider other labor-related measures of informality.

The estimates of the augmented factor model approach show three broad patterns of the informal economy. For countries such as Afghanistan and India, the informal economy has been increasing in the past two decades, and their falling labor participation rate has played a dominant role in indicating such increases. For countries such as China and Türkiye, the informal economy has been declining, with strong growth of the formal economy as the dominant indicator. For countries such as Greece and Italy, the informal economy displays a cyclical behavior, reflecting mostly the interplay between labor participation rate and GDP growth. Indicators such as currency in circulation and electricity consumption, intended to reflect the currency demand approach and electricity consumption approach in the literature, play a lesser role in indicating the dynamics of the informal economy.

With the estimates of the degree of informality, we find that economic development status and governance matter for the dynamics of the informal economy. The size of the government plays a relatively minor role. There is large heterogeneity among different country groups in terms of the degree of informality and the evolution of the informal economy. We find that in the past two decades, advanced economies and emerging markets have seen the degree of informality in overall economic activity steadily declining, while low-income and developing countries have experienced the opposite trend.

The rest of the paper is organized as follows. Section II reviews the relevant literature. In Section III, we introduce the augmented factor model approach. Section IV describes the

data and stylized facts. Section V presents diagnostic results from the augmented factor model as well as the estimated degree of informality. Section VI concludes.

II. Literature Review

There is a vast literature on the informal economy. The literature review here is not meant to be exhaustive, but we discuss several strands of selected studies that are closely related to estimating the degree of informality.

As alluded earlier, the informal economy has various names. Terms such as informal economy (Chen, 2012; Ohnsorge and Yu, 2022), unofficial economy (Choi and Thum, 2005), hidden economy (Giles, 1999), shadow economy (Schneider and Enste, 2000), and underground economy (Capasso and Jappelli, 2013) are commonly used. Sometimes these terms are used interchangeably, but they reflect different schools of thoughts on the nature of the informal economy (Chen, 2012). In this paper, we sidestep the complexity of the definition of the informal economy by following the definition of the survey data in use.

In general, there are two approaches to estimating the degree of informality. One is a direct empirical approach that obtains estimates from micro level data. As data availability and quality get better over time, the literature has evolved from early use of survey data (Gorodnichenko, Martinez-Vazquez, and Sabirianova Peter, 2009; Isachsen and Strøm, 1985; Van Eck and Kazemier, 1988) to recent use of administrative data (Braguinsky, Mityakov, and Liscovich, 2014; Waseem, 2023) and bank credit data (Artavanis, Morse, and Tsoutsoura, 2016). Still, the availability of such data is limited to a few advanced countries and large emerging markets at specific points in time. It is therefore difficult, if not impossible, to assess the dynamics of the informal economy with this approach.

The other is an indirect approach that estimates the size of the informal economy from statistical models. In the early literature, single indicators of the shadow economy were used to estimate its size. For example, Del Boca and Forte (1982); Kaufmann and Kaliberda (1996) use electricity consumption, with the premise that informal firms also use electricity and electricity consumption reflects economic activity more accurately than official GDP. Cagan (1958); Schneider (1986); Tanzi (1983) use currency demand, noting that informal economic activities prefer cash transactions. An obvious problem with the single indicator approach is that it only reflects one particular aspect of the informal economy.

Since Frey and Weck-Hanneman (1984) introduced the MIMIC model (Jöreskog and Goldberger, 1975) to the literature, it has been the prevailing statistical modeling approach

for estimating the size of the informal economy. The attractiveness of the MIMIC model is that it combines multiple determinants and multiple indicators of the informal economy, reflecting more aspects of the informal economy and lending more credibility to its results. Such a modeling approach has by and large stayed the same over the past decades (Schneider and Buehn, 2017; Schneider and Enste, 2000), producing numerous estimates of the size of the informal economy at national and subnational levels (Dell'Anno, 2007; Elgin, Schneider and others, 2016; Medina and Schneider, 2017, 2018, 2019; Vuletin, 2008). However, the MIMIC model is not without criticism. Feige (2016) challenges the MIMIC model on its conceptual flaws, violation of economic and statistical assumptions, and malleability of results. Schneider and Buehn (2017) highlight three most important points of criticism, including the meaning of the latent variable, unstable coefficients, and reliance on external calibration. This paper combines the strength of the two existing approaches in the literature, utilizing micro data while extending the MIMIC model to a more general augmented factor model.

A separate strand of the literature estimates the size of the informal economy using dynamic general equilibrium models (Elgin and Oztunali, 2012; Ihrig and Moe, 2004; Orsi, Raggi, and Turino, 2014). This approach needs to specify explicit causal channels of the informal economy, often through tax rates and enforcement, and therefore likely misses other channels that give rise to the informal economy. Like the MIMIC model, it also requires calibration of the size of the informal economy in a base year from another independent study, which can be arbitrary. Ohnsorge and Yu (2022) provides a recent review of different approaches to estimating the size of the informal economy, their advantages and disadvantages.

III. Measuring the Informal Economy: An Augmented Factor Model Approach

In this section, we develop a new approach to estimating the degree of informality, which we call the augmented factor model approach. The approach consists of two parts. The first part is a factor model of indicators of the informal economy augmented by its observable causes. It summarizes the main channels through which observable causes affect indicators of the informal economy. The second part is a function that maps the estimated factors to direct measures of informality in survey data. Such a mapping makes the model directly comparable to survey data.

The augmented factor model

We consider the following factor model

$$\boldsymbol{y}_t = \boldsymbol{\Lambda} \boldsymbol{f}_t + \boldsymbol{u}_t, \tag{1}$$

where $\boldsymbol{y}_t = (y_{1t}, \cdots, y_{Pt})'$ is a $P \times 1$ vector of indicators of the informal economy; $\boldsymbol{f}_t$ is a $K-$dimensional vector of latent factors that summarizes the informal economy; $\boldsymbol{\Lambda} = (\boldsymbol{\lambda_1}, \cdots, \boldsymbol{\lambda_P})'$ is a $P \times K$ matrix of factor loadings; and $\boldsymbol{u}_t = (u_{1t}, \cdots, u_{Pt})'$ is a vector of errors.

The latent factors $\boldsymbol{f}_t$ are explained by a $Q \times 1$ vector of observable causes $\boldsymbol{x}_t = (x_{1t}, \cdots, x_{Qt})'$ through the following model:

$$\boldsymbol{f}_t = g(\boldsymbol{x}_t) + \boldsymbol{v}_t, \tag{2}$$

where $g(\boldsymbol{x}_t) = E(\boldsymbol{f}_t | \boldsymbol{x}_t)$ is the component of the latent factors that is explained by the causes and $\boldsymbol{v}_t$ is the component of the latent factors unexplained by the causes.

Equations (1) and (2) constitute an augmented factor model. The prevailing MIMIC model[4] in the literature that estimates the size of the informal economy is a special case of the augmented factor model, which assumes that there is only one latent factor and that the relationship between the causes and the latent factor is linear. The augmented factor model is also used in the finance literature to explain stock and bond returns, where $g(x)$ is typically assumed to be nonlinear (Adrian, Crump, and Vogt, 2019).

The augmented factor model has exact identification, where $\boldsymbol{\Lambda}$ and $g(\boldsymbol{x}_t)$ can be identified up to a rotation matrix transformation (Fan, Ke, and Liao, 2021). The loadings and the factors can be estimated through projected principal component analysis: first, regress $\{\boldsymbol{y}_t\}$ on $\{\boldsymbol{x}_t\}$ and obtain fitted value $\{\hat{\boldsymbol{y}}_t\}$; second, obtain $\hat{\boldsymbol{\Lambda}}$ as eigenvectors of the projected indicators $\{\hat{\boldsymbol{y}}_t\}$ and $\hat{g}(\boldsymbol{x}_t)$ as the projected indicators multiplied by the eigenvector matrix.

The explained component of the latent factors, $g(\boldsymbol{x}_t)$, is often interpreted directly as the level or the growth rate of the informal economy in the literature. However, this can be problematic because it is only unique up to an orthogonal transformation. In this paper, we use it as a predictor of the informal economy, which not only yields the same predictions with any orthogonal transformation but has economic grounds as the explained component summarizes the channels through which the causes affect the indicators of the informal economy.

Mapping of factors to survey data

Since empirical data often contains information on the level of the informal economy at infrequent points in time, we propose the following outline to map the explained compo-

[4] See Appendix A for a detailed explanation of the MIMIC model.

nent of the latent factors to the level of the informal economy. Appendix B provides more details on each step of the estimation procedure.

Step 1. Transform the indicators $\boldsymbol{y}_t$ into growth rates, i.e., the first difference in log levels.

Step 2. Regress $\{\boldsymbol{y}_t\}$ on $\{\boldsymbol{x}_t\}$ and obtain fitted value $\{\hat{\boldsymbol{y}}_t\}$. $\{\boldsymbol{x}_t\}$ include country fixed effects to take into account time-invariant reasons for informal economic activity, such as cultural norms. $\{\hat{\boldsymbol{y}}_t\}$ are then standardized to have a mean of zero and variance of one.

Step 3. Obtain the explained components of the factors through principal component analysis:

$$\hat{\boldsymbol{g}}_t := \hat{g}(\boldsymbol{x}_t) = \hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{y}}_t = \boldsymbol{\lambda}_1'\hat{y}_{1t} + \cdots + \boldsymbol{\lambda}_P'\hat{y}_{Pt}, \tag{3}$$

where $\hat{\boldsymbol{g}}_t = (\hat{g}_{1t}, \cdots, \hat{g}_{Kt})'$ is a $K \times 1$ vector.

Each element, $\hat{g}_{kt}$, is a linear combination of the projected indicators, $\{\hat{\boldsymbol{y}}_t\}$. As such it can be interpreted as a growth rate and is therefore additive. The number of factors can be consistently estimated by methods such as AIC-based criteria. For now, we assume it is known and equal to K.

Step 4. Construct indices of the level of the informal economy. For each factor k, let

$$\hat{s}_{kt} = \sum_{t'=1}^{t} \hat{g}_{kt'}, \qquad \text{for} \quad k = 1, \cdots, K \tag{4}$$

be its cumulative sum up to time t. Then $\hat{s}_{kt}$ is an index of the level of the informal economy that can be mapped into survey data. $\hat{\boldsymbol{s}}_t = (\hat{s}_{1t}, \cdots, \hat{s}_{Kt})'$ are $K \times 1$ indices of the level of the informal economy.

Step 5. Map indices into survey data. Let z_t be the degree of informality estimated from survey data. We conduct the following panel regression to map the indices to the data

$$z_t = \boldsymbol{\beta}'\hat{\boldsymbol{s}}_t + \zeta_0 + \varepsilon_t, \tag{5}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_K)'$ is a vector of coefficients before the indices and ζ_0 is a country fixed effect. Equation (5) can be estimated in a cross-country panel data. For any country, because surveys are usually conducted at irregular intervals, z_t will be missing for many countries at time t. However, because $\hat{\boldsymbol{s}}_t$ is non-missing, the prediction of the regression equation (5), $\hat{z}_t = \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{s}}_t + \hat{\zeta}_0$, will be the estimated degree of informality that covers all countries at all times.

Comparison with the MIMIC model

The augmented factor model approach described above has three key advantages compared to the prevailing model in the literature, the so-called MIMIC model, in estimating the size of the informal economy.

First, the MIMIC model is a special case of the the augmented factor model. Specifically, the MIMIC model assumes that there is only one latent variable that linearly depends on the causes and that the indicators are independent of each other conditioning on the latent variable. In other words, it assumes that $\boldsymbol{f}_t$ is a scalar ($K = 1$), $g(\boldsymbol{x}_t)$ is linear, and the covariance matrix of $\boldsymbol{u}_t$ is diagonal.

Such simplifying assumptions can be strong. Crucially, because most of the causes affect both formal and informal economic activities and most of the indicators reflect both as well, it is a strong statement that the single latent variable itself is an index of the informal economy. By contrast, allowing $\boldsymbol{f}_t$ to have more than one factors in the augmented factor model approach has the benefit of capturing many different channels through which the causes affect the indicators. Moreover, commonly used indicators of the informal economy, such as GDP growth and labor participation rate, often have correlated measurement errors, rendering the covariance matrix of $\boldsymbol{u}_t$ not diagonal.

Second, the mapping of the indices to survey data in equation (5) makes the estimated size of the informal economy robust to choices of causes and indicators. This is because the indices are used as predictors of the informal economy and their predictions are ultimately disciplined by survey data. By contrast, the MIMIC model approach, through normalization, assumes that the latent factor has the same unit as one of the indicators, which can lead to erroneous estimates if the normalizing indicator is highly correlated with other indicators.

Third, the augmented factor model approach has better transparency and interpretability. It allows for decomposition of the estimated size of the informal economy by indices and by projected indicators. Such decomposition helps us understand the leading factors that influence the dynamics of the informal economy and the dominant indicators that signal such dynamics.

Decomposition by indices and by projected indicators

As the mapping of of the indices to the survey data is linear in equation (5), one can readily examine the contributions of each index, $\beta_k \hat{s}_{kt}$, to the degree of informality, $\hat{z}_t$. Since the latent variable in the MIMIC model is essentially the index that corresponds to the first principal component of the projected indicators under, such decomposition allows us to examine the extent to which it is related to the survey data.

Furthermore, as each index is a linear combination of the projected indicators, one can assess how the estimated degree of informality is linked to the projected indicators and which projected indicators dominate the dynamics of the informal economy. To see this, let $y^c_{pt} = \sum_{t'=1}^{t} y_{pt'}$ be the cumulative sum of the pth indicator. From equations (3) and (4), we have

$$\begin{aligned} \hat{s}_{kt} &= \sum_{t'=1}^{t} (\hat{\lambda}_{k1}\hat{y}_{1t'} + \cdots + \hat{\lambda}_{kP}\hat{y}_{Pt'}) \\ &= \hat{\lambda}_{k1}\hat{y}^c_{1t} + \cdots + \hat{\lambda}_{kP}\hat{y}^c_{Pt} \qquad \text{for} \quad k = 1, \cdots, K, \end{aligned}$$

or in matrix form,

$$\hat{\boldsymbol{s}}_t = \hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{y}}^c_t.$$

It follows from equation (5) that:

$$\hat{z}_t = \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{s}}_t + \hat{\zeta}_0 = (\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\Lambda}}')\hat{\boldsymbol{y}}^c_t + \hat{\zeta}_0, \tag{6}$$

which allows for the decomposition of the estimated degree of informality into contributions from the cumulative sum of each projected indicator, y^c_{pt}.

IV. Data

A. Enterprise Survey Data

An important step in the augmented factor model approach is to map the estimated informal economy indices into measures of informality in survey data (equation (5)). To this end, we use all available waves of the World Bank Enterprise Surveys (WBES), which cover a wide range of countries with the same questions. We focus on the WBES because it measures informality from the perspective of firms, whose output is related to the concept of GDP and depends not only on labor but on capital and production technology. We also consider other survey-based measures that tend to have only an employment angle.

Specifically, each WBES survey contains two questions that are relevant to the degree of informality:

- Does this establishment compete against unregistered or informal firms?
- Number of permanent, full-time employees at the end of last fiscal year.

The first question allows us to construct a simple measure of informality as the fraction of firms claiming that they compete with unregistered firms. With the answer to the second question, we can adjust the measure of informality by firm size. The adjusted measure of informality gauges the fraction of workers in registered firms that compete with unregistered firms. We focus on the adjusted measure and call it the degree of informality.

The informality in the WBES survey pertains to unregistered firms. It is likely that many registered firms compete with only a few unregistered firm or only a few registered firm competes with many unregistered firms. Unregistered economic activity can also take place in registered firms. With such complications, the degree of informality measured by the WBES should be viewed as the perceived prevalence of informality. While it does not directly measure the size of the informal economy, it is highly correlated with the size.

Table 1 presents the country and year coverage of all available data in the World Bank Enterprise Surveys. The number of surveys is uneven over time and across countries. In total, the WBES covers 154 countries between 2006-2022. About two thirds of the countries have at least two waves in the WBES.

Figure 1 presents the degree of informality using the latest wave for each country. Latin America and Africa tend to have the largest degree of informality. Eastern and Southern European countries have higher degree of informality than Western European countries. In Asia, China and Indonesia stand out as having sizeable informal economies.

B. Other Survey Data

In addition to the perception of the prevalence of unregistered firms, measuring the informal economy from the employment perspective is also common in the literature. Ohnsorge and Yu (2022) provide a database based on various labor force surveys. It has four measures of labor-related informality, including the share of self employment, the share of the informal employment, the share of employment outside the formal sector, as well as the share of the labor force that contributes to a retirement pension scheme.

All of these measures have limited country and year coverage, as is the WBES. We show in Appendix C that they tend to have low correlations with the degree of informality estimated from the WBES, reflecting that they represent different concepts of informality. Because the augmented factor model approach only relies on macro data of observable causes and indicators, the estimated indices of the informal economy are independent of

Table 1. Country and Year Coverage of the WBES

coverage by year			
year	number of countries	year	number of countries
2006	27	2015	9
2007	12	2016	18
2008	7	2017	11
2009	52	2018	8
2010	35	2019	39
2011	6	2020	11
2012	3	2021	5
2013	48	2022	4
2014	12		

coverage by country	
number of surveys	number of countries
At least 1 survey	154
At least 2 surveys	101
At least 3 surveys	50
4 surveys	2

Note: This table presents the country and year coverage by the World Bank Enterprise Surveys. The upper panel shows the number of countries surveyed each year. The lower panel shows the total number of surveys for each country between 2006-2022.

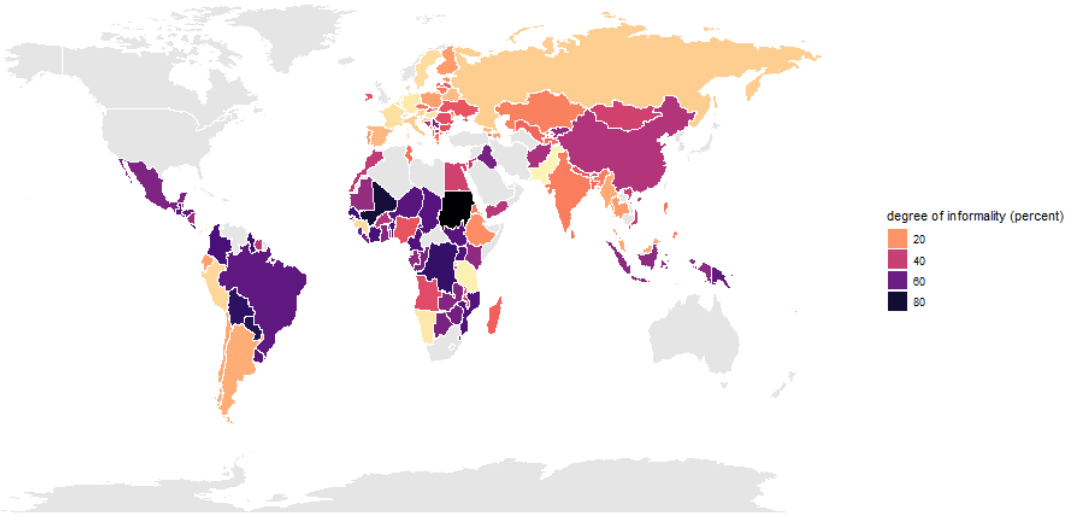
survey data, and they can be used to predict the degree of informality in any survey data, even if the concepts of the informal economy in these survey data differ. As will be clear in Section V, our estimated indices have explanatory power over most of them.

C. Causes and Indicators

The literature has proposed a variety of causes and indicators of the informal economy. The main causes are attributed to development status, tax, trade, unemployment, and governance. Each cause reflects an economic incentive for participation in the informal economy. For example, a higher tax rate could induce more informal economic activities for tax evasion reasons. The leading indicators include per capita GDP growth, currency growth, labor participation rate, and electricity consumption. They reflect various indirect measuring approaches in the past.[5] For instance, a higher labor participation rate indicates more participation in the formal economy, suggesting a smaller informal sector.

[5]Table 1 in Medina and Schneider (2018) provides an overview of rationale behind the main causes and indicators.

Figure 1. Estimated Degree of Informality from the WBES



Note: This figure presents the degree of informality as measured by the fraction of workers in registered firms that compete with unregistered firms in the World Bank Enterprise Surveys. For each country, the latest wave is used.

Table 2 presents the variables that we use in this paper, with the data from the World Bank and the International Monetary Fund. The literature has also considered alternative measures of economic activity, such as satellite-recorded nighttime lights (Medina and Schneider, 2018), as indicators of the informal economy. We add it as a robustness check.

The causes and indicators we use are not meant to be exhaustive. In principle, more causes and indicators can be added. As noted in Section III, one benefit of the augmented factor model approach is that it is constructive, because it allows us to examine the explanatory power of the causes as well as the relevance of indicators.

In total, we have a panel data of 154 countries spanning between 1996 and 2021.[6] Note that estimating the indices of the informal economy from the augmented factor model for a country only requires that the country has cause and indicator variables. It does not require the availability of WBES data. It is only when we map the indices to the survey data that we need the WBES. In total, there are 126 countries for which all cause and indicator variables are available and there is at least one wave of WBES. Table 9 in Appendix E lists all these countries.

[6]Because governance indicators are available biannually before 2002, the panel data covers 1996, 1998, 2000, 2002, and each year after.

Table 2. Causes and Indicators of the Informal Economy

variable	source
Causes	
PPP GDP per capita, unemployment rate	World Development Indicators
Rule of law, control of corruption, government effectiveness, voice and accountability, regulatory quality, political stability and absence of violence/terrorism	Worldwide Governance Indicators
Trade openness, tax-to-GDP ratio, government consumption-to-GDP	World Economic Outlook
Indicators	
PPP GDP per capita	World Development Indicators
Currency in circulation	International Financial Statistics
Labor participation rate (aged 15-64)	World Development Indicators
Electricity consumption	World Development Indicators
Nighttime light	Beyer, Hu, and Yao (2022); Hu and Yao (2022)

V. Unveiling the Informal Economy

In this section, we first analyze the factors driving the dynamics of the informal economy. We show that overall economic activity and the co-movement of formal and informal economic activity matter for the dynamics. Next, we provide some country examples highlighting distinct patterns of the informal economy in various countries. Finally, we investigate the causes of the patterns of the informal economy.

A. Projected Principal Component Analysis

Following the steps described in Section III, we first transform all variables into growth rates. In the second step, we conduct regressions of the indicators of the informal economy on its causes to obtain the projected indicators.[7]

The regression results in Table 3 suggest that both formal and informal economic activity could be channels through which the causes affect the indicators. In fact, the association between the indicators and most causes actually reflects the role of formal economic activity. For example, higher trade openness and lower unemployment rate are associated with higher growth rates of all the indicators. This can be plausibly attributed to effects of a stronger formal economy. Higher government consumption is negatively correlated

[7]More detailed steps are described in Appendix B.

with the indicators, possibly reflecting that the public sector is crowding-out the private sector of the formal economy. The negative coefficients before GDP per capita indicates the natural slowdown of the formal economy as it matures. To some extent, the informal economy provides a countervailing effect on the indicators. For instance, the association between tax and the indicators is not statistically strong. One interpretation is that higher tax reduces formal economic activity but encourages informal economic activity, and the two forces balance out.

The results in Table 3 highlight that the indicators of the informal economy may well be indicators of the formal economy, and the causes of the informal economy also affect the formal economy. To assume there is only one latent variable through which the causes affect the indicators, as the MIMIC model does, is therefore a strong assumption. Moreover, the formal economy and the informal economy are correlated (Ohnsorge and Yu, 2022), and they can be influenced by common factors. Stronger governance, for example, might enhance the formal economy while weaken the informal economy. To the extent that each governance measure affects one more than the other, its effect on the indicators is different. As such it is an empirical question whether the sign of the coefficient before each governance measure is positive or negative. Table 3 shows that strong government effectiveness is associated with slower growth of currency in circulation and stronger growth of labor participation rate, while political stability is positively linked to GDP per capita growth and electricity consumption growth.

In the third step, we conduct a principal component analysis of the projected indicators from the regressions in Table 3.

In Figure 2, panel (a) shows that the first two principal components have eigenvalues greater than 1, suggesting that at least two factors should be chosen.[8] This again confirms that it is more reasonable to assume that there are at least two latent variables through which the causes of the informal economy affect the indicators. Panel (b) shows that cumulatively, the first two principal components explain about 67% of the variance of the projected indicators.

Panel (c) of Figure 2 shows that the first principal component has higher loadings on currency in circulation growth and electricity consumption growth. It is related to the currency-demand approach and the electricity-consumption approach in the literature, and can be viewed as a physical indicator of overall economic activity. The second princi-

[8]This is also consistent with the eigenvalue ratio approach in Lam and Yao (2012), where all eigenvalues are plotted in a descending order and the place that has the steepest gradient is the number of factors to be chosen.

Table 3. Projection of Indicators on Causes

	(1) currency in circulation	(2) labor participation rate	(3) GDP per capita	(4) electricity consumption
tax-GDP ratio	0.21	-0.00012	0.090	0.063
	(0.18)	(0.011)	(0.057)	(0.058)
trade openness	0.057**	0.0057***	0.048***	0.023*
	(0.022)	(0.0021)	(0.013)	(0.012)
government consumption-GDP ratio	-0.16	-0.042***	-0.33**	-0.16*
	(0.31)	(0.013)	(0.13)	(0.096)
(log) PPP GDP per capita	-0.10***	0.0035	-0.023**	-0.028***
	(0.023)	(0.0026)	(0.011)	(0.0090)
unemployment rate	-0.16	-0.065***	-0.18***	-0.18***
	(0.17)	(0.015)	(0.062)	(0.058)
rule of law	-0.012	0.00024	-0.017*	-0.018
	(0.027)	(0.0021)	(0.0091)	(0.011)
control of corruption	0.026	-0.0024	0.0043	0.0053
	(0.023)	(0.0022)	(0.0081)	(0.0097)
government effectivenss	-0.057**	0.0046**	0.0082	-0.0072
	(0.025)	(0.0021)	(0.0085)	(0.010)
political stability	0.0051	-0.00078	0.0072*	0.019***
	(0.012)	(0.00086)	(0.0041)	(0.0065)
regulatory quality	0.0015	-0.00073	-0.0073	0.019**
	(0.021)	(0.0020)	(0.010)	(0.0092)
voice and accountability	0.012	0.00068	0.018*	0.010
	(0.018)	(0.0018)	(0.011)	(0.0075)
Obs	2985	2985	2985	2985
Adjusted R^2	0.043	0.058	0.15	0.098

Notes. This table presents the projection of the indicators of the informal economy on its causes. All indicators are in growth rates (first difference in log levels). All regressions include country fixed effects. Standard errors are in parentheses and clustered at the country level. $*p < 0.10, *p < 0.05, ***p < 0.01$.

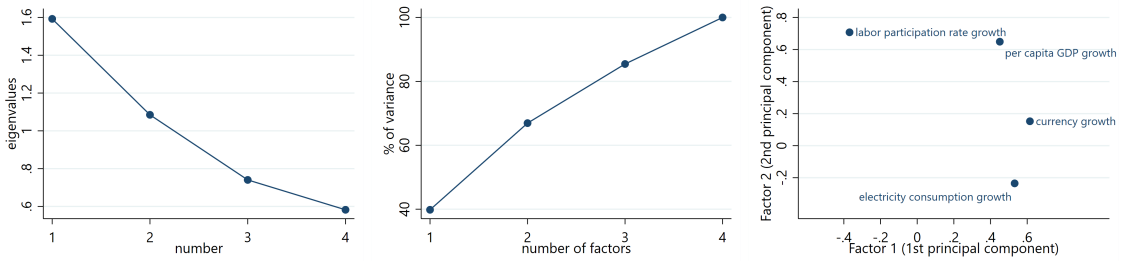
Table 4. Relationship between Estimated Factors and Indicators

	currency growth	labor participation rate growth	per capita GDP growth	electricity consumption growth
loadings				
Factor 1 (1st principal component)	0.61	-0.37	0.45	0.53
Factor 2 (2nd principal component)	0.15	0.71	0.65	-0.24
correlation				
Factor 1	0.78	-0.47	0.57	0.67
Factor 2	0.16	0.74	0.68	-0.25

Notes. This table presents the correlation between the first two principal components and the indicators of the informal economy.

pal component has higher loadings on labor participation rate growth and GDP per capita growth. Since it is orthogonal to the first principal component by construction, it can be viewed as summarizing the co-movement of the formal and the informal economy. Table 4 presents the loadings of the two factors as well as their correlation with the indicators. Consistent with the magnitude of the loadings, the first factor has the highest correlation with currency in circulation and the second factor with labor participation rate.

Figure 2. Projected Principal Component Analysis



(a) eigenvalues **(b) cum. pct. of variation explained** **(c) screeplot**

Notes. This figure presents the keys results of the projected component analysis.

In the fourth step, following equation (4), we construct two indices, each as the cumulative sum of their corresponding factor. We then use the indices to predict the degree of informality as in equation (5). For the dependent variables in Table 5, we include both the adjusted and unadjusted measure of informality in the WBES as well as four other measures of labor-related informality from Ohnsorge and Yu (2022): the share of self-employment in total employment (SEMP), the share of population that does not contribute to any pension scheme (Pension), the share of informal employment (Infemp); and the share of employment outside the formal sector (Infsize).

Interestingly, the first index, which corresponds to the first factor, does not explain the degree of informality in a statistically significant way in all survey data. By contrast, the coefficients before the second index are statistically significant in columns (1), (2), (3) and (6). Recall that the first factor is the latent variable under the assumptions of the MIMIC model. This suggests that the MIMIC model only captures overall economic activity–of which the informal economy is an integral part–but misses the co-movement of the formal and the informal economy, which is an important channel statistically and economically.

Note that columns (1) and (2) have relatively low R^2 compared to columns (3)-(4). This is because the WBES typically has no more than three observations per country, as shown in Table 1, and the observations vary widely in different waves, suggesting possibly large measurement errors. In contrast, employment-based informality measures in columns (3)-(4) have consecutive measurements for each country. While the measurements differ across countries, they are close to each other for the same country in different years. As such country fixed effects are able to absorb the cross-country differences and increase R^2 substantially.

Table 5. Survey Data and Estimated Factors

	(1) WBES	(2) WBES (unadjusted)	(3) SEMP	(4) Pension	(5) Infemp	(6) Infsize
Index 1 (cum. sum. of Factor 1)	0.097	0.092	-0.017	0.13	-0.24	-0.094
	(0.47)	(0.41)	(0.047)	(0.33)	(0.21)	(0.29)
Index 2 (cum. sum. of Factor 2)	-0.85***	-0.51*	-0.11**	-0.23	-0.42	-0.80***
	(0.30)	(0.27)	(0.056)	(0.56)	(0.29)	(0.29)
Obs	209	209	1508	150	313	300
Adjusted R^2	0.16	0.18	0.96	0.93	0.94	0.92

Notes. This table presents the results from the regressions of the degree of informality in survey data on the two indices constructed from the projected principal component analysis. WBES indicates the degree of informality that is adjusted by firm size. WBES (unadjusted) is the degree of informality measured as the fraction of firms claiming competition with unregistered firms. The rest survey data are from the informal economy database of the World Bank (Ohnsorge and Yu, 2022): SEMP is the share of self-employment in total employment; Pension the share of population that does not contribute to any pension scheme; Infemp informal employment; and Infsize employment outside the formal sector. Standard errors are in parentheses and clustered at the country level. $*p < 0.10, *p < 0.05, ***p < 0.01$.

Table 6. Estimated Weights on Projected Indicators

	currency in circulation	labor participation rate	GDP per capita	electricity consumption
weight	-0.04	-0.30	-0.25	0.11

Notes. The estimated degree of informality is a linear combination of the projected indicators (equation (6)). This table presents the weight (coefficient) on each projected indicator.

Equation (6) shows that the estimated degree of informality is a linear combination of the projected indicators. Table 6 presents the weights on the projected indicators. Higher currency in circulation, higher labor participation rate, and higher GDP per capita are associated with lower informality, while stronger electricity consumption reflects higher informality. The weights on labor participation rate and GDP per capita are higher. This is expected because the first principal component, which explains most of the variation in the data, has higher loadings on them.

B. Estimates of the Degree of Informality: Country Examples

Predictions of the augmented factor model

The predictions of the augmented factor model (equation (5)) will be our estimates of the degree of informality. With such estimates, we highlight a few stylized patterns of the evolution of the informal economy and explore their correlation with the formal economy.

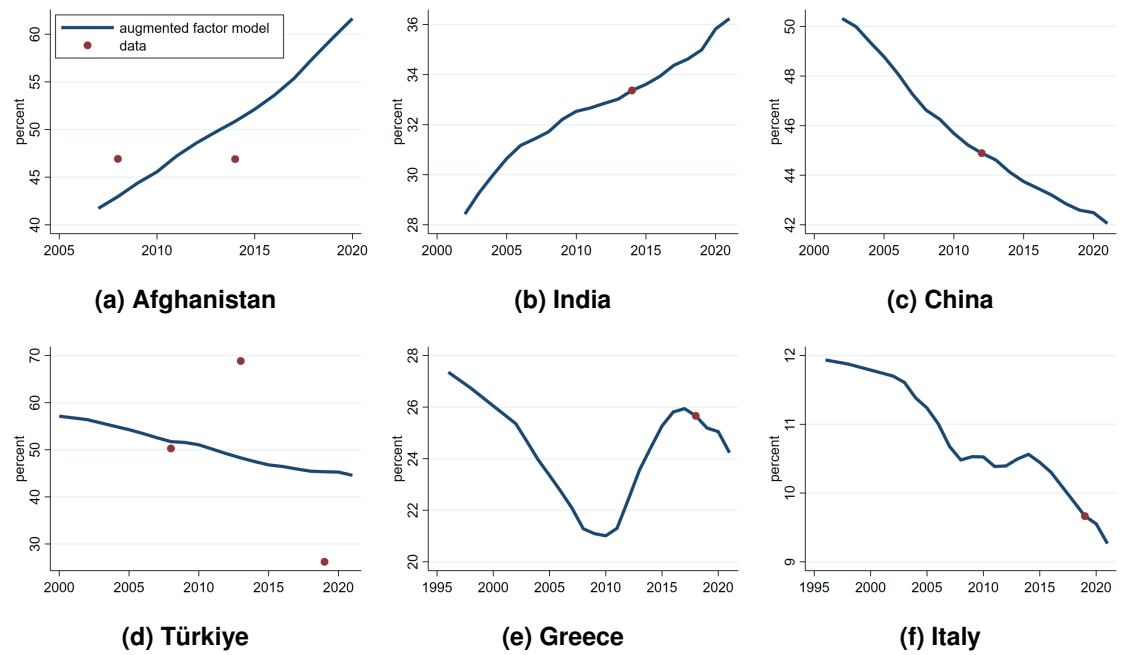
Figure 3 presents three broad patterns of the evolution of the informal economy with six country examples. For Afghanistan and India, the degree of informality has been increas-

ing in the past two decades. For China and Türkiye, it is decreasing. For Greece and Italy, it shows a strong cyclical behavior.

We also place the imputed degree of informality from the WBES along the predictions by the augmented factor model. Such survey data points discipline the predictions by the augmented factor model: for countries with only one wave of the WBES, such as India and China, the survey data point sets the model prediction at the same level as the data in that year; for countries with more than one waves, such as Afghanistan, the survey data points set the model average in these years as the average of these data points.

It is worth noting that there can be sizable measurement errors in the survey data. In addition, as the indices of the augmented factor model are based on macro data, they may not necessarily agree with the survey data. The model predictions can therefore deviate from the survey data substantially. This is evident in the case of Türkiye.

Figure 3. Degree of Informality: Selected Countries



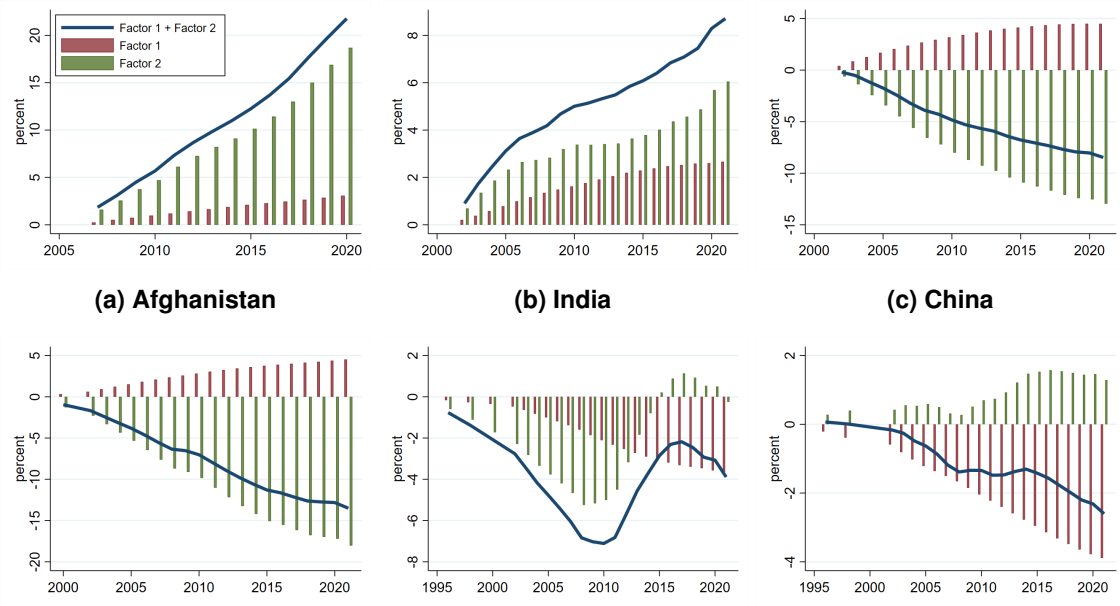
Notes. This figure presents the estimated degree of informality from the augmented factor model as well as data from the World Bank Enterprise Surveys. The degree of informality refers to the fraction of registered firms, adjusted by firm size, claiming that they compete with unregistered firms.

Decomposition by Indices

Equation (5) allows us to decompose the predictions of the augmented factor model by the contribution of each index. Figure 4 presents such decomposition. For Afghanistan, In-

dia, China, and Türkiye), the second index is quantitatively more important in explaining the overall dynamics of the informal economy. For Greece and Italy, the second index is responsible for the cyclical movements of the informal economy. As the second index is related to the co-movement of formal and informal economic activity, this suggests that the transition between the two types of activity could play a bigger role than overall economic activity in shaping the dynamics of the informal economy.

Figure 4. Decomposition of the Degree of Informality by Contribution of Factors: Selected Countries



Notes. This figure presents the predicted degree of informality from the projected principal component analysis as well as the contributions from the first two principal components.

Decomposition by Projected Indicators

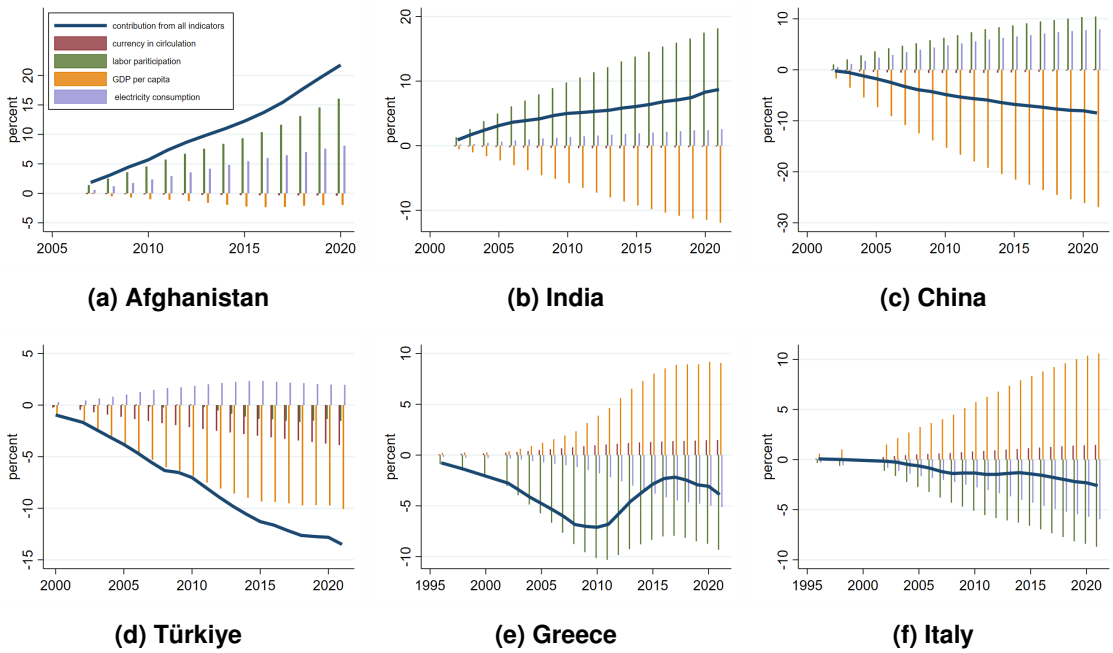
Since the estimate of each factor is a principal component—which is in essence a linear combination—of the projected indicators, it allows us to decompose the estimated degree of informality by the contribution of each indicator (equation (6)). This allows us to examine which projected indicator reflect the dynamics of the informal economy.

Recall that each indicator y_{pt} is a growth rate that is standardized to have a mean of zero and a variance of one. Its cumulative sum: $y^c_{pt} = \sum_{t'=1}^{t} y_{pt'}$ is a level index. Coupled with the coefficients in Table 6, each indicator's contribution to the dynamics of the informal economy can be calculated.

Figure 5 presents the results. For Afghanistan and India, labor participation rate is the leading indicator for the rise in the informal economy. Other indicators influence the dynamics of the informal economy to a lesser extent. For Afghanistan, the increase in electricity consumption also suggests an expansion of the informal economy. For India, the growing GDP per capita suggests a decline in the informal economy, but this indicator is not enough to offset the signal from the declining labor participation rate.

For China and Türkiye, the dominant indicator is per capita GDP growth: strong growth of the formal sector indicates a decline in the informal economy. For Greece, in the late 2000s before the Global Financial Crisis, the rapid increase in labor participation rate suggests a fall in the degree of informality; in the early 2010s during the European debt crisis, the declines in labor participation rate and in official GDP growth both indicate a rise in the informal economy. For Italy, weak GDP growth and rising labor participation rate point to different directions of the informal economy and they largely cancel out each other.

Figure 5. Decomposition of the Degree of Informality by Contribution of Indicators: Selected Countries



Notes. This figure presents the predicted degree of informality from the projected principal component analysis as well as the contributions from the four indicators.

C. Informal Economy: Causes and Patterns

Now that we have obtained estimates of the degree of informality by the augmented factor model approach, we can revisit the causes of the informal economy.

Table 7 shows the results from the regressions of the estimated degree of informality on its causes. Columns (1), (2), and (3) examine the roles of government size, economic development status, and governance, respectively, while column (4) includes all the causes.

Column (1) shows that tax revenue and government consumption do not affect the degree of informality in a statistically significant way. This is perhaps surprising, as much of the literature assumes that tax compliance is the main reason for the existence of the informal economy. However, one needs to distinguish between different definitions of the informal economy. In some studies, such as Orsi, Raggi, and Turino (2014), it is about the underground economy that exists for tax evasion purposes. In this paper, the informal economy is about productive activities that are unregistered. The avoidance of registration has more to do with the cost, time and effort of formal registration and less to do with the size of the government.

Column (2) shows that better performance of the formal economy, as characterized by higher GDP per capita, higher trade openness, and lower unemployment rate, tends to reduce the informal economy. This is consistent with the common view that as an economy develops, the level of informality tends to decrease.

Column (3) shows that governance matters for the degree of informality. The rule of law and regulatory quality are important for reducing informality, whereas a high degree of voice and accountability[9] could lead to high informality. Intuitively, a more free society could lead to more dynamism of the economy, which could boost the informal economy.

Notice that the adjusted R square in Table 7 is high across all columns. This is because to obtain the estimates of the degree of informality, we first use an augmented factor model to extract factors connecting the causes to the indicators, and then use the factors to predict the survey data. The causes therefore explain most of the variation in the predicted degree of informality by construction.

[9] According to the definition by the World Bank, voice and accountability captures perceptions of the extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association, and a free media.

Table 7. Estimated Informal Economy and Causes

	(1)	(2)	(3)	(4)
	degree of informality			
tax-GDP ratio	-1.75			7.37
	(5.12)			(5.68)
government consumption-GDP ratio	0.93			1.32
	(2.58)			(2.89)
(log) PPP GDP per capita		-3.14***		-3.09***
		(0.83)		(0.94)
trade openness		-2.39***		-2.61***
		(0.73)		(0.67)
unemployment rate		0.25***		0.23***
		(0.053)		(0.050)
rule of law			-2.11*	-1.53
			(1.09)	(0.94)
control of corruption			0.74	0.55
			(0.84)	(0.73)
government effectivenss			-1.40*	-1.02
			(0.75)	(0.65)
political stability			0.43	0.28
			(0.50)	(0.53)
regulatory quality			-1.26*	-0.066
			(0.72)	(0.65)
voice and accountability			2.15**	1.83**
			(0.83)	(0.73)
Obs	2498	2498	2498	2498
Adjusted R^2	0.94	0.96	0.95	0.96

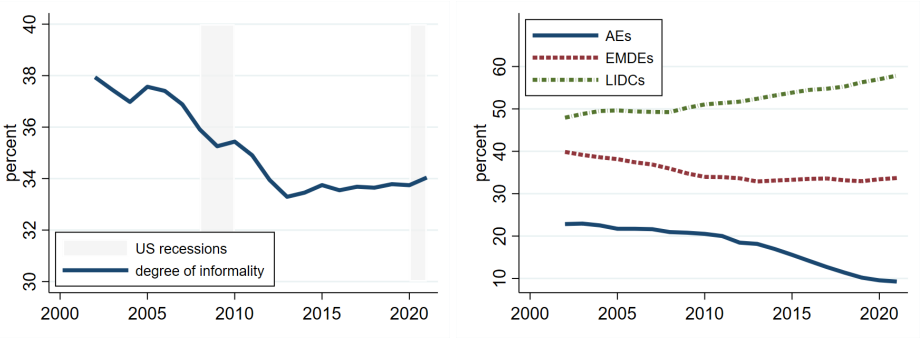
Notes. This table presents the results from the regressions of the estimated informal economy on its causes. All regressions include country fixed effects. Standard errors are in parentheses and clustered at the country level. $*p < 0.10, *p < 0.05, ***p < 0.01$.

To examine how the informal economy has evolved over the past decades, we focus on a subset of 104 countries that have estimates between 2002 and 2021, including 21 advanced economies (AEs), 50 emerging markets and developing economies (EMDEs), and 33 low-income and developing countries (LIDCs). Table 10 in Appendix E presents the details of the country list.

In Figure 6, panel (a) shows that the median degree of informality declined between 2002 and 2013 but stalled afterwards. Notably, during the global financial crisis in 2008-2010, the decline slowed down, and during the COVID-19 pandemic, the informal economy increased slightly. The stagnant informality after 2013 could be related to low global GDP growth, but we leave this question for future research.

Panel (b) shows that countries of different income status differ drastically in terms of the degree of informality. Interestingly, while informality has declined for AEs and EMDEs,

Figure 6. Median Degree of Informality

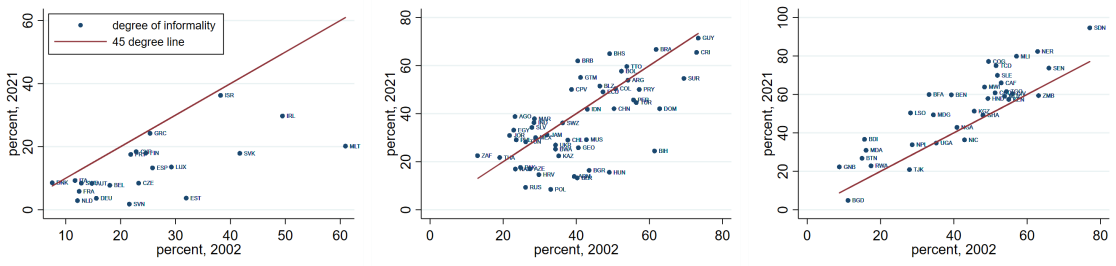


(a) all countries **(b) by country groups**

Notes. This figure presents the patterns of the informal economy in 104 countries. Panel (a) shows the median degree of informality over time. Panel (b) compares the median degree of informality by country groups in 2002 and 2021. AEs = advanced economies; EMDEs = emerging markets and developing economies; LIDC = low income and developing countries. Country classifications are based on the World Economic Outlook, and the three groups are mutually exclusive.

it has actually increased for LIDCs. In fact, as shown in Figure 7, for most LIDCs, the degree of informality is larger in 2021 compared to two decades ago. This is in contrast to previous studies using the MIMIC model, such as Medina and Schneider (2018), that generally find declining informality across all country groups.

Figure 7. Degree of Informality: 2002 vs. 2021



(a) AEs **(b) EMDEs** **(c) LIDCs**

Notes. This figure presents the contrast of the degree of informality between 2002 and 2021 for different country groups. Dots below the 45 degree line indicate a smaller degree of informality in 2021 than in 2002. AEs = advanced economies; EMDEs = emerging markets and developing economies; LIDC = low income and developing countries. Country classifications are based on the World Economic Outlook, and the three groups are mutually exclusive.

VI. Conclusion

In this paper, we develop an augmented factor model approach to estimating the degree of informality. It combines direct measures in survey data with estimated factors from a factor model of indicators of the informal economy augmented by its causes. We show that

the prevailing MIMIC model used in the literature is a special case of the augmented factor model under strong assumptions. While the augmented factor model shows that the dynamics of the informal economy is shaped by the strength of overall economic activity as well as the interplay between the formal and informal economies, the MIMIC model only captures the former. We find that economic development status and governance matter for the dynamics of the informal economy. Over the past two decades, advanced economies and emerging markets have seen the degree of informality steadily declining, however, low-income and developing countries have experienced the opposite trend.

REFERENCES

Adrian, Tobias, Richard K Crump, and Erik Vogt, 2019, "Nonlinearity and flight-to-safety in the risk-return trade-off for stocks and bonds," The Journal of Finance, Vol. 74, No. 4, pp. 1931–1973.

Artavanis, Nikolaos, Adair Morse, and Margarita Tsoutsoura, 2016, "Measuring income tax evasion using bank credit: Evidence from Greece," The Quarterly Journal of Economics, Vol. 131, No. 2, pp. 739–798.

Beyer, Robert, Yingyao Hu, and Jiaxiong Yao, 2022, "Measuring quarterly economic growth from outer space," IMF Working Paper Working Paper No. 2022/109.

Braguinsky, Serguey, Sergey Mityakov, and Andrey Liscovich, 2014, "Direct estimation of hidden earnings: Evidence from Russian administrative data," The Journal of Law and Economics, Vol. 57, No. 2, pp. 281–319.

Cagan, Phillip, 1958, "The demand for currency relative to the total money supply," Journal of political economy, Vol. 66, No. 4, pp. 303–328.

Capasso, Salvatore, and Tullio Jappelli, 2013, "Financial development and the underground economy," Journal of Development Economics, Vol. 101, pp. 167–178.

Chen, Martha Alter, 2012, The informal economy: Definitions, theories and policies (WIEGO Manchester).

Choi, Jay Pil, and Marcel Thum, 2005, "Corruption and the shadow economy," International Economic Review, Vol. 46, No. 3, pp. 817–836.

Del Boca, Daniela, and Francesco Forte, 1982, "Recent empirical surveys and theoretical interpretations of the parallel economy in Italy," The underground economy in the United States and abroad, Lexington (Mass.), Lexington, pp. 160–178.

Dell'Anno, Roberto, 2007, "The shadow economy in Portugal: An analysis with the MIMIC approach," Journal of Applied Economics, Vol. 10, No. 2, pp. 253–277.

———, 2022, "Theories and definitions of the informal economy: A survey," Journal of Economic Surveys, Vol. 36, No. 5, pp. 1610–1643.

Elgin, Ceyhun, and Oguz Oztunali, 2012, "Shadow Economies around the World: Model Based Estimates," Techn. rep., Bogazici University, Department of Economics.

Elgin, Ceyhun, Friedrich Schneider, and others, 2016, "Shadow economies in OECD countries: DGE vs. MIMIC approaches," Bogazici Journal, Vol. 30, No. 1, pp. 51–75.

Fan, Jianqing, Yuan Ke, and Yuan Liao, 2021, "Augmented factor models with applications to validating market risk factors and forecasting bond risk premia," Journal of Econometrics, Vol. 222, No. 1, pp. 269–294.

Feige, Edgar L, 2016, "Reflections on the Meaning and Measurement of Unobserved Economies: What Do We Really Know About the'Shadow Economy'," Journal of Tax Administration (2016) Vol, Vol. 2.

Frey, Bruno S, and Hannelore Weck-Hanneman, 1984, "The hidden economy as an âĂŸunobservedâĂŹvariable," European economic review, Vol. 26, No. 1-2, pp. 33–53.

Giles, David EA, 1999, "Measuring the hidden economy: Implications for econometric modelling," The Economic Journal, Vol. 109, No. 456, pp. 370–380.

Gorodnichenko, Yuriy, Jorge Martinez-Vazquez, and Klara Sabirianova Peter, 2009, "Myth and reality of flat tax reform: Micro estimates of tax evasion response and welfare effects in Russia," Journal of Political economy, Vol. 117, No. 3, pp. 504–554.

Hu, Yingyao, and Jiaxiong Yao, 2022, "Illuminating economic growth," Journal of Econometrics, Vol. 228, No. 2, pp. 359–378.

Ihrig, Jane, and Karine S Moe, 2004, "Lurking in the shadows: the informal sector and government policy," Journal of Development Economics, Vol. 73, No. 2, pp. 541–557.

Isachsen, Arne Jon, and Steiner Strøm, 1985, "The size and growth of the hidden economy in Norway," Review of Income and Wealth, Vol. 31, No. 1, pp. 21–38.

Jöreskog, Karl G, and Arthur S Goldberger, 1975, "Estimation of a model with multiple indicators and multiple causes of a single latent variable," Journal of the American statistical Association, Vol. 70, No. 351a, pp. 631–639.

Kaufmann, Daniel, and Aleksander Kaliberda, 1996, "Integrating the unofficial economy into the dynamics of post socialist economies: A framework of analyses and evidence," Economic transition in Russia and the new states of Eurasia, Vol. 117.

Lam, Clifford, and Qiwei Yao, 2012, "Factor modeling for high-dimensional time series: inference for the number of factors," The Annals of Statistics, pp. 694–726.

Medina, Leandro, and Friedrich Schneider, 2017, "Shadow economies around the world: New results for 158 countries over 1991-2015," CESifo Working Paper Series.

———, 2018, "Shadow Economies Around the World: What Did We Learn Over the Last 20 Years?" IMF Working Papers, Vol. 2018, No. 017.

———, 2019, "Shedding Light on the Shadow Economy: A Global Database and the Interaction with the Official One," Techn. rep., CESifo.

OECD, Paris, 2002, "Measuring the non-observed economy: A handbook," OECD.

Ohnsorge, Franziska, and Shu Yu, 2022, The long shadow of informality: Challenges and policies (World Bank).

Orsi, Renzo, Davide Raggi, and Francesco Turino, 2014, "Size, trend, and policy implications of the underground economy," Review of Economic Dynamics, Vol. 17, No. 3, pp. 417–436.

Schneider, Friedrich, 1986, "Estimating the size of the Danish shadow economy using the currency demand approach: An attempt," The Scandinavian Journal of Economics, pp. 643–668.

Schneider, Friedrich, and Andreas Buehn, 2017, "Estimating a shadow economy: Results, methods, problems, and open questions," Open Economics, Vol. 1, No. 1, pp. 1–29.

Schneider, Friedrich, and Dominik H Enste, 2000, "Shadow economies: Size, causes, and consequences," Journal of economic literature, Vol. 38, No. 1, pp. 77–114.

Tanzi, Vito, 1983, "The underground economy in the United States: Annual estimates, 1930-80," Staff Papers-International Monetary Fund, pp. 283–305.

Van Eck, Robert, and Brugt Kazemier, 1988, "Features of the Hidden Economy in the Netherlands," Review of Income and Wealth, Vol. 34, No. 3, pp. 251–273.

Vuletin, Guillermo, 2008, "Measuring the informal economy in Latin America and the Caribbean," IMF working paper No. 08/102.

Waseem, Mazhar, 2023, "Overclaimed refunds, undeclared sales, and invoice mills: Nature and extent of noncompliance in a value-added tax," Journal of Public Economics, Vol. 218, p. 104783.

APPENDIX A. THE MIMIC MODEL

The Multiple Indicators Multiple Causes (MIMIC) model is the prevailing modeling approach in the literature to estimate the size of the informal economy. It links multiple observable indicators of the informal economy to multiple observable causes of the informal economy through a latent variable. The latent variable is an index of the informal economy that can be used to calculate the size of the informal economy through variable transformation and calibration.

The MIMIC model consists of a structural equation and a measurement equation. Let y_t^* be the scalar latent index of the informal economy, which is assumed to be determined by a $q \times 1$ vector of causes $\boldsymbol{x}_t = (x_{1t}, \cdots, x_{qt})'$ through a linear structural equation:

$$y_t^* = \boldsymbol{\alpha}' \boldsymbol{x}_t + v_t, \tag{7}$$

where v_t is a scalar structural disturbance that captures the component of the informal economy not explained by the causes $\boldsymbol{y}_t$. Let $\boldsymbol{y}_t = (y_{1t}, , y_{Pt})'$ be a $P \times 1$ vector of linear indicators of the latent index of the informal economy. The measurement model follows:

$$\boldsymbol{y}_t = \boldsymbol{\beta} y_t^* + \boldsymbol{u}_t. \tag{8}$$

The disturbances are assumed to be mutually independent:

$$E(v_t \boldsymbol{u}_t') = \boldsymbol{0}', E(v_t^2) = \sigma^2, E(\boldsymbol{u_t} \boldsymbol{u}_t') = \Theta^2, \tag{9}$$

where Θ^2 is a diagonal matrix. The reduced-form equation of the MIMIC model is then:

$$\boldsymbol{y}_t = \boldsymbol{\beta} \boldsymbol{\alpha}' \boldsymbol{x}_t + (\boldsymbol{\beta} v_t + \boldsymbol{u}_t). \tag{10}$$

In essence, the MIMIC model is therefore a regression equation of $\boldsymbol{y}_t$ on $\boldsymbol{x}_t$ with two restrictions. First, the coefficient matrix before $\boldsymbol{x}_t$, i.e., $\Pi = \boldsymbol{\beta} \boldsymbol{\alpha}'$, has rank one. Second, the covariance matrix of the error term is the sum of a rank-one matrix and a diagonal matrix, $\Omega = E[(\boldsymbol{\beta} v_t + \boldsymbol{\varepsilon}_t)(\boldsymbol{\beta} v_t + \boldsymbol{\varepsilon}_t)] = \sigma^2 \boldsymbol{\beta} \boldsymbol{\beta}' + \Theta^2$. Note that if $\boldsymbol{\alpha}$ and σ are multiplied by a scalar and $\boldsymbol{\beta}$ is divided by the same scalar, the reduced-form equation remains unchanged. A normalization is therefore needed in order to pin down $\boldsymbol{\alpha}$ and y^*. In practice, the literature typically assumes that the first indicator has the same unit as y^*. In other words,

$$y_{1t} = y_t^* + v_{1t}. \tag{11}$$

The MIMIC model can be estimated by the maximum-likelihood estimation (see, for example, Jöreskog and Goldberger (1975)).

As a comparison, the augmented factor model generalizes the single latent index y^* in the MIMIC model to be multiple factors. It allows for a nonlinear relationship between the factors and the causes in equation (7). It also does not impose any structure on the covariance matrix Θ^2.

APPENDIX B. ESTIMATION OF THE AUGMENTED FACTOR MODEL

In this section, we describe the basic steps for estimating the augmented factor model.

First, all indicators $\boldsymbol{y}_t$ are transformed into growth rates, i.e., the first difference of log levels.

Second, we regress $\{\boldsymbol{y}_t\}$ on $\{\boldsymbol{x}_t\}$ and obtain fitted value $\{\hat{\boldsymbol{y}}_t\}$. $\{\hat{\boldsymbol{y}}_t\}$ is an estimator of the projected indicators $E(\boldsymbol{y}_t|\boldsymbol{x}_t)$. Here $\{\boldsymbol{x}_t\}$ include country fixed effects to take into account time-invariant reasons for informal economic activity, such as cultural norms. $\{\hat{\boldsymbol{y}}_t\}$ are then standardized to have a mean of zero and variance of one. We now continue to use $\{\hat{\boldsymbol{y}}_t\}$ to denote the standardized projected indicators.

Third, factor loadings $\boldsymbol{\Lambda}$ and the explained components of the factors $g(\boldsymbol{x}_t)$ are estimated through principal component analysis of $\{\hat{\boldsymbol{y}}_t\}$. Specifically, let $\Sigma_{y|x} = E(E(\boldsymbol{y}_t|\boldsymbol{x}_t)E(\boldsymbol{y}_t|\boldsymbol{x}_t)')$ be the variance-covariance matrix of the projected indicators. It can be estimated by $\hat{\Sigma}_{y|x} = \frac{1}{T}\sum_t(\hat{\boldsymbol{y}}_t\hat{\boldsymbol{y}}_t')$. The columns of $\hat{\boldsymbol{\Lambda}}$ are the eigenvectors of the first K eigenvalues of $\hat{\Sigma}_{y|x}$, ranked from high to low. $\hat{g}(\boldsymbol{x}_t) = \hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{y}}_t$.

Fourth, the indices of the informal economy, defined as the cumulative sum of the explained components of the factors, are calculated as follows:

$$\hat{s}_{kt} = \sum_{t'=1}^{t} \hat{g}_{kt'}, \qquad \text{for} \quad k = 1, \cdots, K$$

where $\hat{g}_{kt} = \hat{g}_k(\boldsymbol{x}_t)$ is the kth element of $\hat{g}(\boldsymbol{x}_t)$.

Fifth, we regress survey data $\{z_t\}$ on the indices of the informal economy, including country fixed effects.

$$z_t = \boldsymbol{\beta}'\hat{\boldsymbol{s}}_t + \zeta_0 + \varepsilon_t,$$

where ζ_0 is a country-specific intercept.

The predictions of such regression equation, $\hat{z}_t = \hat{\boldsymbol{\beta}}' \hat{\boldsymbol{s}}_t + \hat{\zeta}_0$, are the estimates of the size of the informal economy.

Note that we have omitted country index throughout for ease of notation. It is straightforward to add an additional subscript i to all the variables above, indicating it is for country i. When calculating $\hat{\Sigma}_{y|x}$, the average should also be taken over all countries in addition to over time.

APPENDIX C. RELATIONSHIP BETWEEN WBES AND OTHER SURVEY DATA

Ohnsorge and Yu (2022) provide a database based on various labor force surveys. It has four measures of labor-related informality, including the share of unemployment in total employment (SEMP, 1990-2018), informal employment (Infemp, 2000-2018), employment outside the formal sector (Infsize, 1999-2018), as well as share of the labor force that contributes to a retirement pension scheme. We transform the last measure to be the share of the labor force that does not contribute to a retirement pension (Pension, 1990-2010), which then becomes a measure that is increasing in the degree of informality.

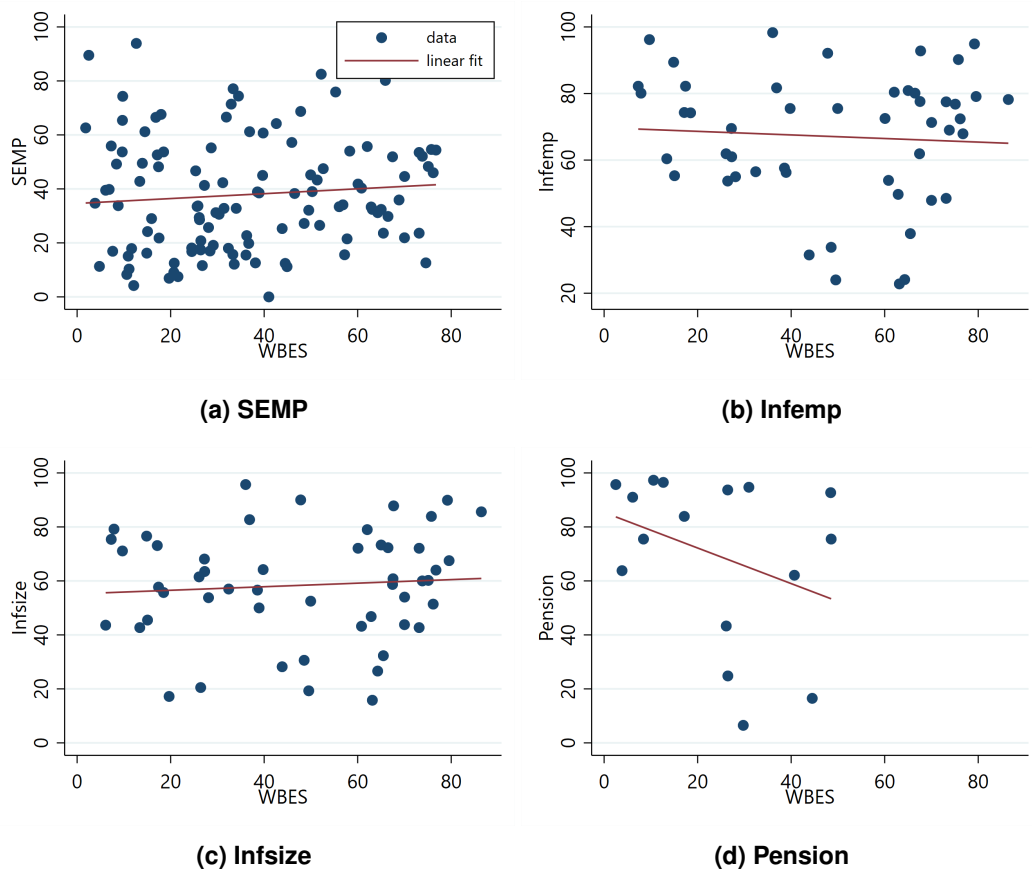
Figure 8 presents the scatter plots of each measure of labor-related informality against the degree of informality from the WBES. It is clear that the degree of informality from the WBES is only weakly correlated with the labor-related measures of informality, suggesting that they could reflect different concepts of the informal economy. However, as the augmented factor approach is designed to fit the data, the estimated degree of informality will be closely aligned with the survey data in use.

APPENDIX D. ADDING NIGHTTIME LIGHTS

The literature has also considered alternative measures of economic activity, such as satellite-recorded nighttime lights (Medina and Schneider, 2018), as indicators of the informal economy. In this section, we include nighttime light growth as another indicator of the informal economic activity.

We use data from Hu and Yao (2022), and Beyer, Hu, and Yao (2022). Because nighttime light data are derived from different satellite systems before and after 2013, we unify them using separately estimated elasticities. In particular, we divide nighttime light growth by

Figure 8. Degree of Informality in Survey Data: WBES vs. Labor Surveys

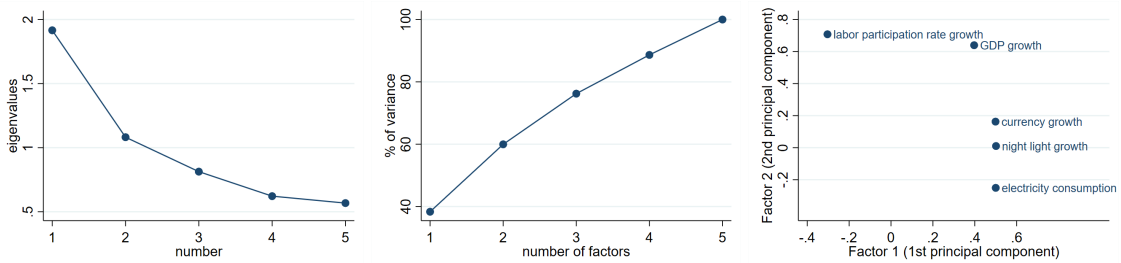


Notes. This figure contrasts the degree of informality in the World Bank Enterprise Surveys (WBES) against four measures of informality in labor-related surveys. SEMP indicates the share of unemployment in total employment, Infemp informal employment, Infsize employment outside the formal sector, and Pension the share of the labor force that does not contribute to a retirement pension.

1.3 (Hu and Yao, 2022) before 2013 and by 1.55 (Beyer, Hu, and Yao, 2022) to obtain a unified measure.

Figure 9 shows that with nighttime lights added, it is still optimal to choose two principal components. Panel (c) shows that the second factor's loading on nighttime light is almost zero. However, because the second factor is more important in predicting the degree of informality, as shown in Table 8, the estimated degree of informality with nighttime lights will not be much different from that without nighttime light. Figure 10 confirms that the estimates of the degree of informality with and without nighttime light are well aligned.

Figure 9. Projected Principal Component Analysis with Nighttime Light



(a) eigenvalues **(b) cum. pct. of variation explained** **(c) screeplot**

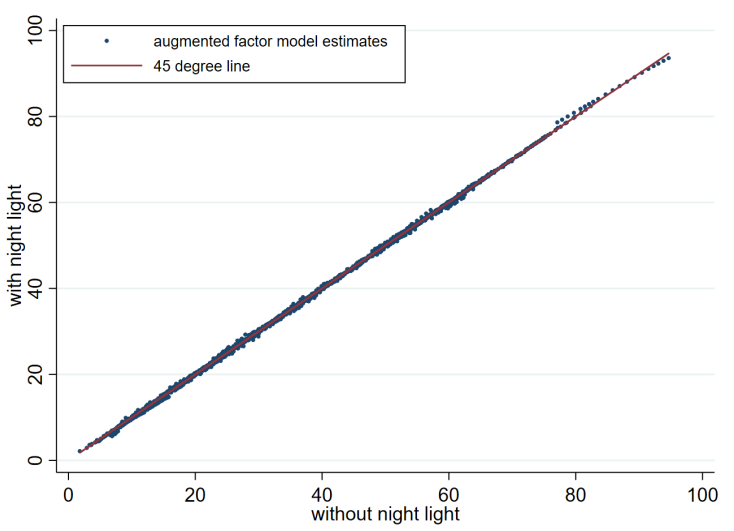
Notes. This figure presents the keys results of the projected component analysis with five indicators, including nighttime light.

Table 8. Survey Data and Estimated Factors with Night Lights

	(1) WBES	(2) WBES (unadjusted)	(3) SEMP	(4) Pension	(5) Infemp	(6) Infsize
Index 1 (cum. sum. of Factor 1)	0.066	0.077	-0.023	0.12	-0.26	0.062
	(0.46)	(0.41)	(0.045)	(0.35)	(0.21)	(0.32)
Index 2 (cum. sum. of Factor 2)	-0.83***	-0.50*	-0.12**	-0.28	-0.40	-0.80***
	(0.30)	(0.28)	(0.056)	(0.57)	(0.28)	(0.29)
Obs	209	209	1508	150	313	300
Adjusted R^2	0.16	0.18	0.96	0.93	0.94	0.92

Notes. This table presents the results from the regressions of the degree of informality in survey data on the two indices constructed from the projected principal component analysis with five indicators, including nighttime light. WBES indicates the degree of informality that is adjusted by firm size. WBES (unadjusted) is the degree of informality measured as the fraction of firms claiming competition with unregistered firms. The rest survey data are from the informal economy database of the World Bank (Ohnsorge and Yu, 2022): SEMP is the share of self-employment in total employment; Pension the share of population that does not contribute to any pension scheme; Infemp informal employment; and Infsize employment outside the formal sector. Standard errors are in parentheses and clustered at the country level. $*p < 0.10, *p < 0.05, ***p < 0.01$.

Figure 10. Estimated Degree of Informality with and without Nighttime Light



Notes. This figure contrasts the estimated degree of informality by the augmented factor model with and without nighttime light.

APPENDIX E. COUNTRY GROUPS

In this section, Table 9 presents the list of 126 countries for which all cause and indicator variables are available and there is at least one wave of World Bank Enterprise Survey (WBES). Table 10 presents the list of 104 countries for which all cause and indicator variables are available between 2002 and 2021 and there is at least one wave of WBES.

Table 9. List of Countries with Causes, Indicators, and WBES Data

AFG	BTN	EST	ITA	MUS	SLB	URY
AGO	BWA	FIN	JAM	MWI	SLE	UZB
ALB	CAF	FRA	JOR	NAM	SLV	VNM
ARG	CHL	GAB	KAZ	NER	SRB	VUT
ARM	CHN	GEO	KEN	NGA	SUR	ZAF
AUT	CIV	GHA	KGZ	NIC	SVK	ZMB
AZE	CMR	GIN	LBN	NLD	SVN	
BDI	COD	GMB	LSO	NPL	SWE	
BEL	COG	GNB	LUX	PAK	SWZ	
BEN	COL	GRC	LVA	PAN	TCD	
BFA	CPV	GTM	MAR	PER	TGO	
BGD	CRI	GUY	MDA	PHL	THA	
BGR	CYP	HND	MDG	POL	TJK	
BHS	CZE	HRV	MEX	PRT	TLS	
BIH	DEU	HUN	MLI	PRY	TTO	
BLR	DNK	IDN	MLT	ROU	TUN	
BLZ	DOM	IND	MMR	RUS	TUR	
BOL	ECU	IRL	MNE	RWA	TZA	
BRA	EGY	IRQ	MOZ	SDN	UGA	
BRB	ESP	ISR	MRT	SEN	UKR	

Notes. This table presents the list of countries for which all cause and indicator variables are available and there is at least one wave of World Bank Enterprise Survey (WBES).

Table 10. List of Countries in Each Income Group

group	AEs	EMDEs			LIDCs	
	AUT	AGO	GEO	SUR	BDI	MLI
	BEL	ARG	GTM	SWZ	BEN	MOZ
	CYP	ARM	GUY	THA	BFA	MWI
	CZE	AZE	HRV	TTO	BGD	NER
	DEU	BGR	HUN	TUN	BTN	NGA
	DNK	BHS	IDN	TUR	CAF	NIC
	ESP	BIH	IND	UKR	CIV	NPL
	EST	BLR	JAM	ZAF	COD	RWA
	FIN	BLZ	JOR		COG	SDN
	FRA	BOL	KAZ		GHA	SEN
	GRC	BRA	MAR		GNB	SLE
	IRL	BRB	MEX		HND	TCD
	ISR	BWA	MUS		KEN	TGO
	ITA	CHL	NAM		KGZ	TJK
	LUX	CHN	PAK		LSO	UGA
	MLT	COL	PER		MDA	ZMB
	NLD	CPV	PHL		MDG	
	PRT	CRI	POL			
	SVK	DOM	PRY			
	SVN	ECU	RUS			
	SWE	EGY	SLV			
# countries	21		50		33	

Notes. This table presents the list of countries used in Section V.C, using their ISO code. AEs = advanced economies; EMDEs = emerging markets and developing economies; LIDC = low income and developing countries.

PUBLICATIONS