Impact of COVID-19 shock on a segmented labour market: Analysis using a unique panel dataset

Das, Satadru and Ghosh, Saurabh and Mazumder, Debojyoti and Tushavera, Jitendra

Reserve Bank of India

March 2023

Impact of COVID-19 shock on a segmented labour market: Analysis using a unique panel dataset

Satadru Das[1a], Saurabh Ghosh[1], Debojyoti Mazumder[1] and Jitendra Tushavera

Abstract

This paper studies the impact of economic crisis caused by the COVID on the Indian labour market using the Periodic Labour Force Survey (PLFS). The unique dataset offers the opportunity to analyse sectoral transition and mobility of workers in response to a crisis due to its rotational panel framework. We employ transition matrices, non-parametric cumulative distribution functions, and machine learning techniques to identify the impact of COVID shock on formal and informal sector workers and whether this impact was heterogeneous. We find that labour market outcomes, both in terms of employment status and income, became even more divergent between the formal and informal sectors during the first wave of pandemic and remained divergent in the recovery phase. The classification analysis highlights that the sector in which the worker was employed (formal or informal sector), was an important predictor of income loss during the first wave.

JEL Classification: J31, J46, J62

Keywords: Segmented Labour Market, Informality, COVID scarring

---

Disclaimer: The opinions expressed in this paper are those of the authors and do not represent the Reserve Bank of India.

[a] Corresponding author. Email-id: satadrudas@rbi.org.in

[1] Strategic Research Unit, DEPR, Reserve Bank of India

1. Introduction

There is a strand of economic theory called labour market segmentation which argues that there exists systematic and persistent wage differentials between workers due to structural barriers to labour mobility. Such hypothesis is opposed to the neo-classical view of the labour market which assumes unrestricted mobility of labour, and wage differences between workers is explained by productivity differentials. According to the segmented labour market theory (SLM), the restrictions to mobility of workers are caused by regulatory, social, and geographic barriers[2], although some recent papers have identified that segmentation can be an intrinsic feature of labour market without any exogenous barriers.[3]

Several recent studies have analysed the disproportionate effect of scarring during recessions on a segmented labour market. Genda et al. (2010) compared the effect of the 2008 financial crisis on Japanese and US labour markets and found that less educated workers in Japan took relatively more time to recover from the recession's effects compared to their US counterparts because of segmentation in the Japanese labour market. Cockx and Ghirelli (2016) analysed the effects of recession on the Belgian labour market which is segmented because of labour regulations. They showed that earnings of workers in the primary market (permanent workers with high degree of labour regulations and social security) are affected by recession although they face comparatively less uncertainty of employment, while workers in the secondary market (contractual workers with less or no social security) do not face decrease in earnings but are exposed to higher risk of unemployment. A study on the labour market in Spain also demonstrated differential effect of recession due to segmentation (Fernandez-Kraz and Rodrigues-Planas, 2017). The present paper is related to this strand of literature as we analyse the effect of COVID scarring on the segmented labour market in India.

The labour market segmentation in India is usually defined within the context of formal and informal employment by the research literature. There are two alternative but inter-related definitions of the formal and the informal labour markets in India. The first definition is related to the nature of contract between the workers and firms, and the social security benefits workers are entitled to under such contracts. Workers hired under formal contracts and entitled to social security benefits are generally considered formal workers while those who do not have such contracts are considered part of informal workforce (Neog and Sahoo, 2016a and 2016b). The second definition is related to the enterprise ownership. Unincorporated enterprises owned by

[2] Refer to Leontaridi (2002) for an extensive survey of the evolution of SLM theory.

[3] For example, Altmann et al (2014) have argued that incomplete information during contract formation may also lead to segmentation in labour markets.

households are broadly classified as informal sector firms. Indian company laws require firms with excess of 10 workers (20 workers if the firm does not use electricity) to register with government and therefore, comes under the purview of a number of national and state-level labour laws. These smaller firms are not obligated to have formal contracts with workers or pay social security benefits. Such firms are considered part of the informal sector while bigger firms are part of the formal sector. Workers employed in these firms are considered informal sector workers.

Several papers have argued that the restrictive nature of Indian labour laws is responsible for the slow growth of productivity and low employment generation in the formal sector (e.g. Besley and Burgess, 2004; Kochhar et al, 2006; Hasan and Jandoc, 2012 etc.). Moreover, Hsieh and Klenow (2009) argued in their influential paper that the arbitrariness of implementing labour laws on firms with ten or more workers may create a barrier to optimization of firm size and lead to misallocation and lower growth of productivity. Amirapu and Grechter (2014) estimated that Indian labour laws increase the per-worker cost by 35% for firms which hire more than 10 workers compared to firms which hire lesser. They hypothesized that the additional cost labour laws for the formal sector firms is responsible for the existence of a large number of informal, micro and small firms in India.

The wage gap between formal and informal workers in India is well-documented in the literature (Mazoomdar 1975; Sastry 2004; Mehrotra 2012 and 2013; Neog and Sahoo, 2016a and 2016b). However, a wage gap may not imply segmented labour market as long as there are no barriers to worker mobility between the formal and the informal segments (Dickens and Lang, 1985; Pratap and Quintin, 2006). The arbitrariness of enforcement of labour laws based on firm size, as discussed above may create such a barrier. This may be true for India due to the following two reasons. First, informal firms face a disincentive to scale up and become formal sector firms due to the above-mentioned labour law and as a result most of the informal firms remain informal irrespective of their profitability; moreover, smaller informal firms may remain competitive even in the presence of large formal firms due to the lower per worker costs. Second, salaried formal sector workers may have work contracts which make it costly for firms to dismiss less productive workers and hire productive workers from the informal sector.

As discussed, the difference in labour laws is triggered by firms' characteristic. Therefore, in this paper we define the informal sector by enterprise type instead of the nature of job contract. We use a unique dataset, called the Periodic Labour Force Survey (PLFS) to analyse the effect of COVID scarring on both formal and informal sectors. The PLFS is a large, representative, quarterly survey initiated in 2017. One of the defining feature of the survey is that it consists of a rotational panel of urban workers i.e. a subsample of workers are interviewed in each quarter for four consecutive quarters. This feature allows us to track the mobility of each worker in each of the sector and how they were affected during COVID and their subsequent recovery from it. We create transition probabilities of employment and income distribution of workers in each sector. Then, we contrast these around the COVID period. Lastly, we employ machine learning classification techniques to identify the characteristics of the workers which determined their vulnerability to the COVID shock and the strength of subsequent recovery. Recent developments in big data techniques facilitate incorporating linear, non-linear relational relationships as dictated by the dataset. We refresh this classification mainly because of two reasons. First, adverse productivity shocks often help in identifying complex associations, which could be extremely policy relevant. Second, the availability of a pan-India survey based big panel data which is available pre and post pandemic periods, enables us to test the hypotheses using appropriate methodology and underline the empirical findings.

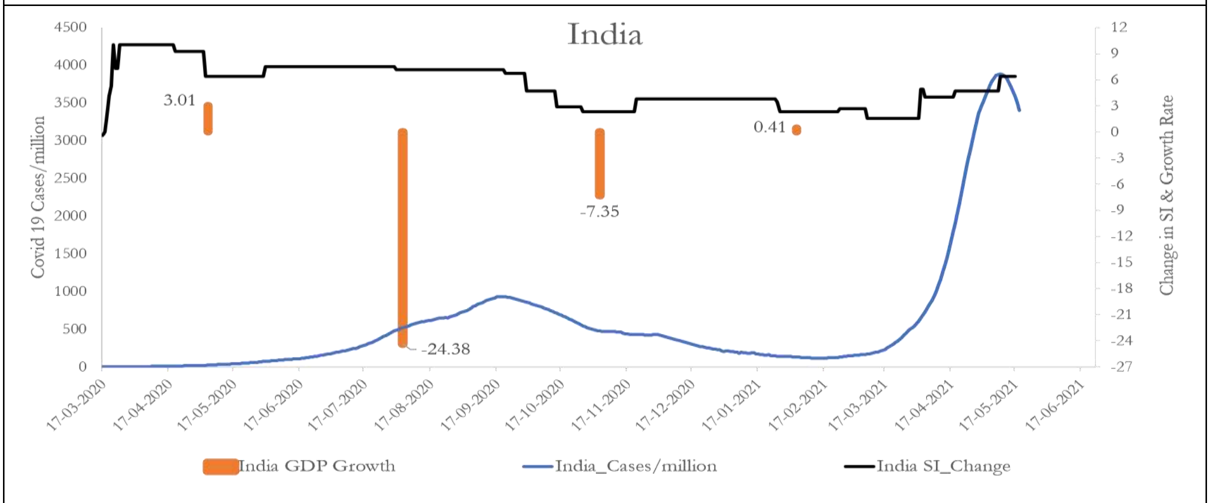
The analysis suggests the share of informal sector employment increase after the pandemic. Few industries-services creates the space for informal sector employment the most, in certain occupation categories. This trend seems to be stable even in the recovery phase after the first wave of COVID-19. Employing transition matrices and non-parametric cumulative distribution functions, it is shown that labour market outcomes, both in terms of employment status and income, became even more divergent between the formal and informal sectors during the first wave of pandemic and remained divergent in the recovery phase. We use classification technique to understand the characteristics of this divergences in the labour market outcomes. The analysis indicate that the sector in which the worker was employed (formal or informal sector), was an important predictor of income loss during COVID lockdown. This indicator was also an important predictor in the recovery but to a lesser degree. Moreover, characteristics like education, family education, social groups were also important predictors of income fall (increase) during COVID lockdown (subsequent recovery). These variables may also play crucial role in allocating workers into formal and the informal sectors in the first place.

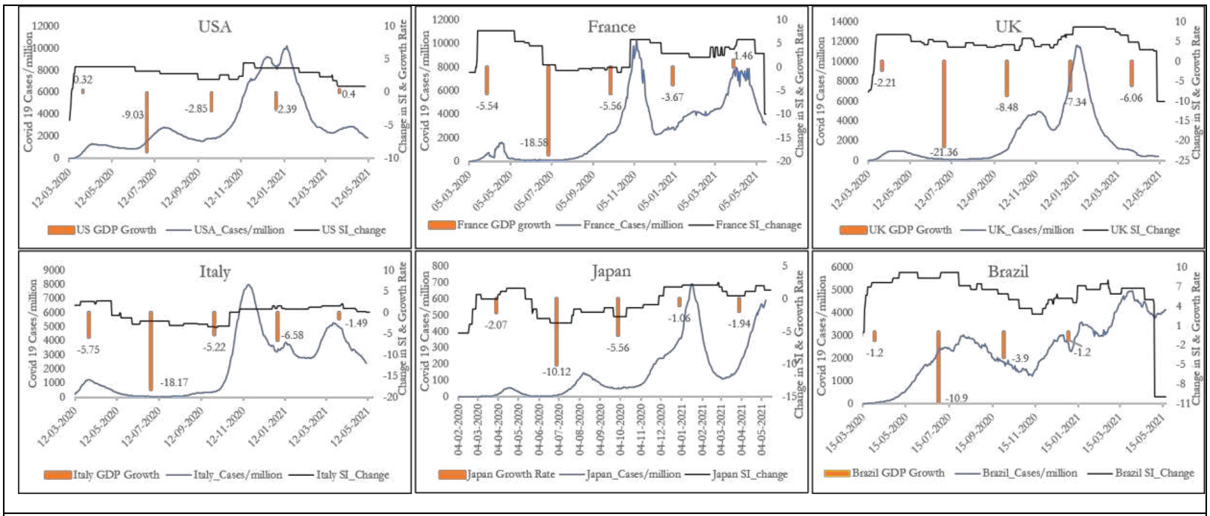
The plan of the paper is as follows. Section 2 briefly provides a background of COVID-19 shock in India. Section 3 explains the PLFS dataset which we use. We explore this unique dataset for our present purpose in section 4. Using the panel feature of the PLFS data we analyse the heterogeneous employment and income impact of the pandemic shock on formal and informal sector workers in section 5. Section 6 using machine learning techniques classifies the characteristics of the divergent labour market outcomes. Section 7 concludes the paper.

2. Impact of COVID-19 on India

The COVID-19 pandemic resulted in both loss of life as well as disruptions in economic activities across the world. It has claimed more almost three million lives and has forced governments in many countries to impose stringent restrictions on movement and gathering to stem the spread of the disease. Absent these restrictions, researchers have estimated that the death toll could have been higher by at least five times (Ferguson, et al. 2020). This would put the COVID-19 pandemic among the largest disasters in the past hundred years. India has been among the countries badly affected by the pandemic. It had the second highest number of recorded infections and third highest deaths, although it was more moderately placed in terms of per-capita infection and deaths (John Hopkins Coronavirus Resource Center). While many countries imposed a localised or graded lockdown, the Government of India imposed a complete lockdown across all part of the country during March 25, 2020 to April 14, 2020 to stop the spread of the virus. In terms of strictness, India's lockdown was ranked among the severest in the world by the University of Oxford COVID-19 Government Response Tracker (Mehrotra, 2020). Post April 14, 2020, the lockdown restrictions were localised and started lifting in a staggered manner as more data on the COVID cases became available. Specifically, the districts were classified into three zones: red, orange and green depending on the new COVID cases and the active cases using the reported numbers at the district level. The lockdown restrictions were then imposed with varied intensity based on the classification of the districts with red districts facing the toughest restrictions while the green districts facing the least restrictions.

Figure 1: Multi-country scenario of GDP Growth, COVID-19 Infection Spread and Stringency Index till mid 2021





Note: GDP data is sourced from MOSPI, IHME is the source of COVID-19 infection data and Stringency Index is provided by Oxford University.

Above figure summarizes the three domains together of various countries. It includes real GDP growth, COVID-19 cases per million of population and change in stringency index with respect to the middle of March 2020 level of stringency[4]. In terms of Convid-19 infection spread and the stringent restriction, the relation is mixed. However, the countries which experienced more stringent measures, also experienced a decline in GDP growth. India experienced most strict and prolonged restriction measures and a very sharp fall in GDP, along with the misery of widespread of pandemic.

The economic fallout of the pandemic and the collapse in global trade was exacerbated by the lockdown and other restrictions on movement imposed by the Government to limit the spread of the disease. India's Gross Domestic Product (GDP) is estimated to have decreased by 7.3% in 2020-21. The policymakers in India were not oblivious to the economic impact of such restrictions, however, given the limitation of health infrastructure in India to tackle such a pandemic, they prioritized saving lives over the economic fallout (Department of Economic Affairs, Ministry of Finance, 2021). However, the lower mortality has come at a significant economic cost. The lockdown or stay-at-home orders have led to loss of employment, decline in income, and fall in spending (Baek et al. 2020; Baker et al. 2020; and Gupta et al. 2020). Chetty et al. (2020) have shown that some of the unemployment caused by lockdown can be long-lasting and re-opening of economies did not have a significant impact in reducing such unemployment. The economic impact of the pandemic may have also increased inequality as evidence shows that it affected less educated workers and less efficient industries more (Alstadsæter et al. 2020; and Béland et al. 2020; and Mongey et al. 2020). Also, lockdown or limitation in activities in certain businesses or areas can have large overall macroeconomic effects because of networks and forward and backward

[4] We do a base change exercise for the stringency index to represent the stringency index and GDP growth together for all the 7 countries. We consider the stringency level of the first half of 2020 as the base and the estimated the rate of change with respect to that. The duration is set matching the PLFS data which is available till June 2021.

linkages (Barrot et al. 2020; Baqaee and Farhi 2020). The closure of businesses can also have negative impact on solvency and liquidity of businesses which may prolong the crisis even after lockdown is lifted (Carletti et al. 2020; Gourinchas et al. 2020; Schivardi and Romano 2020).
Some papers suggest that while lockdowns as a measure to counter pandemics may have some short-run costs, they do not necessarily have negative medium to long-run effects. For instance, Correria et al. (2020) show that stricter government interventions in cities in the USA during 1918-19 influenza pandemic had positive impacts in the medium run even though they were costly in the short run. Bodenhorn (2020) shows that mandated business closures in response to the Spanish flu did not necessarily result in more business failures in the USA. The long to medium-run impact of a pandemic on the economy is, therefore, debatable. It is important to analyse the economic recovery from the crisis to identify early signs of how much the economic landscape will be different in the post-pandemic world, and to rectify any economic uneven-ness which might have become exacerbated due to the crisis.

3. Data Description

We use the quarterly Periodic Labour force Survey (PLFS) data published by the National Statistical Office (NSO) of India for our analysis. From 2017-18 onward, this survey is the official source of labour force and employment data in India. Earlier, the official Indian labour force data was the quinquennial employment-unemployment surveys of the National Sample Survey Organisation (NSSO)[5] which was used to provide cross-sectional information. PLFS was initiated with an objective to provide detailed information about the workforce at a higher frequency for both rural and urban India. Additionally, for the first time in India, this data offers a rotational panel design, restricted to the urban population where every quarter a fresh set of households is surveyed and revisited for four consecutive quarters. Four rounds of PLFS unit-level survey data have been released so far, spanning both rural and urban regions for the periods July 2017- June 2018, July 2018-June 2019, July 2019 - June 2020 and July 2020- June 2021. The questionnaire contains household and individual information on various indicators like household size, type, religion, social group and usual consumer expenditure, sex, age, marital status, education, occupation, activity status, wages, income etc.

The PLFS uses a stratified multistage survey design for households' selection of both rural and urban areas. At the first stage, stratification is conducted within each NSS region according to the size class of the town as per the Population census of 2011 to form a 'stratum'. Then, First Stage

[5] NSSO was merged with Central Statistical Office (CSO) and formed NSO in 2019.

Units (FSU)[6] are selected from these strata according to the probability proportional to size with replacement (PPSWR) approach, size being the number of households in the Urban Frame Survey (UFS)[7] block in urban area and village in rural areas. After the identification of the FSU, based on population intervals, a suitable number of sub-blocks and hamlets are marked in urban and rural areas respectively, such that the population is equalized. Then, using simple random sampling, two sub-groups or hamlets are selected. Finally, within the selected sub-blocks and hamlets, a second stage stratification (SSS) is conducted in the sub-blocks and hamlets based on the number of members in the household having a level of general education as secondary or above. From each SSS, sample households are selected using simple random sampling without replacement.

This stratified multi-stage survey has two types of visits, the 'first visit schedule' and the 'revisit schedule'. Around 11.3 thousand urban and 13.5 thousand rural households are considered for each first visit schedule. The survey starts with the first visit for 25% FSUs of the annual allocation in the first quarter. In the subsequent quarters, 25% of new households are added. For rural households, the data is collected only once at the first-visit level. The urban households are surveyed at both levels such that the rotational panel structure can be set up. Therefore, 25% of the urban FSUs of the annual allocation, which is visited at the first visit level, are revisited in the next three quarters. That is, in case of the urban India, P_{11} (say) is the number of households who are surveyed at visit 1 ('first visit schedule') for 1^{st} panel. In the next quarter, P_{21} is the number of households who are revisited (visit 2) from the 1st panel (if there is no attrition then $P_{11} = P_{21}$) and P_{12} are the new households who are visited for the first time as the second panel. Similarly, in the subsequent quarter there are three sets of visits: third visit of the first panel (P_{31}), second visit of the second panel (P_{22}) and first visit of the third panel (P_{13}). Thus, the next quarter is the fourth and the last revisit of the households from the first panel (P_{41}). Also, third, second and first visit for the panels 2, 3 and 4, respectively. Table 1 describes the urban panel for one survey year.

Table 1: Rotational Panel Structure of the PLFS data for urban India

	Visit 1	Visit 2	Visit 3	Visit 4
Quarter 1	P_{11}			
Quarter 2	P_{12}	P_{21}		
Quarter 3	P_{13}	P_{22}	P_{31}	
Quarter 4	P_{14}	P_{23}	P_{32}	P_{41}

[6] https://mospi.gov.in/documents/213904/0/concepts_golden.pdf/e98fc072-8660-edd9-f179-ce95674f4ca5?t=1615539414160
[7] https://mospi.gov.in/web/mospi/urban-frame-survey-ufs

Several concerns with respect to the PLFS data has also been raised in the literature (see, Menon & Nath, 2022; Abdul & Sahoo, 2021). One of the serious issues is, 2017–18 and 2018–19 data release uses different sets of first-stage unit (FSU) numbers. As a result, it is very difficult to convincingly merge the data from these two years.

For our purpose, this data is the only source which we can be used. The official definition of the informal sector (mentioned in the introduction section) is an enterprise-based definition and the first visit schedule of the PLFS survey collects the enterprise information from the employee. Collecting the enterprise's information from the workers, along with other individual-level characteristics, is a unique feature of the PLFS data set in India. Since we are contrasting formal and informal sector labour market outcomes around the pandemic period, PLFS data which are available both pre and post first wave of COVID-19, is the best fit for our analysis. We use the first visit data for overall India to understand the complexion of Indian informal sector employment around the pandemic period. Later, while we exploit the revisit data to create a panel structure for urban India, the first visit information is used to identify whether a worker is an informal sector worker or a formal sector worker. Since, the PLFS data also provides information about the income of the individuals both in the first visit and the revisit level, using the panel structure, cumulative distribution of relative income is constructed for the formal and informal sector workers. We contrast the distributions of the pandemic time with the pre pandemic and post pandemic episodes. This captures the extent of the negative income shock and the rate of recovery. Going further ahead, we try to classify the characteristics of the individuals who were crucially affected during COVID-19 shock or recovered quickly afterwards. A large number of observations in the PLFS data set enables us to undertake the classification exercise using the machine learning procedure which is a well-established methodology.

4. Exploratory Data Analysis: Formal vis-à-vis Informal Sector Employment

The definition of the informal sector remained a point of debate in the literature. Chakraborty (2010) documented and discussed the merits of different definitions of the informal sector in length. The debate was further stirred after a recent article in SBI Ecowrap (Ecowrap, 2021) Here, we mostly stick with the definition given by NSO in the PLFS survey report[8], which was conceptualized according to the ICLS resolution of ILO. The definition suggests, "unincorporated enterprises owned by households, (i.e., proprietary and partnership enterprises including the informal producers' cooperatives) are largely considered as informal sector enterprises". To match

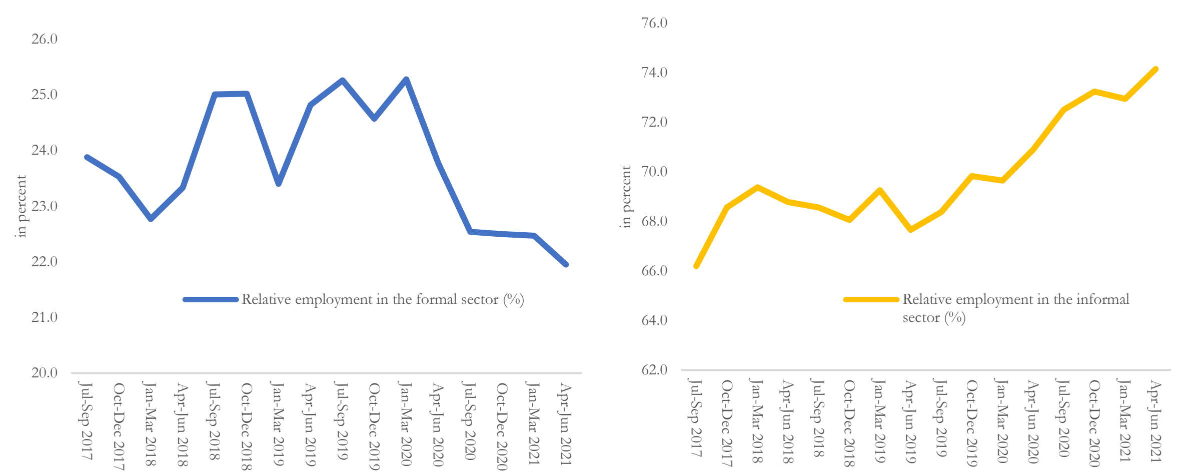
[8] https://www.mospi.gov.in/documents/213904/301563/Annual_Report_PLFS_2019_20m1627036454797.pdf/18afb74a-3980-ab83-0431-1e84321f75af

that definition in PLFS data, the report considered "proprietary and partnership enterprises" to be informal sector enterprises. Whereas, government and public sector enterprises, autonomous bodies, public/private limited companies, cooperative societies, and trust/other non-profit institutions are considered as formal sector enterprises. Following this definition, first, we find out the share of workers employed in the informal and formal sectors for each quarter from 2017-18 to 2019-20. Then, we go deep to understand the sectoral composition of the broad trend. Finally, we try to understand the transition in work status, given the worker was employed in the informal sector vis-à-vis the formal sector.

4.1. Formal and Informal Sector Employment Share: Trend and Distribution

4.1.1. Trend: Pre and Post COVID-19 Scenarios

The informal and formal sector employment is defined as the total number of individuals who work in the informal sector and formal sector, respectively. To understand the trend of informal sector employment share, we consider the first visits of the three PLFS surveys (from 2017-18 to 2019-20) in quarterly frequency. The first visit survey design not only represents samples from both rural and urban areas but also allows us to identify whether the worker works in the informal sector or formal sector as their primary activity status. We divide the number of employed persons in the (in)formal sector by the total number of employed persons to get a measure of the (in)formal sector employment share. We find that the relative employment share of the informal sector historically dominates over the formal sector employment share in India. More than 65 per cent of the labour force on average works in the informal sectors. On average the non-agricultural relative employment share of the informal sector is almost three times more than the formal sector employment share. Given this backdrop, if we look at the employment share in April to June of 2020 which reflects the time-period of the first wave of the COVID-19 pandemic and nationwide lockdown, then we find an even wider gap between formal and informal employment share. The formal employment share drooped down to 23.8% and the informal employment share spiked up to 70.9% during the first wave of the pandemic. Interesting to note that even after the first wave of the pandemic also the share of the informal sector remained elevated. April to June quarter of 2021 witnessed the informal sector employment share at 74.2% which is much higher than the pre-COVID level. Consequently, the formal sector employment share fell to 22% in that quarter.



(a) Formal Sector Employment Share (b) Informal Sector Employment Share

Figure 1: Trend in Employment Share of Informal vis-à-vis Formal Sectors

4.1.2. Industries and Services wise Distribution

As we understand the trend of the informal sector employment share, the question arises that how the incidence of informal sector employment is distributed among different industries and services. We use the National Industry Classification (NIC) category to construct 15 industry-service sectors, leaving agricultural and public administration. First, we find the informal employment concentration. That is, within the informal sector employment which industries and services are contributing most. Second, we go deep into each industry and service category to find out the formal and informal sector employment distribution.

Among the fifteen, four industries and services have the highest concentration of the informal sector workers. These four sectors are Wholesale and Retail Trade, Construction, Manufacturing and Transport and Storage. That remain consistent over per and post COVID-19 years. However, within these four industries-services, Construction became the leader in terms of informal sector employments in 2020-21, whereas in earlier years, Wholesale and Retail Trade had highest concentration of informal sector employments.

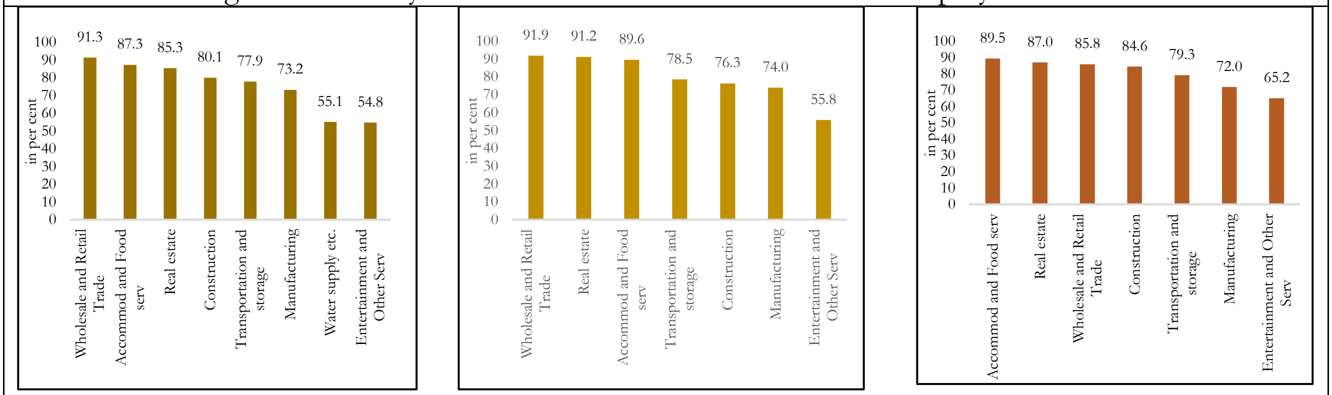
Table 2: Composition of Informal Sector Employment by Industry and Services (in Percentage)

	{1}	{2}	{3}	{4}	{5}	{6}	{7}	{8}	{9}	{10}	{11}	{12}	{13}	{14}	{15}
2017-18	0.5	23.3	0.1	0.3	23.1	24.5	10.3	4.4	0.7	0.7	0.5	2.7	2.7	1.2	4.8
2018-19	0.6	22.2	0.1	0.4	24.4	25.2	10.0	4.1	0.6	0.6	0.4	2.6	2.6	1.1	5.1
2019-20	0.3	21.1	0.1	0.4	25.3	26.1	10.0	4.1	0.7	0.6	0.4	2.7	2.4	1.1	4.6
2020-21	0.5	20.7	0.2	0.3	27.0	25.4	9.8	4.1	0.7	0.6	0.5	2.4	2.0	1.3	4.6

The list of industry and services: 1. Mining, 2. Manufacturing, 3. Electricity, gas etc., 4. Water supply and utilities., 5. Construction, 6. Wholesale and Retail Trade, 7. Transportation and storage, 8. Accommodation and Food services, 9. Information and communication, 10. Financial services, 11. Real estate, 12. Professional and admin support services, 13. Education Services, 14. Health Services, 15. Entertainment and Other Services.

Wholesale and Retail Trade, Accommodation and Food Services, Real estate, and Construction are the industries-services where the share of the informal sector workers is highest (figure 3). In each of these industries, the relative informal sector employment share is on average more than 80 per cent. The ranking in terms of informal sector employment share among these industries-services is not stable over time. For formal sector employment, the sectoral distribution is reported in figure (4). The sectors, like Electricity, Financial Services, Information Technology, and Education occupy the highest employment share under the formal sector in all three years.

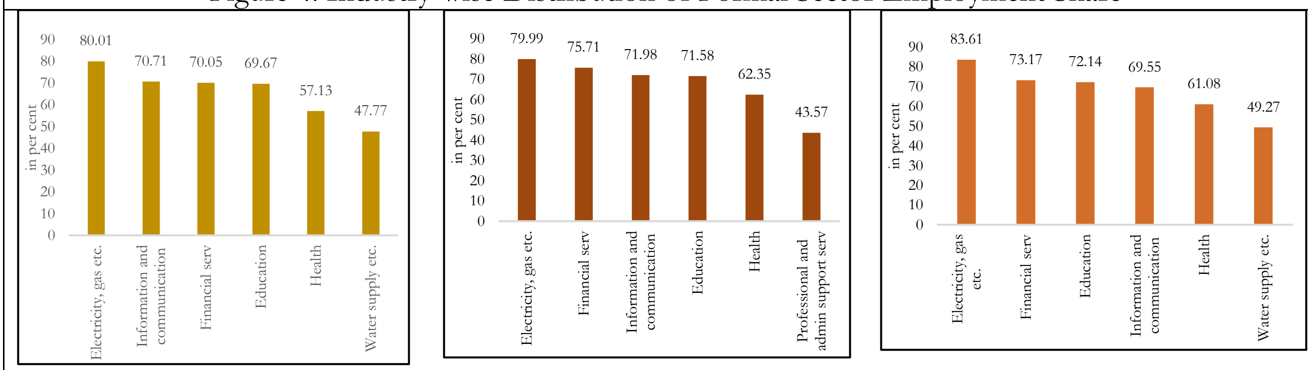
Figure 3: Industry-wise Distribution of Informal Sector Employment Share



(a) 2017-18 (b) 2018-19 (c) 2019-2020

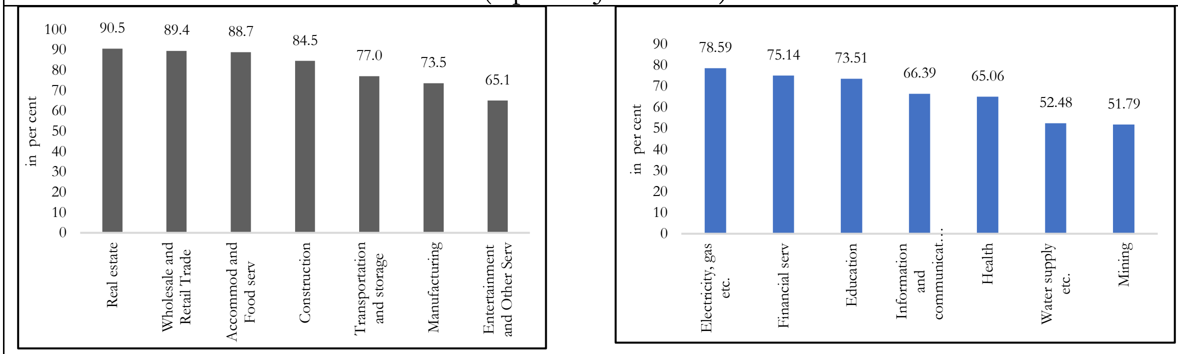
In contrast with the annual average, the sectoral distribution of informal employment share for April to June 2020 demands additional attention. This exercise tells us which industries-services cause the rise in informal sector employment share and the fall in formal sector employment share during the first wave of the pandemic. The Construction and Entertainment sector are the major contributor to the rise in the informal sector employment share in the April to June of 2020. Other sectors maintained the status quo in terms of informal sector employment share. The fall in the formal sector employment share was driven by Electricity and Information and Communication in the top 7 sectors ranked in terms of their formal sector employment share.

Figure 4: Industry-wise Distribution of Formal Sector Employment Share



(a) 2017-18 (b) 2018-19 (c) 2019-2020

Figure 5: Industry-wise distribution of Informal and Formal Sector Employment Share (April to June 2020)

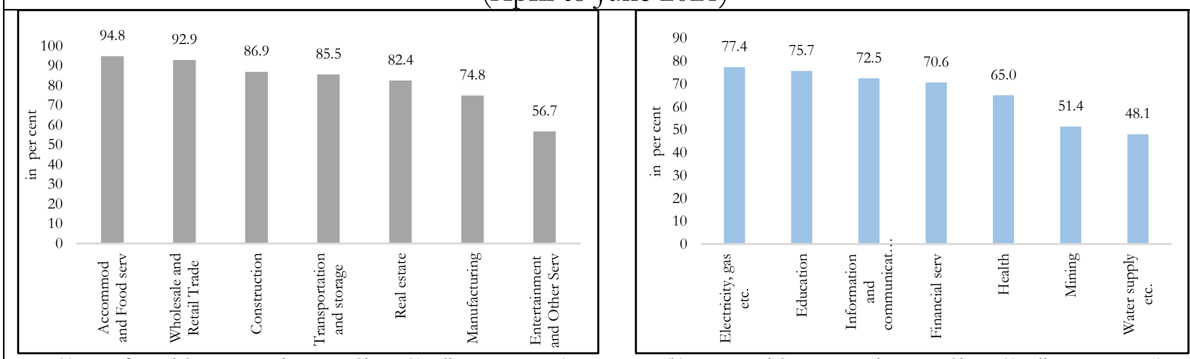


(a) Informal Sector Employment Share (April to June 2020) (b) Formal Sector Employment Share (April to June 2020)

However, shifts away from the formal sector also happened in sectors which were not ranked within the top 7. Sectors like Professional and administrative support services, Manufacturing, Transportation and storage, Real estate etc. witnessed a more than one per cent fall in the formal sector employment share from April to June of 2020 compared to the annual average of 2019-20. If we compare the April to June quarter of 2019 with the same quarter of 2020, we see in addition to Electricity and Information and Communication, there was a shift away from formal sector employment in Accommodation and Food Services, too. As it was anticipated, an additional push in the formalization was visible in the Health sector during the first wave of COVID-19.

Comparing the same quarter of the next year, i.e., April to June 2021, we try to understand how the sectoral distribution altered during the second wave phase (Figure 6). Accommodation and Food Services and Wholesale Retail Trade employed the largest share of the informal sector workers from April to June 2021, whereas Real Estate slipped behind compared to the first wave of the pandemic. However, Construction and Manufacturing, both kept the share of informal sector employment remained more or less unchanged.

Figure 6: Industry-wise distribution of Informal and Formal Sector Employment Share (April to June 2021)

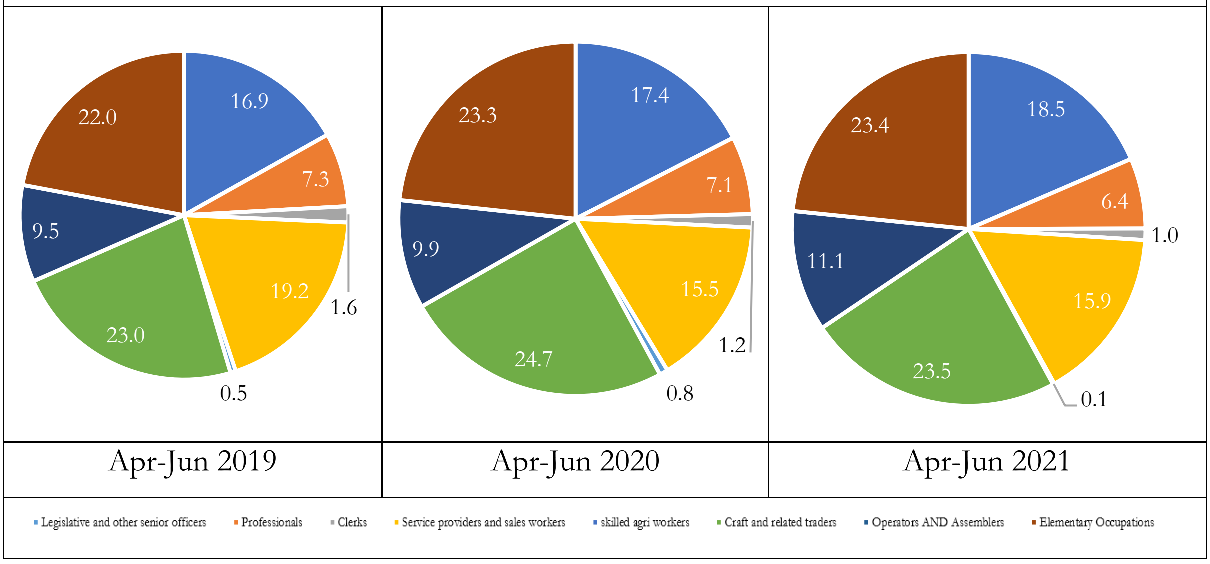


(a) Informal Sector Employment Share (April to June 2021) (b) Formal Sector Employment Share (April to June 2021)

4.1.3. Occupation-wise Distribution

Like industries, it also important to understand how the informal sector employment is distributed within different occupations and how that alters pre and post pandemic. The National Classification of Occupation (NCO) codes is provided at the individual level in the PLFS data. We consider the first digit classification to construct eight occupation categories. According to the occupation categories, the annual representation of the informal sector employment share is broadly stable. However, we check the occupational distribution of informal sector employment for April to June quarter which characterise the first wave of COVID-19 and subsequent lockdown in 2020.

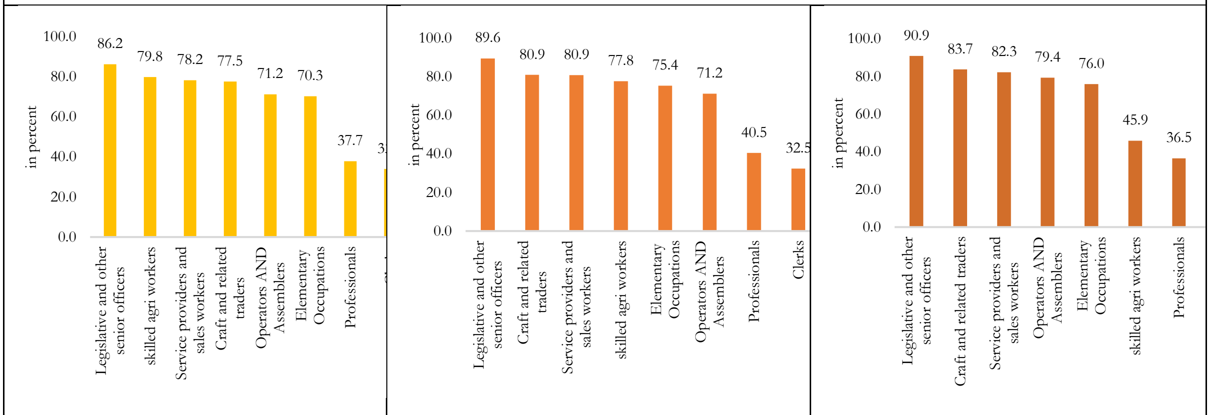
Figure 7: Informal Sector Employment share by Occupation



The informal sector workers' major concentration remains in two occupations: trading and elementary occupation. Post COVID, that increases notably in assembling and legislative occupations and reduces among service providers and professionals.

The share of the informal sector employment among different occupations for the April to June quarter is shown in the Figure 8. The share of the informal sector employment in most of the occupation categories went through a rise after COVID-19 shock. Post pandemic, the "Operators and assemblers" witnessed a very sharp increase in the informal sector employment share. The Legislative and other senior officers (which includes proprietary enterprises and owners of shops as well according to NCO definition), Craft and related traders, Service providers have high informal sector employment share in pre and post pandemic. However, the share increases in all these occupations after pandemic. Among professionals and clerks the share of informal sector employment shrunk in April to June quarter in 2020 and 2021 compared to 2019.

Figure 8: Occupation-wise Informal Sector Employment Share



Apr to Jun 2019 | Apr to Jun 2020 | Apr to Jun 2021

The exploratory data analysis suggests that informal sector employment is highly concentrated among few industries and occupations, both in pre and post COVID periods. Construction, Accommodation and food service, Retail and wholesale trade, Manufacturing and Transport are the industry categories which employ most of the informal sector workers, and also fuel the rise of overall share of informal sector employment after the pandemic shock. In occupation categories, Skilled agricultural workers and Professionals worked less in the informal sector after pandemic which was compensated by the increase of Shop-owners, Craftsmen, Elementary workers, and Service providers. This analysis indicates that a shock like COVID-19 squeezes the space for formal sector employment and pushes labourers towards informal sector employment. Few industries-services creates the space for informal sector employment the most, in certain occupation categories. This trend seems to be stable even in the recovery phase after the first wave. Now, it is interesting to ask how the labour market outcomes of the workers of the informal sector altered compared to their formal sector counterparts, which we explore in the next section.

5. Analysis of rotational panel data

5.1. The Transition in Work Status

In the previous subsections, we tried to understand the trend and sectoral division of informal sector employment. Then, we contrasted those broad trends with the quarter when the first wave of COVID-19 struck in India. Now, we look at the transitional dynamics of the labour market. That is, we examine if there is any heterogeneous trend in workforce transition from t to t+1 given the worker was at formal vis-à-vis informal sector at time period t, and also what happened to these trends during the first wave of the pandemic.

To find the transition matrix the panel structure of the PLFS data is used. We start this exercise by following the NSS definition of the informal and formal sectors. As mentioned in the data description, the panel in PLFS surveys only urban households. Also, the question about the enterprise type which determines whether the sector is formal or informal is asked only on the first visit of the PLFS survey. However, the working status is asked in the first visit as well as in the subsequent revisits under the current weekly status (CWS). The set of working statuses is

$$S = \{self\ employed, salaried\ employed, casual\ employed, sick\ but\ employed,$$

$$not\ working\ (but\ employed), unemployed, no\ participation\}.$$

The strategy which is followed here is to consider only the first visit and the first revisit to the households. The first visit is considered as period t and the revisit is considered as period t+1. Given that the first visit questionnaire enables us to identify whether the worker is associated with the sector $x \in X$ where $X = \{formal, informal\}$ we can ask the following question: what is the value of $P(s'_{t+1} \in S | s_t \in S\ and\ x_t \in X_t)$? That is, what is the probability of a worker switching (or, not switching) his/her work status from t to t+1 if the worker was working in informal sector vis-à-vis formal sector at period t? The following table Table 3 displays the probability of switching to unemployed from different working statuses for formal and informal sector workers. We consider only salaried and casual employment among the formal sector workers as these two employment statuses combines almost entire formal sector employment. Table 4 represents the probability of switching to not working from self, salaried and casual employment for informal sector worker and salaried and casual employed for formal sector workers. The not working status is important during the pandemic because there was a large number of workers who went out of payroll, but their jobs were not terminated. That means, the worker with not working status was supposed to join their old job, once the lockdown-like restrictions would ease out. If we compare both the tables (Table 3 and 4) it is important to note that the percentage of people who moved from working status to not working status was quite high during the first and second wave of the pandemic in contrast with the percentage of people who moved to unemployment from different working status. The job destruction rate (that is, the transition probability to unemployment) for the informal sector worker was higher in all categories compared to formal sector workers. The main brunt of the job loss was borne by the casual workers both in the formal sector and in the informal sector. It is important to note that the job destruction rate for casual workers remained elevated. Casual workers did not receive the not working status (i.e., retaining the job) as much as it did by other working categories. Both these two tables quantify the extent of negative shock faced by the workers during the first and second waves of the pandemic.

Table 3: Transition Probabilities to Unemployment (Represented in Percentage)

	(Given Informal sector employed at t by ps)			(Given Formal sector employed at t by ps)	
Time: t to t+1	Self employed to unemployed	Salaried employed to unemployed	Casual employed to unemployed	Salaried employed to unemployed	Casual employed to unemployed
July-Sept 17 to Oct-Dec 17	1.90	2.35	7.32	1.29	2.70
Oct-Dec 17 to Jan-Mar 18	1.54	2.58	6.47	1.76	7.64
Jan-Mar 18 to Apr-June 18	1.29	2.09	4.79	1.46	5.76
July-Sept 18 to Oct-Dec 18	1.46	2.43	7.40	1.11	6.86
Oct-Dec 18 to Jan-Mar 19	1.43	1.51	3.86	0.74	7.67
Jan-Mar 19 to Apr-June 19	0.86	1.58	4.81	0.74	4.14
Apr-June 19 to July-Sept 19	0.62	1.47	4.31	0.67	1.47
July-Sept 19 to Oct-Dec 19	0.35	1.31	3.24	0.39	4.79
Oct-Dec 19 to Jan-Mar 20	1.07	1.19	9.68	0.51	4.76
Jan-Mar 20 to Apr-June 20	7.41	10.15	50.64	4.83	38.08
Apr-June 20 to July-Sept 20	0.00	0.25	5.24	0.86	4.49
July-Sept 20 to Oct-Dec 20	0.65	0.92	3.97	0.42	4.35
Oct-Dec 19 to Jan-Mar 21	1.05	1.37	6.86	1.45	1.82
Jan-Mar 21 to Apr-June 21	2.33	4.32	23.85	0.97	21.96

Table 4: Transition Probabilities to *Not-Working* (Represented in Percentage)

	(Given Informal sector employed at t by ps)			(Given Formal sector employed at t by ps)	
Time: t to t+1	Self employed to Not-Working	Salaried employed to Not-Working	Casual employed to Not-Working	Salaried employed to Not-Working	Casual employed to Not-Working
July-Sept 17 to Oct-Dec 17	1.78	0.72	1.01	1.23	1.33
Oct-Dec 17 to Jan-Mar 18	1.68	0.75	0.70	0.77	0.00
Jan-Mar 18 to Apr-June 18	1.92	0.88	0.33	1.52	0.34
July-Sept 18 to Oct-Dec 18	0.92	1.32	0.24	2.43	3.18
Oct-Dec 18 to Jan-Mar 19	0.66	0.37	0.00	1.26	0.00
Jan-Mar 19 to Apr-June 19	0.86	1.10	0.05	2.00	0.00
Apr-June 19 to July-Sept 19	1.34	0.66	0.27	0.81	0.00
July-Sept 19 to Oct-Dec 19	1.32	1.59	0.64	0.71	0.00
Oct-Dec 19 to Jan-Mar 20	5.19	4.78	0.46	4.25	0.52
Jan-Mar 20 to Apr-June 20	38.12	48.53	2.84	35.51	1.25
Apr-June 20 to July-Sept 20	0.00	0.57	0.00	0.56	0.04
July-Sept 20 to Oct-Dec 20	0.67	0.28	0.08	0.85	0.00
Oct-Dec 19 to Jan-Mar 21	1.34	1.25	0.11	0.73	0.00
Jan-Mar 21 to Apr-June 21	13.55	15.71	0.44	10.82	0.00

5.2. Impact on Workers' Income

In this section, we use the urban panel of the PLFS data to track the effect of COVID shock on the income of the formal and informal sector workers. We add income all seven days a week (reported as income from current weekly status in PLFS data) and one-fourth of the monthly salaried income of each person such that we find the weekly income for an individual. Following the same individual, we try to find the answer, whether the income of that individual is reduced in the next quarter. We create a variable which is a ratio of income of the same individual but of two different time points. Formally, $X_{ij;t,t-k}$ is the ratio of income of the individual i employed in the sector $j \in \{informal, formal\}$ at quarter t and quarter $t-k$. That is,

$$X_{ij;t,t-k}\begin{cases} > 1, if\ income\ increases\ at\ t\ from\ t-k \\ < 1\ if\ income\ decreases\ at\ t\ from\ t-k \\ = 1\ if\ income\ unchanged\ at\ t\ from\ t-k. \end{cases}$$

We construct the distribution of $X_{ij;t,t-k}$. Table 5 provides a basic snapshot of the segmented nature of the Indian labour market in terms of heterogeneity in income dynamics. Even in per-COVID period income of the informal sector workers are more volatile than the formal sector workers. However, given the pandemic shock in the second quarter of 2020, 70% of the informal sector workers witnessed a fall in their income during the first wave whereas 30% of the formal worker witnessed the same in the urban areas of India. During the later quarters of the year 2020, more than 20% and 26% of the informal sector workers experienced income recovery which is higher compared to the formal sector workers. The percentage of informal sector workers who observed a decrease in income remained elevated in the third and fourth quarters of 2020.

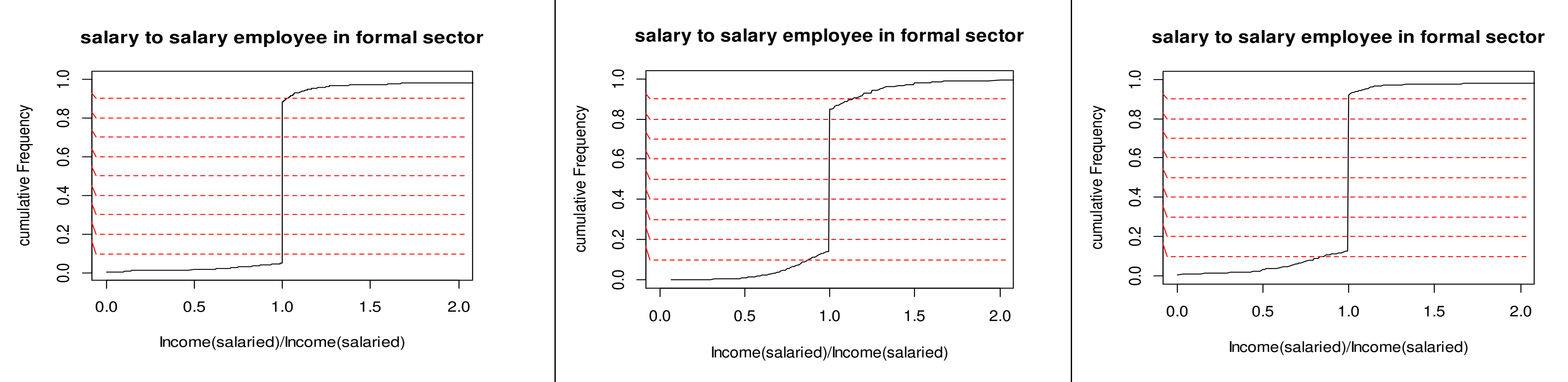
Table 5: Change in income of formal and informal workers

	Pre-COVID Apr-Jun 2019	**Post-COVID Apr-Jun 2020**	**Post-COVID Jul-Sep 2020**	**Post-COVID Oct-Dec 2020**
Formal				
Income increase	15%	8%	10%	15%
Income decrease	18%	31%	22%	20%
Informal				
Income increase	22%	9%	20%	26%
Income decrease	30%	70%	50%	43%

Next, we draw the cumulative income distribution charts for the workers whose working status remain unchanged given the COVID-19 shock (We also do the same for workers whose status changed to "Not working but employed", the charts are reported in the Appendix). Figure (9) and (10) shows that workers who were in the salaried work-status and remained salaried after the pandemic shock were to some extent insulated, both in the formal as well as the informal sector.

We compare the distribution of quarterly income change during the first wave of the pandemic (from 2020-Q1 to 2020-Q2) with the quarter before (2019-Q4 to 2020-Q1). We also present the income change distribution between 2019-Q1 and 2019-Q2 in case there are seasonal variations. It must be noted that each rotational panel is interviewed only for 4 consecutive quarters. Hence, the sample for the 2019-Q2 belongs to a different panel to the one from 2020-Q2, the same individuals are not interviewed in 2019-Q2 and 2020-Q2. The results indicate that for the formal sector salaried workers who remained salaried during the pandemic, the percentage of those for whom income declined quarter-on-quarter was similar to the same in the baseline. On the other hand, for the informal sector, this percentage increases from about 10% in the baseline to 20% during the pandemic.

Figure 10: Effect of first wave of COVID-19 on income distribution of formal sector salaried workers



Change in income between 2019-Q4 to 2020-Q1	Change in income between 2019-Q1 to 2019-Q2	Change in income between 2020-Q1 to 2020-Q2

Figure 11: Effect of first wave of COVID-19 on income distribution of informal sector salaried workers

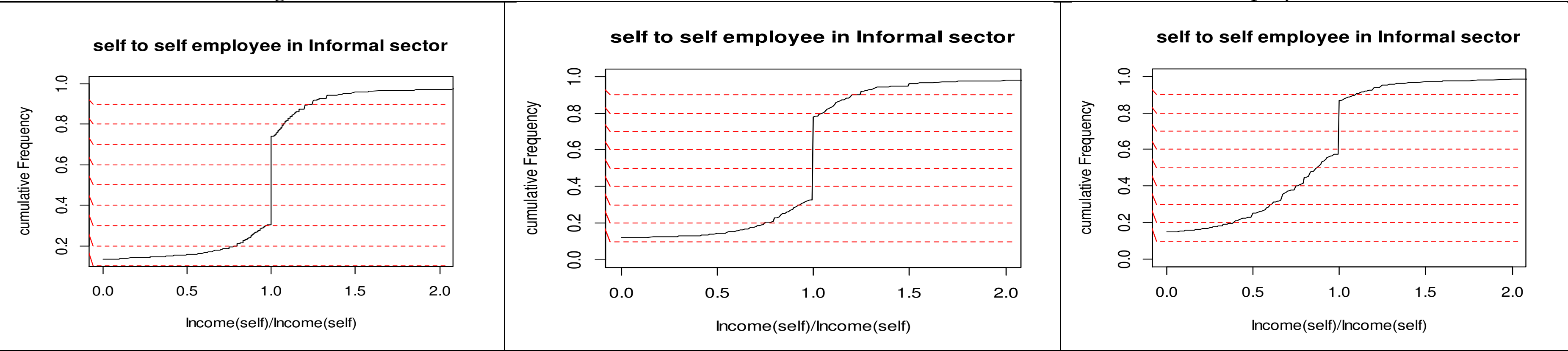
Change in income between 2019-Q4 to 2020-Q1	Change in income between 2019-Q1 to 2019-Q2	Change in income between 2020-Q1 to 2020-Q2

Next, we analyse the self-employed category which represents the largest segment of the labour force. The self-employed workers belong almost exclusively to the informal sector. As before, we compare the distribution between three periods i.e. 2020-Q1 to 2020-Q2, 2019-Q4 to 2020-Q1, and 2019-Q1 to 2019-Q2. It is apparent from figure (12) that income is more volatile for the self-employed category when compared to the salaried category in all the periods. The percentage of workers whose income increased (decrease) in the pre-pandemic was about 25% (30%) for the self-employed, while the corresponding numbers for the salaried class is 20% (15%). As expected the self-employed category is characterized by higher risk and higher reward. Secondly, the self-employed category had suffered a sharp change in the income change distribution during the pandemic's first wave. The percentage of workers whose income declined quarter-on-quarter increased to almost 60% from 30% in the baseline. The percentage for whom income increased during the same period was about 12% as compared to 22% in the pre-pandemic period.

Lastly, we analyse the income change distribution during the recovery period for the salaried segments in the formal and informal sectors, and self-employed in the informal sector (figure 13). To catch the recovery in income, we look at the change in income between 2020-Q2 and 2020-Q4. We skip 2020-Q3 because there were still widespread restrictions in place during the period and the economy fully opened up in 2020-Q4. We see that for the salaried employees in the formal sector the income change distribution do not show any significant difference with the pre-pandemic or the pandemic periods. This is expected as the pandemic had little effect on the income of the salaried formal sector workers. For the informal sector salaried employees, two differences are noticeable – the percentage of workers whose income increased is 30% which is higher than the pre-pandemic (20%) or pandemic levels (10%), and second, the percentage of workers whose income declined is also lower than the baseline (10% compared to 20%). As seen in figure (11) there was a moderate impact of pandemic on the informal sector salaried workers with percentage of workers whose income declined rising by 10%. Therefore, the distribution indicates salaried workers mostly recovered within 2020-Q4. Finally, in the self-employed category the percentage whose income increased rose to almost 50% as compared to 20-25% in the baseline scenario. However, as written above the percentage of workers whose income declined during pandemic had increased to 60% from 25% baseline. Therefore, the income distribution shows that informal sector self-employed had not yet fully recovered by end of 2020.

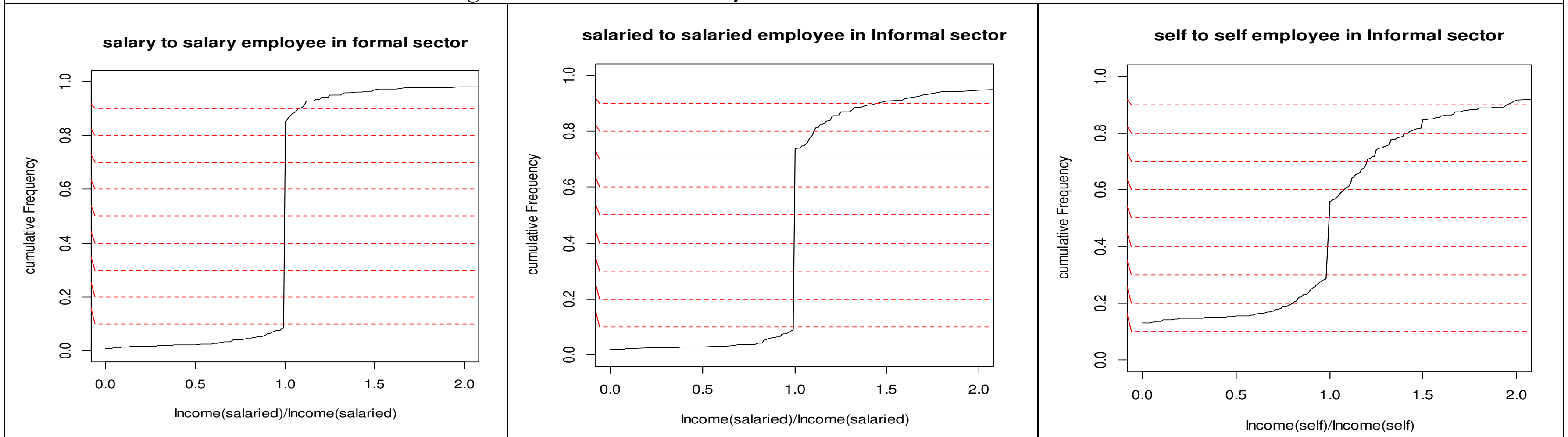
The analysis shows that not all workers' income were similarly impacted by COVID-19. While formal sector salaried workers were largely immune to the shock, the salaried in the informal sector faced moderate rise in percentage of workers whose income declined. The most severely impacted were self-employed. The evidence also shows recovery may also be uneven.

Figure 12: Effect of first wave of COVID-19 on income distribution of informal sector self employed



Change in income between 2019-Q4 to 2020-Q1	Change in income between 2019-Q1 to 2019-Q2	Change in income between 2020-Q1 to 2020-Q2

Figure 13: Income recovery from the first wave of COVID-19



Change in income between 2020-Q2 to 2020-Q4	Change in income between 2020-Q2 to 2020-Q4	Change in income between 2020-Q2 to 2020-Q4

6. Classification Analysis and Empirical Findings

In the earlier sections our exploratory data analysis presented stylized facts on the differential effect of a generalized shock on employment and income in a segmented labour market. In this section we attempt to formalize the classification with appropriate empirical analyses. We use the PLFS data to evaluate the available characteristics of workers in order identify workers whose income recovered faster after the adverse shock. Availability of different round of PLFS data which can be used to map with different phases of COVID shock gives us a unique opportunity to test our hypotheses, and to appropriately compare them with counterfactuals.

For the rest of this section, we have adapted machine learning based classification techniques, rather than using conventional methodologies that depend on assumptions on functional forms and distribution. In a volatile environment like post COVID period it becomes extremely difficult to attribute a particular functional form to the data generating process, and therefore we leave it to Machine learning models. They are getting increasingly popular in predicting outcome/nowcasting as they allow data to choose the appropriate functions that fits the underlying data best and allow non-linearities where appropriate. Availability of large number of samples in the PLFS data help us to use ML techniques without any glitches.

There are numerous classification models that may be found in the literature, each with advantages and disadvantages. We include Logistic classification, Gaussian Naive Bayesian, K-Neighbours Classifier, SGD-Classifier, Nearest Centroid, CART, Decision Tree Classifier, Random Forest Classifier, and Bagging Classifier for our analysis. To briefly mention the main differences between these classifiers, Logistic classification, for instance, uses direct probability estimation using a sigmoid function and chooses the parameter values by minimizing the log-loss or cross entropy functions to arrive at predictions from decision. Gaussian Naive Bayes, on the other hand, assumes that each feature in each class follow a Gaussian (normal) distribution. Naïve indicates the strong assumption that the conditional distributions of each of the features are independent and it applies Bayes Theorem. The major advantages of Gaussian NB class are that it handles both continuous and discrete data and could be used in real time projections. K-Neighbours, on the other hand, is a nonparametric supervised model, which calculates the geometric distance of the features to make the classification. The Nearest centroid method can be viewed an extension, which calculates the centroid of a cluster and then calculates the distance between the centroid. The decision tree classifier (Pang-Ning et al., 2006) creates the classification model by building a decision tree. Decision trees are indeed a very intuitive approach that solve problems with a tree and leaf representation. The *pruning*, or the level at which to stop the decision tree branches depend on a pre-specified criterion. The CART algorithm is a type of classification algorithm that is required to build a decision tree based on Gini's impurity index.

To narrow down to a few of them we tease out classification *accuracy scores* using training data and K-fold cross verifications. Our goal is to extract as much useful information from the data as we can while syphoning out noise, as is the case with any machine learning modelling. A complex model might outperform the training set but may not fit the test data very well because it captures both information and noise well. As a result, we try to assess the final set of chosen models, after carefully evaluating the selection criteria for both the training set and the testing set.

6.1. Workforce classification during income recovery

In this section we fundamentally focus on the question that, which are the workers whose income increased during the recovery phase after the economy re-opened in July 2020, and whether we can classify this increase with the underlying features of the person, i.e., in term of their demographics, geographical location, occupation wish variation etc., during this generalized recovery phase. A COVID shock impacted income of the entire workforce adversely. With vaccination, medication, and hard immunity development as the severity of the pandemic started declining over time, employment and economic activity pan India started recovering. We take PLFS survey round April-June 2020 (V2), when the first COVID wave was sweeping across India and compare it with Oct-Dec 2020 (V4) for the same set of workers.

As mentioned in the earlier section, we ran a horserace among the category variables and evaluate these models in terms of their K-fold cross verification method. To start with, we use 'accuracy score' for each of these models and calculate their standard deviations for selecting these models. Based on their accuracy scores Random Forest (RF) decision tree seem to be the best selection.

Table 6: Classification Model Selection

Classification	Accuracy	SD of Accuracy
LR	0.535816	0.023663
GNB	0.542631	0.026772
KNC	0.476619	0.017381
SGDC	0.472955	0.095779
NC	0.428923	0.040670
CART	0.472193	0.022006
RF	0.577381	0.017994
BC	0.531388	0.014354

We thereafter use RF algorithm to train the data and then calculate the accuracy score based on the training set and for the testing set. While the Train accuracy score is around 0.99 the test score was close to 0.62, indicating possible sample overfitting. To confirm this further, we evaluate the same with the second-best classification model, i.e. GNB, and find similar result as indicated by RF classification. Given that the accuracy score is higher than that naïve unbiased coin toss, if we use the RF-Classification Model then the accuracy score and the other measure of accuracy is reported in the Table 7.

Table 7: Precision, Recall and Accuracy, Random Forest Model

	precision	recall	f1-score	support
0	0.52	0.34	0.41	805
1	0.67	0.70	0.68	1244
2	0.61	0.70	0.65	1474
accuracy			0.62	3523
macro avg	0.60	0.58	0.58	3523
weighted avg	0.61	0.62	0.61	3523

6.1.1. Dimension reduction and feature selection:

The accuracy score difference between the train and test set indicate the possibility of model overfitting and suggest a scope of dimension reduction by selecting the most relevant features. The commonly used feature selection methods with categorical data and target variable are the chi-squared statistic and the mutual information statistic. We use Chi-sq statistics, which uses independence of the test statistics, for feature selection. The variable that are found to be independent of the target variables, i.e. for those Chi-sq reports a very low value, could perhaps be dropped to reduce data dimensionality (Figure 9(a)) and overfitting.

Alternatively, the same can be inferred from the "mutual information for machine learning" method, which indicate reduction in uncertainty regarding one variable given another series information. Figure 9(b) plots mutual information scores for each of the features and the target variable. Both the figure indicates selection of the same set of variables, which include 'state code, employment sector dummy, 'nic_code_dummy', 'nco_code_dummy', and 'emp_status_V1'.

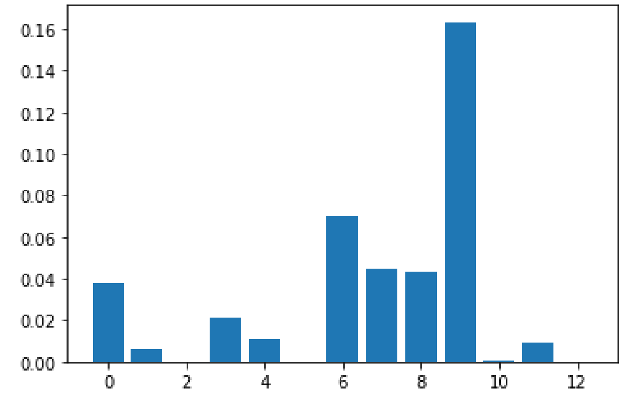
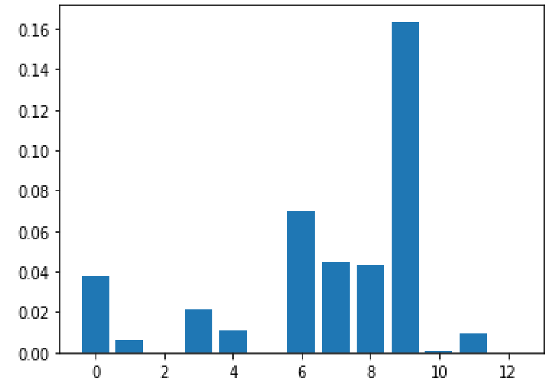
	

Figure 9(a): Mutual independence	Figure 9(b): Mutual information criteria

Name of variables: 0='State_Code', 1='Marital_Status', 2='Age_dummy', 3='Gender_dummy', 4='education_dummy', 5='vocational_training', 6='employment_sector_dummy',7= 'nic_code_dummy',8= 'nco_code_dummy', 9='emp_status_V1',10= 'zone', 11='Social_Group', 12='family_edu_ratio'

This selection of variables helped reducing the difference between Training and Test set accuracy scores in the Random Forest decision classifier. As a second check, we have also used Logit classifier to evaluate the difference between the logit train and test accuracy scores. Table 8(a) and Table 8(b) indicate that there was marginal difference between the two. However, since the accuracy score was better with Random Forest model, we continue with the same for the rest of our analysis. It is evident from the precision scores, recall score and f1 scores that the class with same income (1) and increase in income (2) performed better that the decline in income class (0) during the sample period.

Table 8(a): Training and test scores

Random forest

Train Accuracy Score 0.846

Test Accuracy Score 0.586

****

Logit model

Train Accuracy Score 0.536

Test Accuracy Score 0.541

Table 8(b): Other Scores-Random Forest

	precision	recall	f1-score	support
0	0.41	0.34	0.37	805
1	0.65	0.66	0.66	1244
2	0.58	0.63	0.61	1474
accuracy			0.57	3523
macro avg	0.55	0.54	0.54	3523
weighted avg	0.57	0.57	0.57	3523

6.1.2. Confusion Matrix

Classification scores can be misleading in presence of unequal number of observations in each class or in our cases where we have three classes in your dataset. In such cases a confusion matrix can give better visualisation of the classification model, where it is performing well and the types of errors it is making, for a given set of test data and a supervised model. For us, the classification Matrix is a 3X3 matrix, with row sums giving the actual values of income fall, same income and increase in income, while the columns reports the predicted values in each classes and the errors in classification. Because of the presence of large numbers and imbalance categories, we report both confusion matrix with absolute numbers and with percentage numbers. From both the confusion matrix, it is quite evident that our classification model with, only four variables, i.e. state, industry classification, occupation classification, and employment status, could reasonably classify increase in income and same income for the sample period.

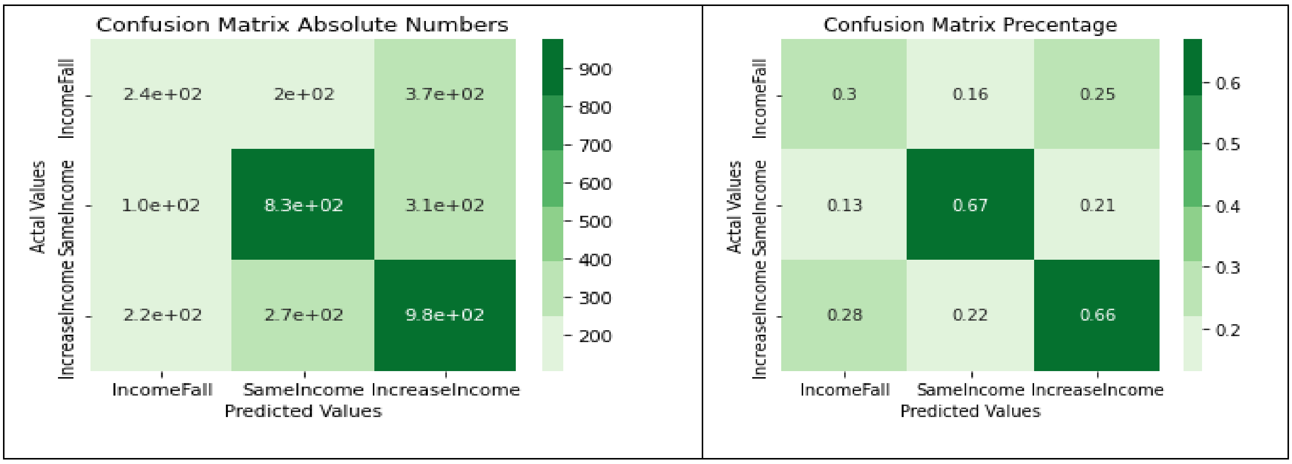


Figure 10: Classification matrix during income recovery

6.1.3. Synthetic Sampling

One disadvantage of classification matrix highlighted in the ML-literature is its application for imbalanced data, wherein the model predicts that each point belongs to the majority class label, as in our case, but the model may not be that accurate. To address such issues, literature suggest use of synthetic sampling, which we explore in this section.

Figure 11(a):Class Distribution Actual

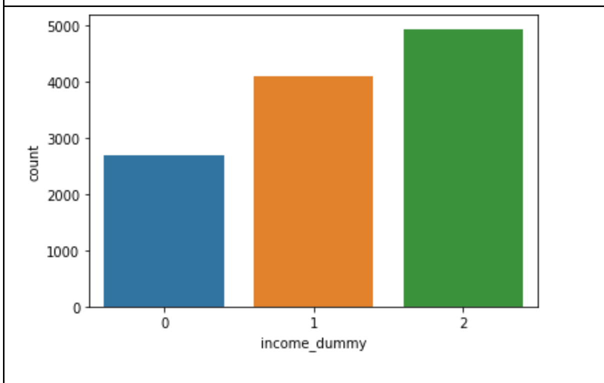
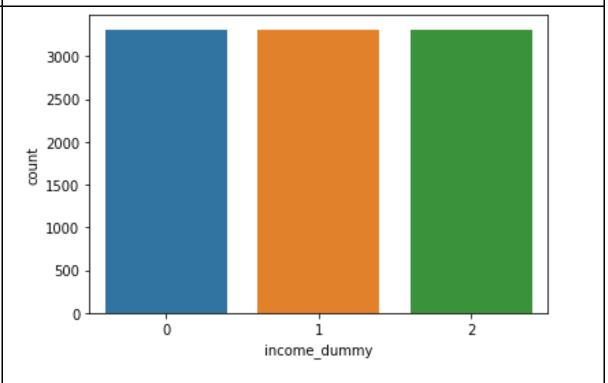


Figure 11(b): Post SMOT Class Distribution



Our sample periods V2-V4 was characterised by large number of agents experiencing increasing income, mainly because of the broad based recovery in the post-pandemic economic activities. Therefore, individuals that have undergone decline in income (Income_dummy=0, if income of the person has actually declined) became the minority class vis-à-vis the majority class (income dummy=1, persons whose income has indeed increased). The problem that often arises in presence of majority class is that the machine learning algorithm is trained to identify the majority class better as compared with the minority class. To address this problem we use Synthetic Minority Oversampling Technique (SMOTE). SMOTE is done in three steps, first select an observation from the minority class, there drawn minority class sample from the nearest neighbours of the selected sample. It thereafter joins the selected observations and the nearest neighbourhood are drawn, and create another sample of minority class from the linear combination. Such resampling will lead to upscaling minority class observations without leading to overfitting problem.

As expected, after SMOT, the classification scores improve significantly for the minority class, that is in classifying those who have undergone a decline in income between V2 and V4 (Table 9). The classification matrix also indicate improvement in the minority class, which reports 12 percentage points increase in the true positive scores. These findings add credence to our classification model, which indicates faced with a rare adverse productivity shock, income recovery of the Indian workforce crucially hinges on features that include state, industry, occupational class and employment status.

Table 9: Classification Scores, Synthetic Sampling

	precision	recall	f1-score	support
0	0.40	0.42	0.41	892
1	0.63	0.66	0.64	1351
2	0.60	0.56	0.58	1632
accuracy			0.56	3875
macro avg	0.54	0.55	0.54	3875
weighted avg	0.56	0.56	0.56	3875

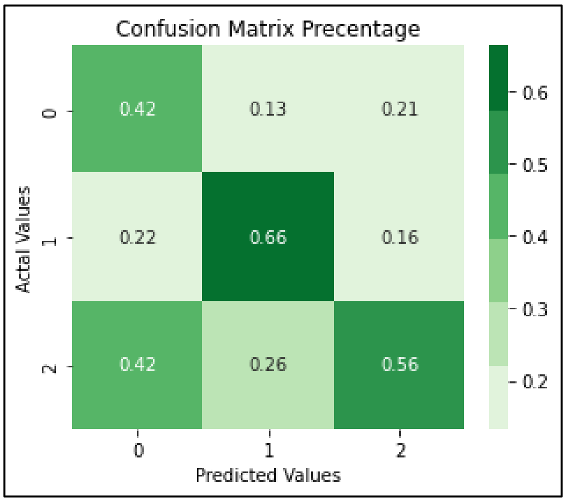


Figure 12: Confusion Matrix, Synthetic Sampling

6.2. Workforce classification during income reduction

While the above section considers workforce classification during income recovery, we in this section consider classifying workforce when there was a generalised decline in income. This can be done using PLFS survey rounds data immediately before lockdown (Jan-March,2020 (V1)) and aftermath (April-June,2020 (V2)) of the Pandemic-lockdown. However, there is a difference between the income recovery period (considered in the Section 6.1) vis-à-vis income decline in this section. Unlike the gradual improvements in income levels during the post-pandemic recovery period, income decline during April to June 2020 was abrupt and broad based, mainly triggered by sudden-stop due to statutory lockdown during the first COVID wave.

We follow all the broad steps in the previous section, however, for brevity, we only highlight the major findings in the subsequent sections. Alike the previous section, we run a large number of ML models to identify the best classification model the captures the pattern underlying our survey data. Our accuracy score and its standard deviation highlight that the GNB model fits the dataset the best. After training the GNB model with 70 per cent of the available observation, our training and testing accuracy turns out to be as follows: GNB Train Accuracy Score is 0.647 and GNB Train Accuracy Score is 0.654.

The percentage confusion matrix, when all features were included in the training set, is in figure. The figure indicates that while the model was well suited for classifying the sudden income decline due to the adverse productivity shock, it didn't do that well when it comes to classifying the increase in income. This is primarily due to the presence of minority class (Figure 11), which failed to train the model adequately for the income increase. We attempt to address this problem in the subsequent sections.

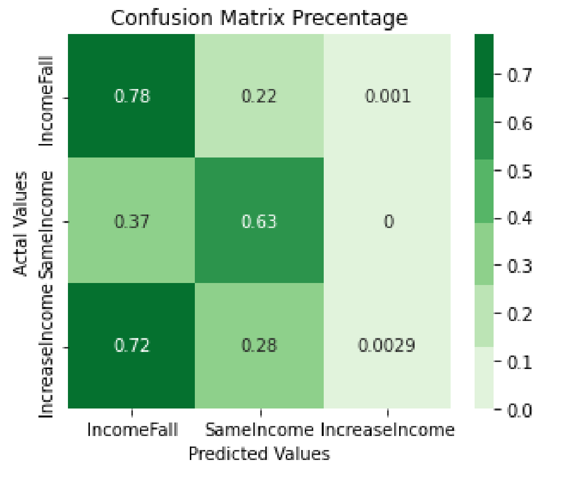


Figure 13(a): Confusion Matrix

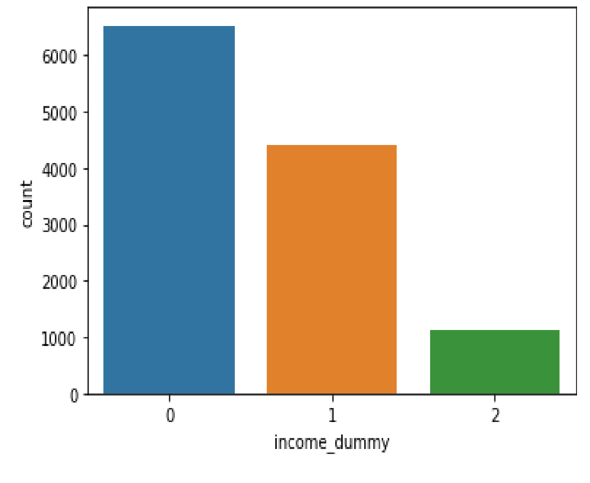


Figure 13(b): Class Distributions

In contrast to the income recovery period (Section 6.1), the accuracy scores here do not indicate significant overfitting when all features were included in the classification. However, for completeness, we repeat mutual independence and mutual information scores for the features during the V1-V2 period. The scores and their implications are summarized in the Figures below:

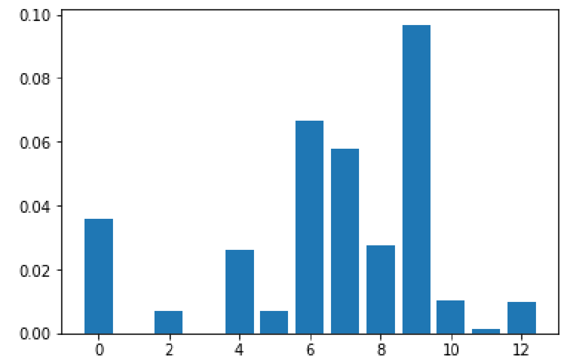


Figure 14(a): Mutual independence (V1-V2)

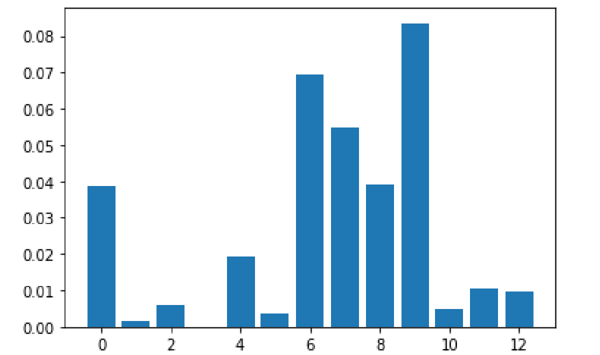


Figure 14(b): Mutual information criteria (V1-V2)

Name of variables: 0='State_Code', 1='Marital_Status', 2='Age_dummy', 3='Gender_dummy', 4='education_dummy', 5='vocational_training', 6='employment_sector_dummy',7= 'nic_code_dummy', 8='nco_code_dummy', 9='emp_status_V1', 10= 'zone', 11='Social_Group', 12='family_edu_ratio'

A few noteworthy points that emerge from the abovementioned feature selection steps are first, education dummy and family education ratio turn out to be significant in classifying individual those underwent a decline in income faced with an adverse productivity shock. Secondly, zones, that classifies COVID affected districts in terms of intensity of lockdowns (Red, Orange or

Green)[9] appears to assume greater significance during the downturn and on consequence income losses. Finally, one of the criterion suggest that inclusion of *social groups* could also be considered as useful consideration for explaining income change during an adverse supply side shock.

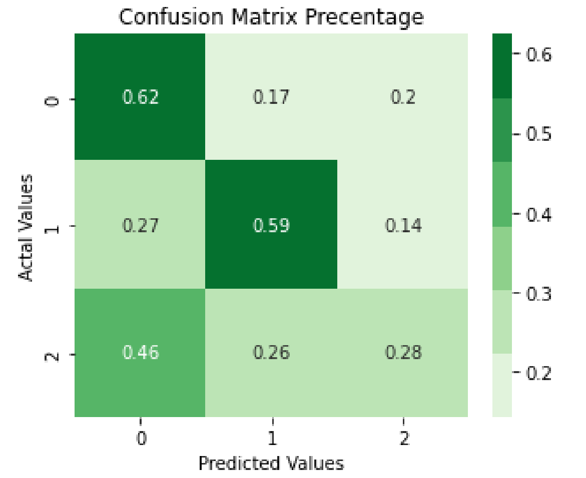


Figure 15(a): Classifying Income Loss with a Subset of Features: SMOT

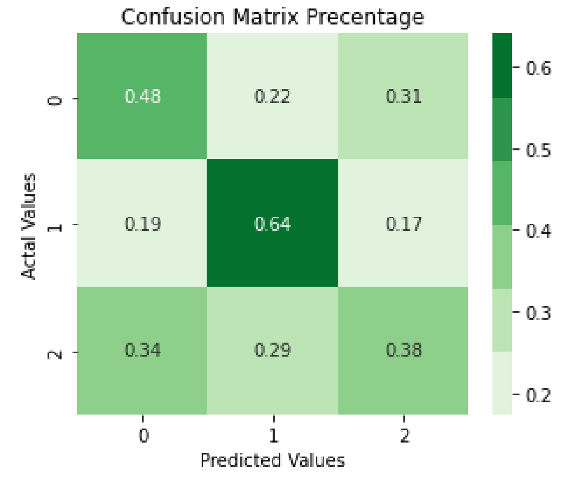


Figure 15(b): Classifying Income Loss with a Selected Features: SMOT

Figure 15(a) reports the confusion matrix during the rapid income fall due to an adverse productivity shock when we use a subset of features selected in Section 6.1.2. These include 'state code, employment sector dummy, 'nic_code_dummy', 'nco_code_dummy', and 'emp_status_V1'. We carried out a resampling based on the SMOT approach to address the class imbalance problem. A comparison with Figure 13(a), indicate clear improvement is classifying the features that indicate the improvement in income by around 22 percentage points.

In the next step, we include the variables that were selected in addition to the abovementioned list for classify income fall during the first pandemic wave. These include education, family education ratio, zone, and social groups. The inclusion in these variables helped in classifying income outcomes. In particular, the classification of the income increases by another 10 percentage points. It indicates that education and better family education perhaps protected income even in the case of rearrest supply shock. The demographic characteristics, for instance, age, sex, marital status, however, failed to add to this classification problem.

[9] Published by MHA, GoI (https://pib.gov.in/PressReleasePage.aspx?PRID=1620095)

6.3. Counterfactual

In search of a suitable counterfactual, we turn to a normal period which is not too far away from the pandemic years. In this vein we use PLFS survey data for Jan-March 2019 and April-June 2019, when available information and anecdotal evidence suggest that the pandemic shock was absent in India. During this period, our best guess is that income rise, or fall will be few and rather random, free from any major classification patterns.

Table 10: Accuracy Scores and Their Standard Deviation

	Accuracy Scores	Std. Deviation
LR	0.513346	0.071161
GNB	0.516298	0.062642
KNC	0.464814	0.068866
SGDC	0.449447	0.139880
NC	0.331355	0.100826
CART	0.438648	0.064024
RF	0.518006	0.066888
BC	0.472271	0.061750

To analyse the same, we follow all the steps noted in the Section 6.1. Our accuracy-score driven model selection indicate that the classification scores closely hover 50 per centage (Table 11). This is in a way an indication of a random job selection pattern. In terms of different scores Random forest model appears to perform marginally better than other class of models. We repeat the experiment with the features selected in Section 6.1 (Features Subset I), and the observed scores and reported in Table 11.

Table 11: Normal Period, Random Forest Model Scores

Feature Subset I		precision	recall	f1-score	support
Train Accuracy Score 0.817	0	0.46	0.40	0.42	1022
Test Accuracy Score 0.534	1	0.63	0.80	0.70	2008
	2	0.33	0.16	0.22	834
Feature Subset II					
rain Accuracy Score 0.952	accuracy			0.56	3864
Test Accuracy Score 0.554	macro avg	0.47	0.45	0.45	3864
	weighted avg	0.52	0.56	0.52	3864

When we extend the model further to include the variables identified in Section 6.2 (i.e. education, family education ration and dependency ratio, Feature Subset II), our model performance improve partially, along with along with evidences of overfitting.

Finally, based on the Random Forest model, the confusion matrix is reported in Figure 16(a) indicate that our classification model can classify those remained in the same income level (the majority class) with 80 percent accuracy. However, accuracy remained somewhat muted in classifying both income increase and income falls. Considering that both categories were minority classes during the normal time, we undertake synthetic resampling (SMOT) as explained in the earlier sections. After addressing the minority classes, when we attempt to classify features in terms income increase and fall, the improvement in the True Positive classifications were at best marginal.

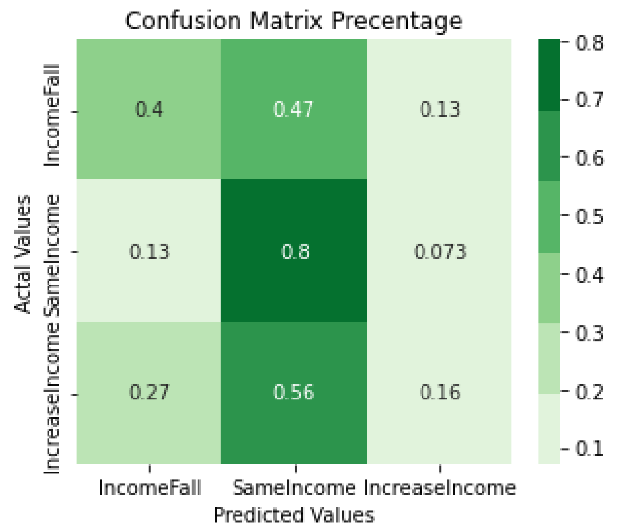


Figure 16(a): Confusion Matrix-Original Sample

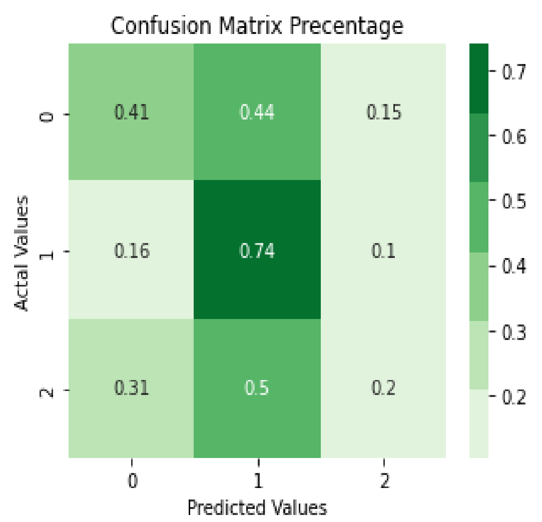


Figure 16(b): Confusion Matrix-SMOT

6.4. Robustness Test

Finally, in this section we consider the robustness of our findings. This is best done a related but different variable as target variable. For our purpose, PLFS data reports unemployment of the workforce besides income related information. We use the change in unemployment status during the recovery period (i.e. V2-V4). During this period most of the surveyed individuals were either in the same employment class or in "Otherwise class", fresh unemployed only accounting for 25 observations (Figure -14(a)). In this section we closely follow the steps mention in Section 6.1, while using changes in employment status as target variable used Feature Subset II and synthetic sampling. The confusion matrix for the same is report Figure 17(b). Employment Status at first survey visit, which reports self-employed, salaried, etc., emerges as the major explanatory variable for this classification model (Figure 17). If we drop that, the accuracy score falls drastically to around 68 per cent, which is also evident from the classification chart (Panel 4 of Figure 17).

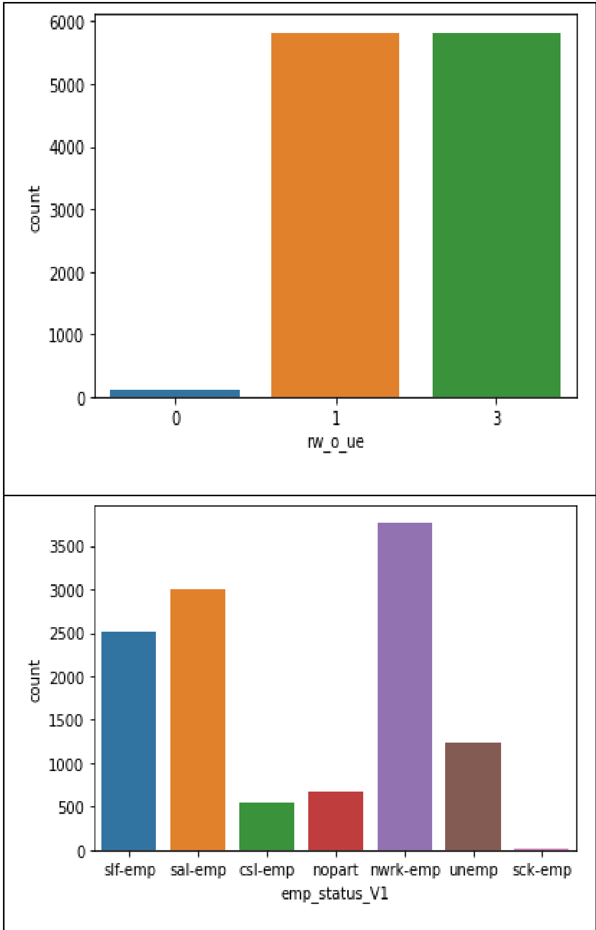


Figure 17(a): Unemployment Distribution

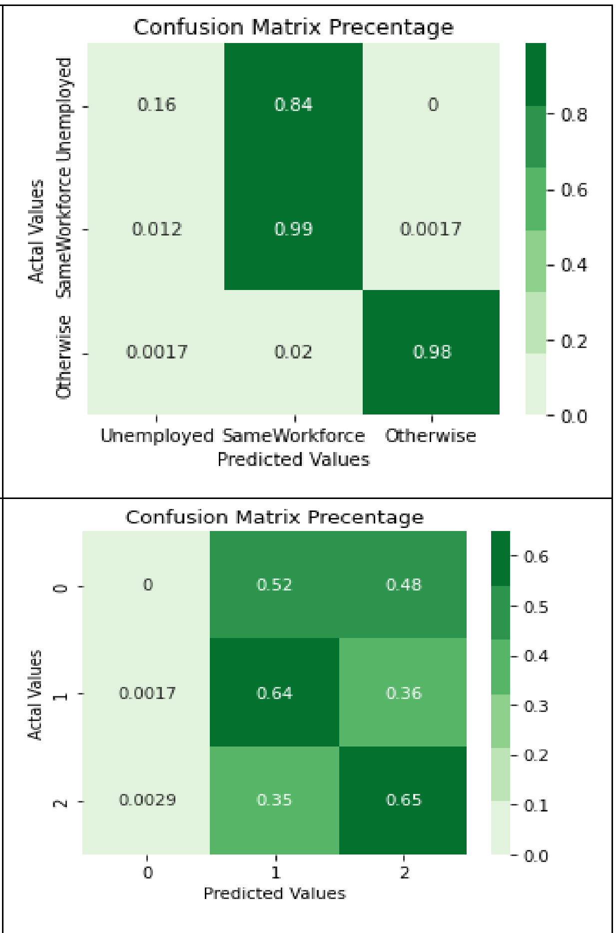


Figure 17(b): Confusion Matrix

6.5. Discussions

Our classification analysis indicates that during the post-pandemic recovery phase, four features assumed importance in classifying income variations (same, rise and fall). They include state, sector, industry classification, occupation classification and employment status at the first visit (feature set I). When we consider an adverse supply shock (pandemic shock) and income losses, a few additional characteristics significantly helped to classify those who experienced income resilience. These variables include education (primary school, secondary school, higher education, etc.), family education ratio (based on general education), inclusion in social groups, and pandemic zone classification (red, green, and orange based stringency of lockdown). Unlike conventional wisdom, we didn't find demographic variables, such as age, gender, and marital status, to be important determinants of income changes in our samples. Moreover, during our counterfactual exercise, that considered pandemic free period, changes in income was not strongly associated with these features.

7. Conclusion

There is considerable debate in the existing literature on the theoretical validity and empirical evidence of segmented labour market. The Indian labour market is also subject to such a debate with empirical evidence showing that there exists substantial difference in wage and human capital between the formal and the informal sectors. However, wage or gaps in human capital do not necessarily indicate segmentation of labour market unless there is a barrier to mobility of workers across sectors. In case of India, the absence of most labour laws in case of smaller firms and stringent labour laws which make hiring and firing costly for larger firms likely act as such a barrier. In this paper we analyse the Indian labour market around the economic shock caused by COVID using a unique rotational panel data to identify if there are any barriers to sectoral mobility which may accentuate the effect of the shock and make the sectoral difference persistent. We find that the proportion of workers who work in the informal sector increased. Additionally, the industry-wise distribution of the informal sector workers went through substantial changes. Labour market outcomes, both in terms of employment status and income, became even more divergent between these two sectors during the first wave of pandemic and in the subsequent quarters too. The loss of income was characterized more to the informal sector workers who experienced a change in their working status from salaried to not-working and self-employed to not-working. The income of the informal sector worker who remained self-employed were also adversely affected. Even during the recovery phase the segmentation in terms of labour market characteristics were visible.

The machine learning exercise confirms the findings of the exploratory data analysis that employment sector which indicate if a worker was employed in the formal or the informal sector, was an important predictor of income loss during COVID lockdown. In the recovery phase, the indicator was also an important predictor but to a lesser degree indicating that the workers who suffered the biggest shock due to COVID may not be the ones who recovered faster. This indicates uneven recovery. Moreover, we find that characteristics like education, family education, social groups were also important predictors of income fall (increase) during COVID lockdown (subsequent recovery). These variables may also play crucial role in allocating workers into formal and the informal sectors. District classification based on the strictness of lockdown was an important predictor for the income decrease during first wave, however they are not an important predictor for the recovery period indicating that regions where scarring was worse were not the ones which recovered faster. Overall, these findings indicate that scarring and recovery were both uneven and the sections which were affected most during the first wave COVID lockdowns were not the ones which recovered faster.

Reference

Abdul-Razak, S., & Sahoo, S. (2021, November 15). *Using the Rotational Panel Data from the Periodic Labour Force Survey: A Cautionary Tale*. Retrieved from SSRN: https://ssrn.com/abstract=3932499

Alstadsæter, A., Bratsberg, B., Eielsen, G., Kopczuk, W., Markussen, S., & Raaum, O. (2020). The first weeks of the coronavirus crisis: Who got hit, when and why? Evidence from Norway. *National Bureau of Economic Research Working Paper*, w27131.

Altmann, S., Falk, A., Grunewald, A., & Huffman, D. (2014). Contractual Incompleteness, Unemployment, and Labour Market Segmentation. *Review of Economic Studies*, 30-56.

Amirapu, A., & Gechter, M. (2020). Labor Regulations and the Cost of Corruption: Evidence from the Indian Firm Size Distribution. *The Review of Economics and Statistics*, 34-48.

Baek, C., McCrory, P. B., Messer, T., & Mui, P. (2020). Unemployment effects of stayat-at-home orders: Evidence from high frequency claims data. *Review of Economics and Statistics*, 1-72.

Baker, S. R., Farrokhnia, R. A., Meyer, S., Pagel, M., & Yannelis, C. (2020). How does household spending respond to an epidemic? Consumption during the 2020 COVID-19 pandemic. *The Review of Asset Pricing Studies*, 834-862.

Baqaee, D., & Farhi, E. (2020). Supply and demand in disaggregated Keynesian economies with an application to the COVID-19 crisis. *National Bureau of Economic Research Working Paper*, w27152.

Barrot, J. N., Grassi, B., & Sauvagnat, J. (2021). Sectoral effects of social distancing. *AEA Papers and Proceedings*, 277-281.

Beland, L. P., Brodeur, A., & Wright, T. (2020). COVID-19, stay-at-home orders and employment: Evidence from CPS data. *GLO Discussion Paper Series*, 559.

Besley, T., & Burgess, R. (2004). Can labor regulation hinder economic performance? Evidence from India. *The Quarterly Journal of Economics*, 91-134.

Bodenhorn, H. (2020). Business in a time of spanish influenza. *National Bureau of Economic Research Working Paper.*

Carletti, E., Oliviero, T., Pagano, M., Pelizzon, L., & Subrahmanyam, M. G. (2020). The COVID-19 shock and equity shortfall: Firm-level evidence from Italy. *The Review of Corporate Finance Studies*, 534-568.

Chakraborty, A. B. (2010). Studies in Indian informal sector. *Doctoral Dissertations. 540.*

Chetty, R., Friedman, J., Hendren, N., & Stepner, M. (2020). How did COVID-19 and stabilization policies affect spending and employment? a new real-time economic tracker based on private sector data. *NBER working paper*, w27431.

Cockx, B., & Ghirelli, C. (2016). Scars of Recessions in a Rigid Labor Market. *Labour Economics*, 162-176.

Correia, S., & Luck, S. (2022). Pandemics Depress the Economy, Public Health Interventions Do Not: Evidence from the 1918 Flu. *The Journal of Economic History*.

Dickens, W. T., & Lang, K. (1985). Testing dual labor market theory: a reconsideration of the evidence. *National Bureau of Economic Research Working Paper*, w1670.

Ferguson, N., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., & .... Ghani, A. C. (2020). Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. *Imperial College London*, 491-497.

Fernandez-Kranz, D., & Rodriguez-Planas, N. (2018). The Perfect Storm: Graduating During a Rcession in a Segmented Labour Market. *ILR Review*, 492-524.

Genda, Y., Kondo, A., & Ohta, S. (2010). Long-Term Effects of a Recession at Labor Market Entry in Japan and the United States. *The Journal of Human Resources*, 157-196.

Gourinchas, P. O., Kalemli-Özcan, Ṣ., Penciakova, V., & & Sander, N. (2020). COVID-19 and SME failures. *National Bureau of Economic Research Working Paper*, w27877.

Gupta, S., Montenovo, L., Nguyen, T. D., Lozano-Rojas, F., Schmutte, I. M., Simon, K. I., & Wing, C. (2020). Effects of social distancing policy on labor market outcomes. *NBER Working paper*, w27280.

Hasan, R., & Jandoc, K. R. (2012). Labor Regulations and the Firm Size Distribution in Indian Manufacturing. *School of International and Public Affairs Working Papers*, WP 1118.

Hsieh, C., & Klenow, P. (2009). Misallocation and manufacturing TFP in China and India. *The Quarterly Journal of Economics*, 124 (4).

Kochhar, K., Kumar, U., Rajan, R., & Subramanian, A. (2006). India's Patterns of Development: What Happened, What Follows. *International Monetary Fund Working Paper*, WP/06/22.

Leontaridi, R. M. (2002). Segmented Labour Markets: Theory and Evidence. *Journal of Economic Surveys*, 63-101.

Mazoomder, D. (1975). The Theory of Share-Cropping with Labour Market Dualism. *Economica*, 261-271.

Mehrotra, K. (2020, May 8). Explained: India enforced one of the strongest lockdowns, here's how it stacks up against other countries. *Indian Express*.

Mehrotra, S., Gandhi, A., Sahoo, B., & Saha, P. (2012). Creating employment in the twelfth five-year plan. *Economic and Political Weekly*, 63-73.

Mehrotra, S., Gandhi, A., Sahoo, B., & Saha, P. (2013). Turnaround in India's EmploymentStory: Silver Lining Amidst Joblessness and Informalisation? *Economic and Political Weekly*, 48(35).

Menon, R., & Nath, P. (2022). A dynamic analysis of Women's labour force participation in urban India. *The Economic and Labour Relations Review.*

Mongey, S., Pilossoph, L., & Weinberg, A. (2020). Which workers bear the burden of social distancing policies? *National Bureau of Economic Research Working Paper Series*, w27085.

Neog, B. J., & Kishore, B. (2016, March 15). *Labour Market Segmentation in the Indian Economy: An Investigation*. Retrieved from SSRN: https://ssrn.com/abstract=2728192

Neog, B. J., & Sahoo, B. K. (2016, December 1). *Segmentation and Mobility in the Indian Labour Market*. Retrieved from SSRN: https://ssrn.com/abstract=2877148

NSSO. (2016). Instructions to field staff, volume 1. Ministry of Statistics and Programme Implementation, Government of India.

Pratap, S., & Quintin, E. (2006). Are labor markets segmented in developing countries? A semiparametric approach. *European Economic Review*, 1817-1841.

Sastry, N. S. (2004). *Estimating informal employment and poverty in India.* New Delhi: United Nations Development Programme.

SBI Research. (2022). Share of informal economy may have shrunk to no more than 20% from 52% in FY18. *Ecowrap*(43).

Schivardi, F., & Romano, G. (2020). A simple method to estimate firms' liquidity needs during the COVID-19 crisis with an application to Italy. *CEPR*.