Employment growth/ skill requirement estimation in India: a non-traditional approach

Tutan Ahmed Assistant Professor, Vinod Gupta School of Management, IIT Kharagpur

Abstract

There is a remarkable lack of regular labour market data in the context of a developing country such as India. Given the lack of regular labour market surveys, a lack of labour market data in the informal sector and the geographical vastness of the country, it is almost impossible to obtain labour market data regionally and regularly. Consequently, there is barely any possibility of obtaining job market forecasts. With the emphasis on skill development initiatives in India, the need for linking skill development initiatives with the labour market is felt quite prominently. Within this context, an initiative has been undertaken in India to develop a job growth and skill requirement forecasting model. It is a data-driven model to be designed with multiple sets of data such as job advertisements in websites, proxy data at the district level, and Government Survey data. Machine learning techniques will be used for prediction of job growth and skill requirement growth. This job forecasting model is likely to be cost-effective, easily replicated across districts and a tool for providing the forecasts for job growth and skill requirement growth regularly and comprehensively.

JEL Codes: J2, C81, C83, C88
Keywords: labour market, India, job forecasting, multiple data sources, Hadoop, machine learning

The author is thankful to Prof. Pulak Ghosh from IIM Bangalore for his encouragement and advice on this work. Early stage support of Mr. Clement Chauvet, UNDP and Kapil Bendre, UNDP was valuable. Mr. Sagar Joshi and Ms. Aakriti Saini helped with data in the early stage of the exercise.

1. Problem statement

A critical issue for labour market information systems in developing countries is that there is a conspicuous absence of real-time information on available job opportunities in the market for policy practitioners (let alone individuals). With the recent formation by the Government of India (GoI) of the Ministry of Skill Development and Entrepreneurship and other State Skill Development Missions, this problem has attracted the attention of policy practitioners at the union as well as at the state government level.

Skill development is a complex business, as it is not confined to the realm of training, but also intricately linked with the labour market. If there is not enough demand for a particular skill in the market, irrespective of the skill level, an individual is unlikely to obtain commensurate value in the labour market. On the other hand, there could be unmet demand in the labour market for skilled manpower. Labour market demand and skill development activity can together be considered as the 'horse and cart', whereby labour market demand acts as the horse, or the driving force for all skilling activities. Knowledge of labour market demand in a real-time manner is therefore imperative for establishing skill development programs.

Policy practitioners are dependent on national household surveys for necessary labour market information. In India, national household surveys for employment/ unemployment are conducted less frequently (once every five years). Moreover, the sample size is often too small to analyse the labour market at a disaggregated level. For example, in India, labour market research is rarely conducted at a district level because of the small sample size (whereas some districts in India have population size more than that of Belgium). Other possible datasets such as those from the Employees' Provident Fund Organisation or Employees' State Insurance Scheme have their limitations regarding coverage and access. The GoIhas recently created a taskforce to address the lack of labour market information (Financial Express Bureau, 2017).

Other institutions in India,such as the Employment Exchange which was created for meeting the labour market information gap, have not been effective in connecting labour market supply and demand (Debroy, 2008). Moreover, the problem of the labour market information gap is compounded in India by the market's heterogeneous and informal nature. The GoI's Ministry of Labour& Employment launched the National Career Service (NCS) portal;however, it has yet toprove successful (Abraham and Sasikumar, 2018). The Ministry of Skill Development and Entrepreneurship has launched the National Labour Market Information System which, too,has yet to make any impact in capturing labour market demand (Abraham and Sasikumar, 2018). However, the issue of making reliable labour market information available to policymakers (and to people, subsequently) has still yet to be addressed.

The outcome of poor coordination between skill development programs and the labour market is manifested in the extremely low placement rate from the Ministry of Skill Development and Entrepreneurship's short-term training programs (Ahmed, 2016; Makkar, 2017). The Ministry and State Skill Development Missions are facing a critical challenge to estimate labour market demand so that these entities can align their skill development activities with existing demand. While some macro-

level predictions have been made in this regard, it is difficult to translate them into an implementation plan, particularly at the district level. The World Bank, in its Skill India Mission Operation project, has identified limited sources of relevant, frequently updated and appropriately disaggregated data to signal industry demand to training suppliers as a critical problem in the skilling ecosystem. A more decentralised, demand-driven approach to skills gap analysis is proposed at the district level.

2. Proposed solution

An initiative was developed (with the initial support of the United Nations Development Programme, India, to address the problem of data discrepancy in the Indian labour market. A concerted effort was made to develop a data-driven model for job/skill requirement forecasting at the district level. This exercise was based entirely on different datasets available for the labour market –for example, job advertisement data in online media, data from different departments of the GoI's District Administration, and survey/Census data. The model, which is an outcome of this exercise, will be more cost-effective compared with traditional survey methods. Moreover, this model can be easily replicated across districts. The GoI, multilateral organisations, private training providers and job-seekers are the targeted beneficiaries of the end product of this exercise. The model is being developed with data from Nagpur district in Maharashtra.

The proposed model will predict job growth for different sub-sectors (as per National Industrial Classification (NIC)) for the various job roles in the respective sectors of a given district. It will also identify skill requirements for respective job roles. These predictions will be on a real-time basis (that is, with a reasonable time interval, for example, one to two months), on a business intelligence tool platform.

3. Model architecture

The principle of collective intelligenceis the fundamental principle for building the model. A novel aspect of this work lies in the extraction and organisation of multiple sources of labour market data for further analysis.

The labour markets in developing countries are extremely heterogeneous, which creates a restriction in labour market analysis due to the lack of available data. However, access to multiple digital datasources, data extraction and data integration technology have made it possible to obtain labour market insights in a real-time manner at regional levels.

Following are the steps which have led to the data-driven model fordeveloping employment/skill requirement forecasting.

3.1 The principle of collective intelligence

This model is based on the assumption that In an extremely heterogenous labor market, different stakeholders in different sectors and sub-sectors have authentic but limited knowledge. . Particular individuals/data sources will be the source of insights for respective sectors. However, it is unlikely that a particular individual/ data source can provide detailed insights of the entire labour market.This leads to a problem similar to that in the parable of a group of blind men and the elephant,in which everyone

illustrates the elephant in their way of fragmented understanding of the totality. Whereas individual illustrations are insufficient in understanding the elephant, a comprehensive and systematic placement of these individual illustrations can provide a complete picture. A similar principle –the principle of collective intelligence – is adopted to address the labour market data intelligence problem.

As this work seeks to predict the employment forecast for an entire district as well as for different sectors of India's heterogenous labour market, comprehensive coverage of different datasets for the respective segments of the labour market is a key challenge. The following few paragraphs will explain the problem and outline the solution, with a delineation of the technology used for the necessary prediction work to follow.

3.2 Data integration framework

A critical aspect to solving the job forecasting problem is to tackle the heterogeneity of labour markets. A heterogeneous labour market has numerous sectors and sub sectors each of which has differences with the rest. It is a challenging exercise to map the data sources for each of the sub-sectors as per the NIC for a given district, and to adjust weights accordingly. Either economic Census or National Sample Survey (NSS) data for employment/unemployment can be used for this purpose. NSS data has been used because it provides more detail about types of employment along with an establishment's details. Existing research has used Labour Force Surveys (LFS) as a framework to capture online job market data (Štefánik, 2012b; Štefánik, 2012a; Jackson, 2007). LFS provide some details of the labour market –self-employment, wage employment, contractual employment, for example – which are important for building this model.

To incorporate different data sources, the NSS data framework on employment/ unemployment surveys has been used. According to the NSS, there is an approximate total of 2,000 sub-sectors of the economy (and therefore corresponding labour markets). These sub-sectors are obtained using the NIC (NIC-2008).[1] Employment details for all sub-sectors are provided in the NSS data.

The prominence of economic activities in these sub-sectors vary from region to region. For example, in Nagpur (the district for this study), as per the NSS 68th round, there are 151 sectors where jobs are available. Observations from other NSS rounds (for example, 66th round, 61st round) show a repetition of these sectors regarding job availability in Nagpur. As the sample size is small, the NSS has been adopted only as the framework and not for weight assignment for employment available in different sectors.

With this framework, multiple data sources are incorporated;for example, website data and district administration data. Multiple job portal data are used for forecasting jobs/skill requirements in the formal sector, and data from different district administrations are used to measure employment generation in the informal sector

1 The NIC is a standard developed by the GoI's Ministry of Statistics and Programme Implementation for the purpose ofclassifying different economic activities in India and maintaining a database for the same. For details, see http://mospi.nic.in/classification/national-industrial-classification.

through suitable proxies. The paragraphs below explain the different data sources available for this exercise.

With the availability of labour market data on the internet, the ways to use these data for labour market analysis are already established (Kureková, Beblavy and Thum, 2014; Askitas and Zimmermann, 2009). A large amount of labour market data is obtained using a web crawler for job availability and skill requirement analysis (Capiluppi and Baravalle, 2010; Jackson, Goldthorpe and Mills, 2005). However, most research into these categories is focused on a single website data source. In this exercise, data from multiple websites has been used to analyse job/skill requirement growth in the formal sector.

3.3 A matrix of data source to labour market segments: core of the solution

A critical aspect to solving the labour market data intelligence problem is to tackle the heterogeneity of the labour market with mapping data sources for each of the above sub-sectors for a given district (151 sub-sectors in the case of Nagpur). However, the mapping is required to be further disaggregated to capture skill requirement and employment forecast. As per the NSS, types of employment in each sector can be further disaggregated as: self-employment, regular/ salaried, casual (government and private). For different sub-sectors, these types of employment vary to a large extent. An illustration of this disaggregation is provided in A1. All together, there are 235 cells after mapping the sub-sector and types of employment.

Post the mapping withNSS data, a major assignment was to map each cell with the relevant job roles. Neither the NSS nor Census captures this data on job roles. The nearest approximation provided by the NSS is the National Classification of Occupations (NCO).[2] To obtain the details of job roles for each economic sub-sector and for each type of employment, NCO classification was added to the above framework. However, only classification and integration of the data sources isn't enough as field level insights are necessary to finalize different job roles to be kept in each cell of NIC-Employment Type-NCO. A consultation between district employment offices, district skill development offices, relevant district departments and the academics was conducted to validate job roles in each cell. For example, the Textile/Apparel sector in Nagpur comprises the following sub-sectors:preparation of cotton fibre (13111),[3] weaving and manufacturing of wool and wool-mixture fabrics (13123), manufacturing of knitted and crocheted synthetic fabrics (13913), manufacturing of all types of textile garments and clothing accessories (14101) and custom tailoring (14105) which are prominent in Nagpur. From A1, it is clear that self-employment is the only mode of employment for custom tailoring (14105), whereas weaving and manufacturing of wool and wool-mixture fabrics (13123) have regular/salaried employment as the only mode of employment. Once the sub-sector and type of employment data as contained in A1are obtained, experts work to incorporate dominant job roles for each of these

2 The National Classification of Occupations is a set of standards created and maintained by the GoI's Ministry of Labour&Employment. For details, see: https://labour.gov.in/sites/default/files/National%20Classification%20of%20Occupations_Vol%20II-B-%202015.pdf.

3 This is the five-digit National Industrial Classification.

sectors. The matrix for the mapped datasource to labour market segment has been prepared based on capturing this local knowledge. A sample of such is provided in A2.

3.4 Formal labour market data sources

Labour market coverage of online job portals is quite limited in India. The NCS portal had only .13 million active job vacancy postings across the country as on 2nd March 2019. On the other hand, every year, as per the 66th and 68th round of the NSS survey, the addition of labour force is around 4.68 million per year (considering usual principal and subsidiary status of employment)(Shaw, 2013). Thus, around one-tenth of the labour force addition is represented by the NCS portal.[4] However, at a regional level, there are multiple job portals such as monster.com, naukri.com, indeed.com, quikr.com, olx.com and urbanclap.com, as examples, which have an extensive coverage over and above the NCS portal. For example, as at 2nd March 2019, for the Nagpur District, the number of job vacancies posted in various portals were: naukri.com= 4241,indeed.com= 815, quick.com= 3432, monsterindia.com= 326, shine.com= 1084 and olx.com= 2146, with an average job-posting duration of one to two months. However, these portals are active mostly in the urban regions. In urban-centric employment prediction, these job portals are quite important (Maksuda, Ahmed and Nomura, in publication).

Newspapers are traditionally a rich source of job advertisement data (Jackson, Goldthorpe and Mills, 2005). Government jobs and some informal sector jobs (for example, private tutor, personal assistant) are traditionally advertised in print newsmedia. With the availability of online versions of print media, these advertisements are digitally available. Examples of these websites for the Nagpur district include: Nokarisandharbha, Employment Newspaper Maharashtra and Nagpur Today. Also, there is an emerging online labour market for the informal sector (for example, maid, driver, security guard, plumber) to cater to the urban areas. Leading websites in this regard are: babajob.com and olx.com, among others. These websites provide data about job demand for the present month or coming months for the various sectors.

3.5 Informal labour market data sources

As mentioned, the labour market in India is extremely heterogeneous, and there is no direct labour market data for the informal labour market. However, it is the informal sector which is dominant in India. A suitable 'proxy' is to be identified from the available district administration data to map employment in the informal sector.

Fortunately, district departments recently started maintaining data in digital format, and the extent of data availability is quite vast. With the availability of various datasets at Nagpur, it was observed that data available under different headings (for example, quantity of production, manpower involved, investment, capital, types of workers, establishment details, details of the infrastructure in the establishments and asset creation) were acting as suitable proxy inputs.

4 For Nagpur District (as at 2nd March, 2019), there were only 2 jobs posted. Link: https://www.ncs.gov.in/job-seeker/Pages/Search.aspx?OT=fheFJjl41aGWG85YSvGqng%3D%3D&OJ=sdm0Lg%2BxO9o%3D

For example, the District Employment Office (DEO) maintains details of employment data for the registered sector. By mapping NSS data to the data available with DEO, it is observed that the DEO data covers approximately the entire formal sector in Nagpur. Moreover, the District Industries Centre maintains data of the establishments at the individual level. This provides an opportunity to obtain input data which corresponds to the employment generation in the formal sector.

For different sectors/sub-sectors, there are multiple types of suitable proxies which are obtained from gleaning the district level data sources. For example, the District Transport Office provides the complete details of different types of vehicles purchased in the district (for example, auto-rickshaw, taxi, bus, jeep, station wagon, personal car).These are suitable proxies for employment generation in the future. Another example is employment generation in the animal husbandry sector. A full detail of production, employment, investment and asset creation in various sub-sectors – for example, poultry, cattle breeding –is available with the District Animal Husbandry department. Different Planning Departments maintain investment data across the departments/sectors or sub-sectors of employment. Mapping these investment and employment-growth details provides a rich source of proxy data and a basis for futuristic predictions. Similarly, for all different departments, different proxy data are being obtained, as these are helpful as a suitable proxy for employment prediction in respective sectors/sub-sectors.

Essentially, to work with the complex problem of labour market prediction, it is necessary to obtain different data sources and map them into the right place to complete the jigsaw puzzle of the labour market and facilitate the prediction of employment/skill-requirement growth with the help of necessary technologies.

3.6 Initial weight assignment for representativeness

It is essential that appropriate weights are allocated to each of these data sources, to take care of the representativeness aspect and therefore to generalise the results.[5] It is important to make a distinction between stock and flow concepts for employment data before explaining these datasets. Stock datasets could be explained as historical datasets, whereas flow datasets provide the most recent information about the labour market (say, for the past quarter). Below is an illustration of the existing datasets for stock and flow measurements. It is to be noted that the weight adjustment is calculated manually for the stock data only. Long periods of stock datasets are likely to provide a stable estimator. Following de Pedraza, Tijdens and de Bustillo (2007), weight is to be allocated applying the post-stratification method (for example, location-based, industry detail-based) to internet data and proxy data for different formal and informal sectors. Manual adjustment of the weights using the flow data would likely cause large fluctuations with little variation. Hence, no manual adjustment of weights is done for flow data. Moreover, a neural network is applied for the prediction exercise, and an essential attribute of neural net is a dynamic weight adjustment with several iterations. This is explained further in the following sections.

5 Neural Net adjusts the consequent weights which are discussed in the following sections.

Table 1: Employment Data Stock and Flow Measurement

	Online Job Market	*Administrative*	*Government Census/ Survey*
Stock	Data repository for job posting/ company registered for a few years (to be accessed after agreement with Job Board Company)	I. Repository of the establishments registered/vacancies posted by District Industrial Centre/ District Employment Office 1. II. Other proxy dataset for informal sector jobs	I. Economic Census (for the details of all establishments) 2. II. National Sample Survey for the adjustment of labour force in different employment type/ industry
Flow	Job posting for last quarter (obtained through web-scraping)	Both of the above sources to be obtained regularly from District Administration	None

3.7 Employment prediction to skill prediction

Many types of research have focused on obtaining the details of skill requirement along with employment requirement details with their predictions (Lenaerts, Beblavy and Fabo, 2016; Capiluppi and Baravalle, 2010; Kureková, Beblavy and Thum, 2014). With web crawling techniques, these research studies have gleaned vacancy data from different websites such as Burning Glass,Monster,EURES, among othersThe crawling technique provides a way to obtain the details of skill requirement specifications along with the specific employment requirement details. However, these details are restricted to individual websites, which are limited to an analysis of a specific segment of the labour market. The complexity level is higher in this present context, as skill requirement details are obtained from multiple websites as well as from proxy datasets. The data de-duplication method (explained below) addresses the complexity that arises from the usage of multiple website data. To specify the skill requirement corresponding to each type of employment in different sub-sectors, field-level data are also incorporated in the form of expert opinions. The details obtained are used as an input to the respective cell for skill details. This can be illustrated as follows. From previous examples of the textile sector, the growth of job roles in this sector is obtained from the NIC-Employment Type-NCO table. For necessary updates, district officials – business experts in this domain – are further consulted, and it is observed that sales and technician jobs in this sector are prominently growing in Nagpur. Thus, local knowledge is being incorporated into this collective intelligence gathering process.

3.8 Data extraction for formal sector labour market: crawling websites

Post the identification of these websites for obtaining job details for different sectors,a web crawler was developed to obtain data automatically with a certain periodicity from the websites as mentioned above. Following Capiluppi and Baravalle (2010), the ‘web

spider', a data extractor module, and an entity recognizer component was developed. The purpose of the web spider is to download vacancies, whereas the data extractor module is responsible for processing and categorizing raw data. Once extracted, the data was fragmented and parsed into smaller segments.The task of the crawler is to extract the most relevant keywords which appear in the job description. Examples of keywords include 'experience', 'education', 'salary', 'sector', 'work location' and 'date of the job advertisement'. Data crawling also identifies sectors for which job advertisements are available in websites (and sectors for which they are not). Naturally, this process is dependent on the existing structure and content of the data;that is,the way the job advertisements are posted on different websites. For example, some websites provide keywords for specific skill set requirements,while others provide details of the job description and the skill requirements are to be selected from the job descriptions. However, when extracting data from multiple websites, there is a high possibility of data duplication and variation of data formats. In fact, these are key challenges for web data extraction. How to address this problem is explained in the following sections.

3.9 Aspects of de-duplication and a'way out'

There are multiple challenges while performing a crawling exercise. A key problem is to obtain a comprehensive set of job advertisement details from multiple websites where the format, structure and wording of the advertisements differ. For example, naukri.com provides keywords for necessary skill requirements whereas other websites provide job descriptions and skill requirements are to be obtained from these descriptions. However, this is not a new problem, as many commercial organizations who routinely gather large databases in business and marketing analysis face this challenge regularly. The challenge here is to identify similar advertisements/job-postings and to determine whether they indicate the same advertisement/job-posting. The Sorted Neighborhood Method (SNM) is one such method used to address this issue. The fundamental problem here is that the data provided by various sources is 'string' in nature. Here, the equality of two values cannot be identified by having some arithmetic equivalence, but rather it requires a set of equational axioms to define the equivalence (Hernández and Stolfo, 1995). The technique used here is to partition the data pooled from different websites using the crawling technology. This pooled data is partitioned into clusters whereby each cluster will have potentially matching records. As proposed by Hernándezand Stoflo (1995), instead of pairwise matching multiple datasets, given it is expensive and time-consuming, clusters are formed, and then equational axioms are followed.

In this present exercise, there are eight to ten principal data sources (websites and online news media). Approximately 10,000 advertisements were obtained over three months (June to August2017), when advertisements from different websites were pooled. Small clusters were then created, following which a series of steps was performed to establish the equivalence of strings. These steps include spell corrections, main keyword check (for example, the title of a job advertisement, company name, experience, salary) and setting an acceptance rule(based on multi-pass or single-pass SNM).

3.10 Prediction using machine learning

To date, labour market/employment growth prediction has been restricted to the realm of linear extrapolation, whereby a fixed equation set by the author is the sole knowledge base for predicting employment growth (Hughes, 1991; Wang and Liu, 2009). However, the machine learning technique has an inherent advantage over other methods, as it learns from the data to adjust necessary weights in its intervening layers of regressions so as to provide the best prediction. The Artificial Neural Network (ANN) technique has an inherent advantage of correcting the prediction mechanism based on the data inputs it receives. The specific neural net used for correcting the prediction algorithm is known as the Back Propagation technique. The programmer is not required to set an equation establishing the relationship between a set of inputs (for example, production, investment, asset generation, the existing number of employees, future investment) and output (that is, employment and skill requirement). The ANN method establishes the relationship between inputs and output variable through the creation of a set of hidden layers/derived variables. Moreover, it continues to assign weights to the hidden layers/derived variables in the process of establishing a relationship between inputs and output variable.

This multiple layer creation is important in this context of employment prediction as the set of inputs and output (employment) enjoy a complex relationship. For each cell derived in the matrix for labour market segments (as explained earlier), the ANN is to run separately to derive the prediction of employment growth for each sub-sector/cell derived (as shown in A2). The prediction models thus developed are used for predicting future employment using the existing algorithm and other future data (for example, investment data which is obtained from the Planning Departments). However, the structure of the network is to be determined (for example,the number of intermediate layers), which is crucial for this exercise. The fundamental advantage of the ANN mechanism is that it keeps on improving the algorithm for each sub-sector with the usage of more data.

These available datasets for building the model are known as training data. Once the model is built with the available training data, it is ready to make predictions using future datasets. However, there is the critical problem of 'overfitting', which may induce significant errors in this prediction model. To address this, the following steps are adopted.

3.11 Validation

In building the model for employment/skill forecasting, there is a critical problem called 'overfitting'. As there are large numbers of categories of sub-sectors and employment type as derived in the matrix (see A1 and A2), a lesser number of observations is available for each category. When the training dataset is small, or the number of parameters in the model is large, there is a possibility that this model will not fit with the rest of the data, whereas it may fit with the training data quite well. This is overfitting.

To eliminate the problem of overfitting, cross-validation is used for choosing the right parameters in the model-building exercise. In the cross-validation process,

a part of the training data is kept separate to run and test the model. In this model, a 'k-fold' cross-validation exercise is to be performed with the test dataset being partitioned into k-subsets. On the other hand, rest of the data (outside of training) is used only to see how well the model that is obtained as an outcome of training and cross-fitting is performing.

3.12 Use of Hadoop platform

A key challenge in dealing with multiple sources of data of differing formats and structures is to sort, process and analyze the data for forecasting purposes. Also, there is a high level of calculation complexity in this forecasting process that necessitates the use of Apache Hadoop cluster information architecture. Hadoop can ingest data from multiple sources and can process data received on a different schedule (for example, web data is to be received more frequently than other proxy datasets from administration). Hadoop chops the data into smaller chunks and computes it through a parallel computation process which is extremely convenient for the existing exercise. Further, giventhe complexity of multiple web crawlers, SNM execution for data de-duplication, running multiple ANN processes for employment/skill requirement prediction for different sub-sector employment types (as per the grids shown in A2) and also the cross-validation process, theHadoop platform is ideal. Moreover, it can deliver insights into employment/skill requirement growth on a real-time basis, which is a key deliverable of this project. The programming language Python has been used for this exercise.

4. The output of the exercise, replicability of the model and regional aspects

As explained at the outset, the output of the exercise is the prediction of employment and skill requirement for each of the sub-sectors and employment type combinations (as shown in A2). These predictions are provided regularly (for example, with a certain interval of a few months) for a given region (Nagpur district in this case) on a business intelligence platform. Going forward, it is expected that this model of employment/skill forecasting will also be replicated in other districts. While most of the components of the data-driven model will remain the same for different districts, a separate exercise for each district will need to develop the Data Integration Framework, to develop the Matrix for Data Sources for Labour Market Segments and to obtain separate field details. The framework may vary from district to district depending on the presence of industry and jobs in the respective district. Further, the data sources may vary for different segments of the labour market in different districts. However, the software tools and methodology for estimating employment growth/skill requirement growth will remain the same, and this will save the cost of this estimation across the districts. Finally, the major problem of lack of employment information in India can be addressed with the use of this exercise, which precisely is the purpose of this model-building.

5. Conclusion

This work is aimed at exploring the possibility of predicting job-growth forecasting with the use of multiple data sources and machine learning techniques. It has explained the process of achieving such an outcome with associated steps to develop a data-driven forecasting model. This model should be a valuable addition to the existing pursuit of obtaining labour market data in a regular manner. Considering the complexity involved with the prediction of the heterogeneous labour market in Asian, African or Latin American countries, this data-driven model may be quite helpful. There has hardly been any information on the informal labour market in India. The use of proxy data, thanks to the digitization of the district administration data repository, would facilitate unlocking this lack of data availability for the informal sector. This model is capable of covering both formal and informal segments of the labour market comprehensively. It is also capable of providing regular job growth updates quite regularly and in a relatively inexpensive way (compared with the labour market survey). Moreover, the use of machine learning technology would ensure a higher accuracy in prediction, as the model is 'self-taught' with data and not merely a prediction which uses a pre-determined algebraic formula.

Table 2: A1Sub-sector and type of employment (manufacturing sector)

		Type of employment			
NIC	*Sub-sector*	*Self-employment*	*Salaried / regular*	*Casual (Govt.)*	*Casual (Private)*
10402	Manufacture of vegetable oils and fats, excluding corn oil	0	0	0	2,628
10613	Dahl (pulses) milling	0	0	0	5,325
13111	Preparation and spinning of cotton fibre including blendedcotton	0	19,584	0	674
13123	Weaving, manufacture of wool and wool-mixture fabrics.	0	2,543	0	0
13913	Manufacture of knitted and crocheted synthetic fabrics	0	2,685	0	0
14101	Manufacture of all types of textile garments and clothing accessories	4,096	0	0	0
14105	Custom tailoring	52,234	0	0	0
16101	Sawing and planing of wood	0	10,648	0	9,572
18112	Printing of magazines and other periodicals, books and brochures	0	2,139	0	0
20238	Manufacture of 'agarbatti' and other preparations	0	0	0	2,023
24109	Manufacture of other basic iron and steel	0	1,307	0	0
24319	Manufacture of other iron and steel casting and products	0	418	0	0
25112	Manufacture of metal frameworks or skeletons for construction	312	0	0	0
25119	Manufacture of other structural metal products	792	0	0	0
25121	Manufacture of metal containers for compressed or liquefied gas	0	2,865	0	0
25933	Manufacture of hand tools such as pliers, screwdrivers, press tools	391	0	0	0
25994	Manufacture of metal household articles	0	1,114	0	0
25999	Manufacture of other fabricated metal products	318	0	0	0
26209	Manufacture of computers and peripheral equipment	0	1,711	0	0
28213	Manufacture of spraying machinery for agricultural use	0	3,583	0	0
29102	Manufacture of commercial vehicles such as vans, lorries	0	2,377	0	0
31001	Manufacture of furniture made of wood	4,608	0	0	0

Table 3: A2Industry sub-sector, employment type and job roles in textile sector (NIC employment type – NCO)

		Self-employment job roles	*Salaried/ wage employment job roles*	*Casual (Govt.)*	*Casual (Private)*
13111	Preparation and spinning of cotton fibre including blended cotton	Home-based yarn manufacturing	Spinning Shift Officer, Weaving Supervisor, Finishing	NA	Contracted out jobs for the workers
13123	Weaving, manufacture of wool and wool mixture fabrics	NA	Marketing,	NA	NA
13913	Manufacture of knitted and crocheted synthetic fabrics	Home-based knitting	Yarn Dying	NA	NA
14101	Manufacture of all types of textile garments and clothing accessories	Home-based/sub-contracted	Polyester Staple Fibre Manufacturer, Sales, Technicians	NA	NA
14105	Custom tailoring	Home-based/self entrepreneurial tailoring	Marketing, Technician	NA	NA

References

Abraham, V. and Sasikumar, S.K.(2018),'Labour Market Institutions and New Technology: the Case of Employment Service in India', *The Indian Journal of Labour Economics, 61*(3), 453-471.

Ahmed, T. (2016),'Labour Market Outcome for Formal Vocational Education and Training in India: Safety Net and Beyond', *IIMB Management Review, 28*(2), 98-110.

Askitas, N. and Zimmermann, K.F. (2009),'Google Econometrics and Unemployment Forecasting', *Applied Economics Quarterly, 55*(2), 107-120.

Brewer, E., Demmer, M., Du, B., Ho, M., Kam, M., Nedevschi, S., Pal, J., Patra, R., Surana, S. and Fall, K.(2005), 'The Case for Technology in Developing Regions', *Computer, 38*(6), 25-38.

Capiluppi, A. and Baravalle, A. (2010), 'Matching Demand and Offer in On-line Provision: aLongitudinal Study of monster.com', in *12th IEEE International Symposium on Web Systems Evolution*, WSE 2010, 13-21.

Debroy, B. (2008),'India's Employment Exchanges: Should they be Revamped or Scrapped Altogether?', in *ISAS Insights*, Institute of South Asian Studies,Singapore.

de Pedraza, P., Tijdens, K.G. and de Bustillo, R.M. (2007), *Sample Bias, Weights and Efficiency of Weights in a Continuous Web Voluntary Survey*, Amsterdam Institute for Advanced Labour Studies, University of Amsterdam, Amsterdam.

.10 May ,FLCRED.

Hernández, M.A. and Stolfo, S.J. (1995), 'The Merge/Purge Problem for Large Databases', *ACM SIGMOD Record, 24*(2), 127-138.

Hughes, G. (1991),*Manpower forecasting: A Review of Methods and Practice in Some OECD Countries*, Report No. 1, FÁS, Dublin.

Jackson, M. (2007),'How Far Merit Selection? Social Stratification and the Labour Market', *The British Journal of Sociology, 58*(3), 367-390.

Jackson, M., Goldthorpe, J.H. and Mills, C.(2005),'Education, Employers and Class Mobility', *Research in Social Stratification and Mobility*, 23, 3-33.

Kureková, L.M., Beblavy, M. and Thum, A.E. (2014),'Using Internet Data to Analyse the Labour Market: a Methodological Enquiry', in *IZA Discussion Papers 8555*, IZA – Institute of Labour Economics, Bonn.

Lenaerts, K., Beblavy, M. and Fabo, B.(2016),'Prospects for Utilisation of Non-vacancy Internet Data in Labour Market Analysis – an Overview', *IZA Journal of Labor Economics, 5*(1), 1.

Makkar, S. (2 September 2017),'Why India's Skill Mission has Failed', in *Business Standard*, New Delhi.

Maksuda, N., Ahmed, T., Nomura, S. (in publication),'What Skills are Demanded Today in Pakistan? A Job Market Seen Through a Lens of an Online Job Portal – ROZEE.PK', The World Bank Discussion Paper Series.

Patel, N., Klemmer, S.R. and Parikh, T.S. (2011), 'An Asymmetric Communications Platform for Knowledge Sharing with Low-end Mobile Phones, in *Proceedings of the 24th annual ACM symposium adjunct on User Interface Software and Technology*, 16-19 October, Santa Barbara, 87-88, ACM, New York.

Shaw, A. (2013),'Employment Trends in India', *Economic & Political Weekly, 48*(42), 23.

Štefánik, M. (2012) (a), 'Internet Job Search Data as a Possible Source of Information on Skills Demand (with results for Slovak university graduates)', in *Building on Skills Forecasts –Comparing Methods and Applications*, 246.

Štefánik, M.(2012)(b),'Focused Information on Skills Demand Using Internet Job Search Data (with results for Slovak university graduates)', in *Conference Proceedings from the 11th Comparative Analysis of Enterprise Data & COST Conference 2012*, Institute of Economic Research, Slovak Academy of Sciences, Bratislava.

Wang, X. and Liu, Y. (2009), 'ARIMA Time Series Application to Employment Forecasting',in *Proceedings of 2009 4th International Conference on Computer Science & Education*, IEEE, New Jersey.