Employment & Skill Requirement Forecasting for Regional Labor Market in India: Methodology with Traditional and Non-traditional Data

Tutan Ahmed[1]

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 survey, 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. Naturally, there is hardly any possibility of obtaining job market forecasts. With the emphasis on skill development initiatives in India, the need for linking skill development initiatives with labour market is felt quite prominently. With this context, an initiative is taken 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, etc. Machine learning techniques are to be used for prediction of job growth and skill requirement growth. This job forecasting model is likely to be cost-effective, easily replicable across districts and a tool for providing the forecasts for job growth and skill requirement growth regularly and comprehensively.

Key Words: Labour Market, India, Job-Forecasting, Multiple data sources, Hadoop, Machine-Learning

1. Problem Statement:

2.

A critical issue of labor market information system in developing countries is that there is a conspicuous absence of real-time information on available job opportunities in the market to the policy practitioners (let alone individuals). With the formation of the new Ministry of Skill Development & Entrepreneurship of Government of India and other State Skill Missions, this problem has received 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 just confined to the realm of training but also quite intricately linked with the labor market. If there is not enough demand for particular skill in the market; irrespective of the skill level, an individual is unlikely to get commensurate

[1] Faculty at Vinod Gupta School of Management IIT Kharagpur

value in the labor market. On the other hand, there could be unmet demand in the labor market for the skilled manpower. The labor market demand and skill development activity can together be considered as the horse and cart duo where labor market demand acts as the horse or the driving force for all skilling activities. This makes the knowledge of labor market demand in a real-time manner imperative for setting up skill development programs.

Policy practitioners are dependent upon national household surveys for necessary labor market information. In India, national household survey for employment/unemployment is conducted less frequently (once in 5 years). Moreover, often the sample size is too small to analyze the labor market at a disaggregated level. For example, in India, hardly any labor market research conducted at a district level because of the small sample size (whereas some districts in India has a size more than that of Belgium). Other possible datasets like the Employee Provident Fund (EPF) or Employee State Insurance (ESI) have their limitation regarding coverage and access. The government of India has created a task force recently to address the issue of the labor market information unavailability (Financial Express Bureau, 2017).

The other institutions like Employment Exchange which is created for meeting the labor market information gap have not been quite efficient to connect the labor market supply and demand (Debroy, 2008). Moreover, the problem of labor market information gap is compounded in India given the heterogeneous and informal nature of the labor market. Ministry of Labor, Government of India has launched National Career Services (NCS) Portal which hasn't been successful yet (Abraham, Sasikumar; 2017). Ministry of Skill Development, Government of India, has launched the National Labor Market Information System (LMIS) which is yet to make any impact to capture labor market demand (ibid.). Some efforts are made in some of the leading institutes in the United States and India (Patel et al, 2011, Brewer et al, 2005) to bring digital technology solution to address the last mile information delivery problem. However, making reliable labor market information available to policymakers (and to people - subsequently) is still unaddressed.

The outcome of poor coordination between skill development programs and the labor market is manifested in the extreme low placement rate from the short-term training programs of the Ministry of Skill Development (GoI) (Ahmed, 2016; Makkar, 2017).

Ministry of Skill Development & Entrepreneurship and other State Skill Missions are facing a critical challenge to estimate labor market demand so that these entities can align their skill development activities with the existing demand. Whereas some macro-level predictions are made

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

2. Proposed Solution:

An initiative was taken (with the initial support of the United Nations Development Program (UNDP), India) to address the problem of data discrepancy in the Indian labor market. A concerted effort was made to develop a data-driven model for job/skill requirement forecasting at the district level. This exercise was entirely based on different datasets available for the labor market - e.g., job advertisement data in online media, data from different departments of the District Administration, and survey/ CENSUS data. The model, which is an outcome of this exercise, is going to be more cost-effective as compared to the traditional survey method. Also, this model will be easily replicable across districts. The government, multilateral organizations, 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 model will predict of job growth for different sub-sectors (as per National Industrial (NIC) classification) for the different job roles in the respective sectors in a given district. It will also bring out skill requirements for respective job roles. These predictions will be on a real-time basis (i.e., with a reasonable time interval, e.g., one/two months) on a Business Intelligence Tool platform.

3. Model Architecture:

In the following paragraphs, the architecture of the model is explained. The principle of collective intelligence, the fundamental principle for building this model is explained in following paragraphs. A novel aspect of this work lies in the extraction and organization of multiple sources of labor market data for further analysis.

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

Followings are the steps in this exercise which leads to the data-driven model building for employment/ skill requirement forecasting.

3.1. The principle of Collective Intelligence:

This model building is based on the assumption that stakeholders in the labor market have their share of truth. Particular individuals/ data sources can share insights of the respective sector. However, it is unlikely that a particular individual has all the insights of the labor market.

It leads to a problem similar to that of the parable of a group of blind men and the elephant where 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. Similar principle, i.e. principle of collective intelligence, is adopted to address this labor market data intelligence problem.

As this work is braving the challenge of predicting the employment forecast for the entire district as well as for different sectors of the heterogenous labor market, comprehensive coverage of different datasets for the respective segments of the labor market is a key challenge. Next few paragraphs explain the problem and the way out. Following this, the technology used for the necessary prediction work is delineated.

3.2. Data Integration Framework:

A critical aspect to solve the job forecasting problem is to tackle the heterogeneity of labour markets with the use of mapping data sources for each of the above sub-sectors for a given district and adjusting weights consequently. Either Economic Census or National Sample Survey data for employment/ unemployment can be used for this purpose. National Sample Survey data used as it provides more details about types of employment along with the establishment details. Some researches have already used Labour Force Survey (LFS) as a framework to cover online job market data (Štefánik 2012b; Štefánik 2012a; Jackson 2007). Moreover, LFS provides some of the details of the labor market i.e. self-employment, wage employment, contractual employment etc. which are important for building this model).

To incorporate different data sources, National Sample Survey (NSS) data framework on employment-unemployment surveys is used. As per NSS, there are a total of around 2000 sub sectors of the economy (and therefore corresponding labor markets). These sub-sectors of the economy is obtained using National Industrial Classification (NIC-2008)[2]. Employment details for all these sub-sectors are provided in National Sample Survey.

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

With this framework, multiple data sources are incorporated, e.g., website data, district administration data, etc. Multiple job portal data are used for forecasting of jobs/ skill requirement in the formal sector, and data from different district administration are used to measure employment generation in the informal sector through suitable proxies. Paragraphs below explain different data sources available for this exercise.

With the availability of labor market data in the internet, ways to use these data for labor market analysis are already in place (Kurekova et al, 2014; Askitas, Zimmermann, 2009). A large amount of labor market data is obtained using crawler for job availability and skill requirement analysis (Capiluppi, Baravalle, 2014; Jackson et al, 2005). However, most research works of these categories are focused on a single website data. In this exercise, data from multiple websites is used to analyze job/ skill requirement growth in the formal sector.

3.3. A Matrix of Data Source to Labor Market Segments: Core of the Solution

A critical aspect to solve the labor market data intelligence problem is to tackle the heterogeneity of labor market with the use of 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 NSS, types of

---

[2] It is an standard developed by the Ministry of Statistics and Program Implementation (Government of India) with the purpose of classifying different economic activities in India and maintaining a database for the same. For details: http://mospi.nic.in/classification/national-industrial-classification

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 Annexure - I. Altogether, there are 235 cells after mapping the sub-sector and types of employment.

Post this mapping with NSS data; a major assignment is to map each cell with the relevant job roles. Neither NSS nor CENSUS captures this data on job roles. The nearest approximation provided by NSS is National Code of Occupation (NCO)[3]. To obtain the details of the job roles for each economic sub-sectors and for each type of employment, NCO classification is added to the above framework. Further, for each cell of NIC-Employment Type- NCO, a consultation between district employment office, district skill development office, relevant district departments and the group of experts is conducted to validate job roles in each cell. For example, for Textile/Apparel sector in Nagpur, following sub-sectors - Preparation of cotton fibre (13111[4]), 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), Custom Tailoring (14105) are prominent in Nagpur. From Annexure- I, it is visible that self-employment is the only mode of employment for Custom Tailoring (14105), whereas Weaving and Manufacturing of wool and wool mixture fabrics (13123) has regular/ salaried employment as only mode of employment. Once, the Annexure - I is obtained, experts have worked on to incorporate dominant job roles for each of these sectors. Based on capturing of this local knowledge, the Matrix for the mapped datasource to labor market segment is prepared. A sample of such is provided at Annexure - II.

3.4. Formal Labor Market Data Sources:

Labor market coverage of online job portals is quite limited in India. National Career Services (NCS) Portal has a total of .48 million job vacancy posting for across the country for the year 2016-17. On the other hand, every year, as per the 66th and 68th round of NSS survey, the addition of labor force is around 4.68 Million per year (considering usual principal and subsidiary status of employment)[5]. Thus, around 1/10th of the labor force addition is represented by NCS

---

[3] National Code of Occupation is a set of standards created and maintained by the Ministry of Labor and Employment (Government of India). For details: https://labour.gov.in/sites/default/files/National%20Classification%20of%20Occupations_Vol%20II-B-%202015.pdf

[4] This is five digit National Industrial Classification.

[5] Shaw, A. (2013). Employment trends in India. *Economic & Political Weekly*, *48*(42), 23.

portal[6]. However, at a regional level, there are multiple job portals like monster.com, naukri.com, indeed.com, quikr.com, olx.com, urbanclap.com, etc. which has an extensive coverage over and above NCS portal. For example, as on 28th August 2017, for Nagpur District, job vacancy posted in different portals are: naukri.com: 2068,indeed.com: 1060, quick.com: 2904, monsterindia.-com: 455, shine.com: 2038, olx.com: 1073, with an average job posting duration of one to two months. However, these portals have their presence mostly in the urban regions. In urban-centric employment prediction, these job portals are quite important (Maksuda et al, in publication process).

Newspapers are traditionally a rich source of job advertisement (Jackson et al, 2005) data. Government jobs, some informal sector jobs (e.g., Private Tutor, Personal Assistant, etc.) are traditionally advertised in print news-media. With the availability of online versions of print media, these advertisements are digitally available. Some of these websites for Nagpur District are (https://nokarisandharbha.wordpress.com/, http://employment-newspaper.com/maharashtra/, http://www.nagpurtoday.in/nagpur-jobs). Also, there is an emerging online labor market for the informal sector (e.g., maid, driver, security guard, plumber, etc.) for catering to the urban areas. Leading websites in this regard are - babajob.com, olx.com, etc. These different websites provide data about job demand of present month or coming months for the various sectors.

3.5. Informal Labor Market Data Sources:

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

Fortunately, different district departments have started maintaining data in digital format recently, and the extent of data availability is quite vast. With the availability of various data sets at Nagpur, we observe that data available under different heads (e.g., quantity of production, manpower involved, investment, capital, types of workers, establishment details, details of the infrastructures in the establishments, asset creation, etc.) are acting as suitable proxy inputs.

---

[6] For Nagpur District (as on 7th October, 2017), there are only 18 jobs posted.

Source: https://www.ncs.gov.in/job-seeker/Pages/Search.aspx?lm=igQpyimn46SdE2OoYkwMeSRZgiRZFRKZPOd9ZlJqnlv3djPspH9Ex61hK1ff%2FsxsXYgG4xXXYQ4%3D&OT=fheFJjl41aGWG85YSvGqng%3D%3D&Source=https://www.ncs.gov.in/Pages/default.aspx

For example, District Employment Office (DEO) maintains a detail 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, District Industrial Center (DIC) 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 is obtained from gleaning the district level data sources. For example, District Transport Department provides the complete details of different types of vehicles purchased in the District (e.g., Auto-Rickshaw, Taxi, Bus, Jeep, Station Wagon, Personal Car, etc.).These are suitable proxies for the employment generation in the future. Another example is employment generation in Animal Husbandry sector. A full detail of production, employment, investment, asset creation in various sub-sectors, e.g., poultry, cattle breeding, etc. are available with the District Animal Husbandry department. District Planning Department maintains different investment data across the departments/ sectors or sub-sectors of employments. 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 suitable proxy for employment prediction in respective sector/ sub-sector.

Essentially, to work with a complex problem of labor market prediction, it is required to obtain different data sources and map them into the right place to complete the jigsaw puzzle of the labor 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, initially[7], to bring representativeness characteristics and therefore to generalize the results. It is important to bring a distinction between stock and flow concept for employment data before explaining these datasets. Stock datasets could be explained as the historical datasets whereas flow dataset provides most recent information about the labor 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

[7] Neural Net adjusts the consequent weights which is discussed in following sections.

adjustment is manually done for the stock data only. Long period of stock datasets are likely to provide a stable estimator. Following Pedraza, Tijdens, and Muñoz de Bustillo, Steinmetz (2007), weight is to be allocated using post-stratification method (e.g., location-based, industry detail based, etc.) to internet data and proxy data for different formal and informal sectors. Manual adjustment of the weights using the flow data would likely to make large fluctuation with little variation. Hence, no manual adjustment of weights is done for flow data. Moreover, neural network is applied for the prediction exercise and an essential attribute of neural net is to dynamic weight adjustment with several iterations. It is explained further in the following sections.

Employment Data Stock & Flow Measurement			
	Online Job Market	Administrative	Government CENSUS/ Survey
Stock	Data repository for job posting/ company registered for 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 II. Other Proxy dataset for Informal Sector Jobs	I. Economic Census (for the details of all establishments) II. National Sample Survey for the adjustments of labor 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 to obtain the details of skill requirement along with employment requirement details along with their predictions (Lenaerts et al, 2016; Capiluppi, Baravalle; 2010; Kurekova et al, 2014). With web crawling techniques, these works have gleaned vacancy data from different websites like “Burning Glass,” “Monster,” “EURES” etc. Crawling technique provides a way to obtain the details of skill requirement specifications along with the

specific employment requirement details. However, all these works are restricted to individual websites which are limited to an analysis of a specific segment of the labor market. Complexity level is higher in this present context as skill requirement details are obtained from multiple websites as well as form proxy datasets. Data de-duplication method (explained below) takes care of the complexity that arises from the usage of multiple website data. For specifying skill requirement corresponding to each type of employment in different sub-sectors, field level data are also incorporated in the form of expert opinions. Obtained details are used as an input to the respective cell for skill details. An illustration is as follows. From previous examples of textile sector, the growth of job roles in this sector is obtained from NIC-Employment Type-NCO table. For necessary updates, district officials, business experts in this domain are further consulted and it is observed that sales, 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 Labor Market: Crawling Websites

Post the identification of these websites for obtaining job details for different sectors; web crawler is developed for obtaining data automatically with a certain periodicity from the websites as mentioned above. Following Capiluppi & Baravalle (2010), the web spider, a data extractor module, and an entity recognizer component is developed. Purpose of 'web spider' is to download vacancies whereas the "data extractor" module is responsible for processing and categorizing raw data. Once extracted, the data is fragmented and parsed into smaller segment.The task of the crawler will be to extract the most relevant "keywords" which appear in the job description. These keywords are like experience, education, salary, sector, work location, date of the job advertisement, etc. Data crawling will also identify sectors for which job advertisements are available in websites (and sectors for which it isn't). Naturally, this process is dependent upon the existing structure and content of the data i.e. upon the way the job advertisements are posted in different websites. For example, some websites provide keywords for specific skill set requirements whereas others provide a detail of the job description and the skill requirements are to be picked from the job descriptions.

However, while 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. The way to address this problem is explained in the following sections.

3.9. Aspects of De-duplication and Way Out:

There are multiple challenges while performing 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 are different. For example, naukri.com provides keywords for necessary skill requirement 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-posting and to figure out if these are indicating the same advertisement/ job-posting. Sorted Neighborhood Method (SNM) is one such method used to address this problem. The fundamental problem here is that the data provided by various sources is "String" in nature. Here, equality of two values can't be identified by having some arithmetic equivalence, but it requires a set of equational axioms to define the equivalence (Hernández, Stolfo, 1995). The technique being used here is to partition the data pooled out from different websites using the crawling technology. This pooled data is partitioned into clusters where each cluster will have potentially matching records. As proposed by Hernández & Stoflo (1995), instead of pairwise matching of multiple datasets, as that is hugely expensive and time-consuming, clusters are formed, and then equational axioms are followed.

In this present exercise, we have 8-10 principal data sources (websites and online news media). An approximate of around 10,000 advertisements are obtained for three months (June- August 2017) when advertisements from different websites are pooled. Small clusters are created following which a series of steps are performed to establish the equivalence of strings. These steps are - spell corrections, main keyword check (e.g., the title of job advertisement, company name, experience, salary, etc.), setting an acceptance rule, etc. (based on multi-pass or single-pass SNM).

3.10. Prediction using Machine Learning:

Till date, labor market/ employment growth prediction has been restricted to the realm of linear extrapolation where a fixed equation that was set by the author used to be the sole knowledge base for predicting employment growth (Hughes, 1991; Wang & Liu, 2009). However, machine learning technique has an inherent advantage above 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. Artificial Neural Network (ANN) technique has an inherent advantage of correcting the

prediction mechanism based on the data inputs it receives. Specific neural net used for correcting the prediction algorithm is known as Back Propagation technique. The programmer is not required to set an equation establishing the relationship between a set of inputs (Production, Investment, Asset generation, the existing number of employees, future investment, etc.) and output (i.e., employment and skill requirement). 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 labor market segments (as explained earlier), ANN is to run separately to derive the prediction of employment growth for each sub-sector/ cell-derived (as it is in Annexure - II). The prediction models thus developed are used for prediction of future employment using the existing algorithm and other future data (e.g. investment data which is obtained from District Planning Department). However, the structure of the network is to be determined (e.g., number of intermediate layers, etc.) which is crucial for this exercise. The fundamental advantage of 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 is 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 a critical problem of “overfitting” which may induce significant error in this prediction model. To address it, following steps are adopted.

3.11. Validation:

In this model building for employment/ skill forecasting, there is a critical problem called “overfitting. As there are large numbers of categories of sub-sector and employment type as derived in the matrix (Appendix - I and Appendix - II), less number of observations are available for each category. When the training data set is small, or the number of parameters in the model is large, there is a possibility that this model is not going to fit with the rest of the data whereas it may fit with the training data quite well.

It is called “overfitting”. To get rid of this “overfitting” problem, cross-validation is used for choosing the right parameters in the model building exercise. In “Cross-validation” process, a part of the training data is kept separately 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 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 to deal multiple sources of data of different formats and structures is to sort, process, and analyze the data for forecasting purpose. Also, there is a high level of calculation complexity in this forecasting process that necessitates the use of Apache Hadoop cluster information architecture. As Hadoop can ingest data from multiple sources and it can process data received on a different schedule (e.g., web data is to be received more frequently than other Proxy datasets from Administration). Hadoop chops the data into smaller chunks and computes data through parallel computation process which is extremely convenient for the existing exercise. Moreover, the complexity of multiple web crawling, SNM execution for data de-duplication, running multiple ANN processes for employment/ skill requirement prediction for different sub-sector- employment type (as per the grids of Appendix-II) and also cross-validation process; Hadoop platform is ideal. Moreover, it can deliver insights of employment/skill requirement growth on a real-time basis which is a key deliverable of this project. Programming language Python is used for this exercise.

4. The output of the Exercise, Replicability of the Model and Regional Aspects:

As explained at the beginning, the output of the exercise is the prediction of employment and skill requirement for each of the sub-sectors and employment type combination (as per Annexure II). These predictions are provided regularly (e.g., with a certain interval of few months) for a given region (Nagpur district in this case) on a Business Intelligence platform. Going ahead, it is expected that this model of employment/ skill forecasting be replicated in other districts too. While most of the components of the data-driven model building will remain same for different districts, separate exercise for each district will require to develop the Data Integration Framework, to develop the Matrix for Data Sources for Labor Market Segments and to obtain separate field details. The framework may vary from district to district depending upon the presence of industry and jobs in the respective district. Also, the data sources may vary for different segments of the labor market in different districts. However, the software tools and methodology for esti-

mation of employment growth/ skill requirement growth will remain 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 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 turn out to be quite helpful. There has hardly been any information on the informal labour market in India. Use of proxy data, thanks to the digitization of district administration data repository, would facilitate unlocking this problem of lack of availability of data 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 (as 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 that uses a pre-determined algebraic formula.

References:

1. Financial Express Bureau (2017). "Task force to look into creation of reliable employment data base", Financial Express, New Delhi, May, 10

2. 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), pp. 453-471.

3. Patel, N., Klemmer, S.R. and Parikh, T.S., 2011, October. 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 (pp. 87-88). ACM.

4. 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), pp.25-38.

5. Kureková, L.M., Beblavy, M. and Thum, A.E., 2014. Using internet data to analyse the labour market: a methodological enquiry (No. 8555). IZA Discussion Papers.

6. Askitas, N. and Zimmermann, K.F., 2009. Google econometrics and unemployment forecasting. Applied Economics Quarterly, 55(2), pp.107-120.

7. Lenaerts, K., Beblavý, 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), p.1.

8. Capiluppi, A. and Baravalle, A., 2010, September. Matching demand and offer in on-line provision: A longitudinal study of monster. com. In Web Systems Evolution (WSE), 2010 12th IEEE International Symposium on (pp. 13-21). IEEE.

9. Jackson, M., Goldthorpe, J.H. and Mills, C., 2005. Education, employers and class mobility. Research in social stratification and mobility, 23, pp.3-33.

10. Hernández, M.A. and Stolfo, S.J., 1995, June. The merge/purge problem for large databases. In ACM Sigmod Record (Vol. 24, No. 2, pp. 127-138). ACM.

11. Ahmed, T., 2016. Labour market outcome for formal vocational education and training in India: Safety net and beyond. IIMB Management Review, 28(2), pp.98-110.

12. Makkar, S (2017). “Why India's skill mission has failed”. Business Standard, New Delhi, September, 2

13. Maksuda, N., Ahmed, T., Nomura, S. (under publication process) 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

15. Hughes, G., 1991. Manpower forecasting: A review of methods and practice in some OECD countries (No. 1). FÁS.

16. Wang, X. and Liu, Y., 2009, July. ARIMA time series application to employment forecasting. In Computer Science & Education, 2009. ICCSE'09. 4th International Conference on (pp. 1124-1127). IEEE.

17. 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.

18. Debroy, B., 2008. India's Employment Exchanges: Should They be Revamped Or Scrapped Altogether?. Institute of South Asian Studies.

19. Jackson, M., 2007. How far merit selection? Social stratification and the labour market 1. The British journal of sociology, 58(3), pp.367-390.

20. Štefánik, M.(a), 2012. Internet job search data as a possible source of information on skills demand (with results for Slovak university graduates). Building on skills forecasts—Comparing methods and applications, p.246.

21. Štefánik, M., 2012 (b). Focused information on skills demand using internet job search data (with results for Slovak university graduates). Institute of Economic Research, Slovak Academy of Sciences.

22. Shaw, A., 2013. Employment trends in India. Economic & Political Weekly, 48(42), p.23.

Annexure I: Sub-Sector and Type of Employment (Manufacturing Sector)

		Type of Employment			
NIC	Sub-Sector	Self-Employ-ment	Salaried/ Regular	Casual Govt	Casual Pvt.
10402	Manufacture of vegetable oils and fats excluding corn oil	0	0	0	2628
10613	Dal (pulses) milling	0	0	0	5325
13111	Preparation and spinning of cotton fiber including blended* cotton	0	19584	0	674
13123	Weaving, manufacture of wool and wool mixture fab-rics.	0	2543	0	0
13913	Manufacture of knitted and crocheted synthetic fab-rics	0	2685	0	0
14101	Manufacture of all types of textile garments and clothing accessories	4096	0	0	0
14105	Custom tailoring	52234	0	0	0
16101	Sawing and planing of wood	0	10648	0	9572
18112	Printing of magazines and other periodicals, books and brochures	0	2139	0	0
20238	Manufacture of “agarbatti” and other preparations	0	0	0	2023
24109	Manufacture of other basic iron and steel	0	1307	0	0
24319	Manufacture of other iron and steel casting and prod-ucts	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	2865	0	0
25933	Manufacture of hand tools such as pliers, screw-drivers, press tools	391	0	0	0
25994	Manufacture of metal household articles	0	1114	0	0
25999	Manufacture of other fabricated metal products	318	0	0	0
26209	Manufacture of computers and peripheral equipment	0	1711	0	0
28213	Manufacture of spraying machinery for agricultural use	0	3583	0	0
29102	Manufacture of commercial vehicles such as vans, lorries	0	2377	0	0

31001	Manufacture of furniture made of wood	4608	0	0	0

Annexure II: Industry 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 (Pvt.)
13111	Preparation and spinning of cotton fiber including blended* cotton	Home based yearn 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, Yarn Dying	NA	NA
13913	Manufacture of knitted and crocheted synthetic fabrics	Home based knitting	Polyester Staple Fibre Manufacturer, Sales, Technicians	NA	NA
14101	Manufacture of all types of textile garments and clothing accessories	Home based/ Subcontracted	Marketing, Technician	NA	NA
14105	Custom tailoring	Home based/ Self Entrepreneurial tailoring	NA	NA	NA