NIPFP Working paper series

What do we know about changing economic activity of firms?

No. 249
22-January-2019
Radhika Pandey, Amey Sapre and Pramod Sinha

National Institute of Public Finance and Policy
New Delhi

What do we know about changing economic activity of firms?

Radhika Pandey
radhesp@gmail.com

Amey Sapre
amey07@gmail.com

Pramod Sinha
pramod.sinha@gmail.com

30th December, 2018

Abstract

Identification of primary economic activity of firms is a prerequisite for compiling several macro aggregates. In this paper we take a statistical approach to understand the extent of changes in primary economic activity of firms over time and across different industries. We use the history of economic activity of over 46000 firms spread over 25 years from CMIE Prowess to identify the number of times firms change the nature of their business. Using the count of changes, we estimate Poisson and Negative Binomial regression models to gain predictability over changing economic activity across industry groups. We show that a Poisson model accurately characterizes the distribution of count of changes across industries and that firms with a long history are more likely to have changed their primary economic activity over the years. Findings show that classification can be a crucial problem in a large dataset like the MCA21 and can even lead to distortions in value addition estimates at the industry level.

JEL: E00, E01

Keywords: Economic Activity, Manufacturing, India, Poisson Distribution

Radhika Pandey, Amey Sapre and Pramod Sinha are at the National Institute of Public Finance and Policy (NIPFP), New Delhi. This paper is an extended version of an earlier article 'The problem of identifying economic activity of firms', *Ideas for India*. The views expressed in the paper are those of the authors. No responsibility for them should be attributed to NIPFP. We are thankful to Dr. Pronab Sen for comments and valuable suggestions.

1 Introduction

In this paper we take a statistical approach to understand changes in the economic activity undertaken by firms. The purpose is two fold; first to quantify the extent of changes at an industry level, and second to gain predictability over changes across different industries and time period. Identification and classification of economic activity are are at the core of any statistical information system and are also important pre-requisites for compiling industry and macro aggregates. In the Indian case, some routinely used classification systems are the National Industrial Classification (NIC) for classifying economic activity of firms, National Product Classification for Service Sector (NPCSS), Indian Trade Clarification Harmonized System ITC-HS codes for trade in products, etc. that are often integrated with entity level identifiers such as Corporate Identity Number (CIN) for identification purposes.

The problem of identification gains importance when existing classification systems may not entirely serve the purpose for various statistical exercises. For instance, identification of the primary economic activity of a firm is of crucial importance from a national accounting perspective since macro aggregates such as Gross Value Addition (GVA), savings, capital stock etc., are required to be computed at a disaggregated industry level. While it may seem straight forward to compute an aggregate at the industry level, the issues with any given classification system are of accuracy and reliability. Problems emerge because firms can have multiple economic activities and even diversify across industries as part of their usual business strategy. In such cases, existing classification systems fall short of capturing actual changes in firms' business activities and may lead to a misclassification problem.

In the Indian context, classification of economic activity at the firm level primarily follows the NIC system. Firms are assigned different NIC codes at the time of their incorporation which eventually serves as a basis for classifying firms into industries for any estimation exercise. However, it is well known that such codes once assigned do not change even when the firm changes its economic activity or diversifies into other activities. This process poses several challenges for the statistical system as the task of relying on NIC codes for identification may lead to misclassification of firms for computing any industry level aggregate.

If firms are misclassified, the outcome of any estimation exercise can have several inconsistencies, such as (i) incorrect inclusion or exclusion of firms in an industry, (ii) distortion in industry level value addition or GDP estimates and (iii) incorrect diagnosis and policy formulation for industries. The consequences are; (i) mis-classified firms introduce volatility in levels and growth rates that is spurious as it does not reflect actual movements in value addition, (ii) it leads to a mis-representation of the GVA-to-Output (GVA/GVO) ratio for various industries.

In recent years, particularly after the introduction of the 2011-12 series of the National Accounts, the problem of identification of economy activity has taken a new dimension. In the 2011-12 series the Gross Value Addition (GVA) estimates of Private Corporate Sector i.e. the organized part of manufacturing and services are computed using the MCA21 dataset which was introduced in 2008. The current identification of primary economic activity of firms rests on; (See CSO 2015b for additional information) (i) identifying top three revenue generating products (ii) information from reported ITC-HS product codes wherever available, (iii) product details available in the MGT-7/9 forms submitted along with annual financial statements, (iv) checking websites of firms in absence of product information and finally, (v) using NIC codes contained in the CIN of the firms to assign them in a particular sector.

The issues with such a strategy are well known and have been a part of the debate on the manufacturing sector (See CSO (2015a), CSO (2015b), Nagaraj (2015a), Nagaraj (2015b), Sapre & Sinha (2016), Manna (2017) for a brief survey of issues). What is unknown is the extent of possible misclassification and the impact it may have on industry level GVA estimates.

We take the debate further by first highlighting the extent of the classification problem, which may be one of the sources of inaccuracies in GVA estimation. Second, in absence of any information on how frequently firms change their economic activity, gaining some predictability of changing economic activity can be useful in formulating an identification strategy. We build a simple empirical model to estimate the count of changes across different size classes and industries to ascertain whether changes in economic activities can be predicted with some accuracy.

2 Identification of Economic Activity: Problems and Concerns

In principle, there are two main reasons why classification of economic activity is a concern. First, based on assigned NIC codes, the initial and current activity of any firm may be different. Second, firms may change their nature of business at any point in time and also repeatedly. Taken together, the impact of changes in economic activity on any estimation process cannot be assumed to be negligible. Firms of different sizes move in and out of any industry, which makes it complicated to have an error-free list of firms that are currently and have always been in one economic activity. In such a situation, data users may naively believe that industry level aggregates are based on firms that have their primary economic activity in one particular industry and ignore the possibility that current and past business of the firms could be different.

Classifying firms requires studying (i) industry in which the firm was registered at the time of incorporation, (ii) the history of economic activities undertaken by the firm, (iii) diversification into other industries and (iv) top revenue generating products of the firm.

Within the available avenues to classify a firm in any industry, it is widely accepted to assign the *primary* economic activity based on the criterion of maximum revenue contribution of *products*. In order words, if a firm has revenue coming from multiple products, the firm is assigned the

economic activity (such as manufacturing or services) based on the product that contributes the maximum revenue in a given industry. For the organized manufacturing and services sector firms in the MCA21 database, this process is based on the identifiers as mentioned earlier in CSO (2015b) along with their financial statements.

Part of problem with this strategy is that the reported ITC-HS are product level codes that do not distinguish between manufacturing and trading as an economic activity. The criterion of maximum revenue contribution despite being the most practical also has its own limitations. For example, if changes in revenue sources are frequent, the entire enterprise is likely to either get included or excluded on a yearly basis in any industry. These changes can lead to abrupt shifts in levels and growth rates of value addition. Given these problems, the relevant question in this context is: what is the extent of the problem in the Indian scenario?

As earlier, the MCA21 dataset which is used for estimation of value addition in the organized part of manufacturing and service sector has more than 11 lakh firms (See MCA(2018) for details). The extent of changes in economic activity of large and small firms available in the dataset is presently unknown. In absence of such information, several classification errors could get introduced in the GVA estimation process. The task at hand is to first understand the extent of changes in firms' primary economic activity and attempt to gain some predictability of such changes. We build on this premise to show the extent of changes that happen at a firm level, which can later be used to on a large dataset to test for accuracy.

2.1 The Extent of the Problem

Presently, the MCA21 dataset is unavailable in public domain. To draw a parallel, we use a comparable dataset from CMIE Prowess to explore data of nearly 46000 listed/unlisted Private and Public Ltd. firms spanning over 25 years (1988-2018). The dataset covers a substantial part firms that submit financial returns in the e-XBRL format in the MCA and also has a wide span in terms of coverage across sectors. In addition, a history of economic activity of each firm is generated for each year of data availability based on the product schedule details.

Identifying changes in primary economic activity is a two-step process. The first step is a scrutiny of product schedules of the firm and the corresponding revenue generated from each product. Second, within available products, primary activity is assigned to the one with the *maximum revenue contribution.* The question that follows is: how frequently does primary activity change? Since businesses can change their primary activity in any year, we begin by arranging firms as per years of data availability. Table 1 shows the number of times primary activity changed for firms with different years of data availability.

For firms with one year of data, the primary activity is their current activity and hence shows no change. Next, out of 5422 firms with two years of data, 296 firms showed a change as compared to their previous year's activity. Similarly, in case of firms with four years of data, 553 show a change once, while 38 firms changed primary activity twice.

Table 1: Changes in primary economic activity of firms over time

Data Avail. (Yrs)	Change in primary activity (number)					Total	At least 1 change	% (D/C)	Cum (Col. D)
A	B					C	D	E	F
	0	1	2	3	4				
1	1025	0	0	0	0	1025	0	0	0
2	5126	296	0	0	0	5422	296	5	296
3	4314	536	26	0	0	4876	562	12	858
4	2957	553	38	0	0	3548	591	17	1449
5	3883	823	66	2	0	4774	891	19	2340
6	2697	755	72	0	0	3524	827	23	3167
7	2014	625	95	7	0	2741	727	27	3894
8	1787	669	109	6	0	2571	784	30	4678
9	1851	762	120	5	0	2738	887	32	5565
10	1605	716	129	7	0	2457	852	35	6417
11	1150	499	91	6	0	1746	596	34	7013
12	1097	463	93	11	0	1664	567	34	7580
13	795	363	85	7	0	1250	455	36	8035
14	646	306	56	5	0	1013	367	36	8402
15	513	240	48	7	0	808	295	37	8697
> 15	3850	1946	627	84	1	6508	2658	41	11355
Total	35310	9552	1655	147	1	46665	11355	24	–

As we move on the number of years, changes become more frequent and the extent of the problem appears clearly. In Columns D and E we add the number of firms with changes in primary activity (i.e. excluding no change) for each year and compute its share to get a sense of the magnitude.

The numbers reveal some interesting aspects of firms' businesses; (i) over time, firms are frequent in changing their business activities and such changes are not apparent in an aggregate GVA estimate, (ii) sources of revenue keep changing, which from a data point of view alters the primary business of the firm, (iii) with a longer time span, the possibility of a large number of firms having repeatedly changed their primary activity increases.

If we consider the simplest case, Figure 1 plots the share of firms that have changed their primary activity at least once over the years. As earlier, the percentage of firms that show a change increases substantially once we have a longer time span. For example, firms that have eight years data, 30% of them have changed their activity once. To delve deeper at the industry level, we explore firms that have ten years of data and changed their primary activity once. Table 2 shows a cross tabulation of firms changing their primary economic activity from their initial year to their present one over a period of 10 years.

As an illustrative case, out of the 220 firms in manufacturing, 99 had moved to non-financial services, while 57 moved into manufacturing from non-financial services at least once in the past 10 years.

Figure 1: Percentage share of firms with at least one change in primary activity

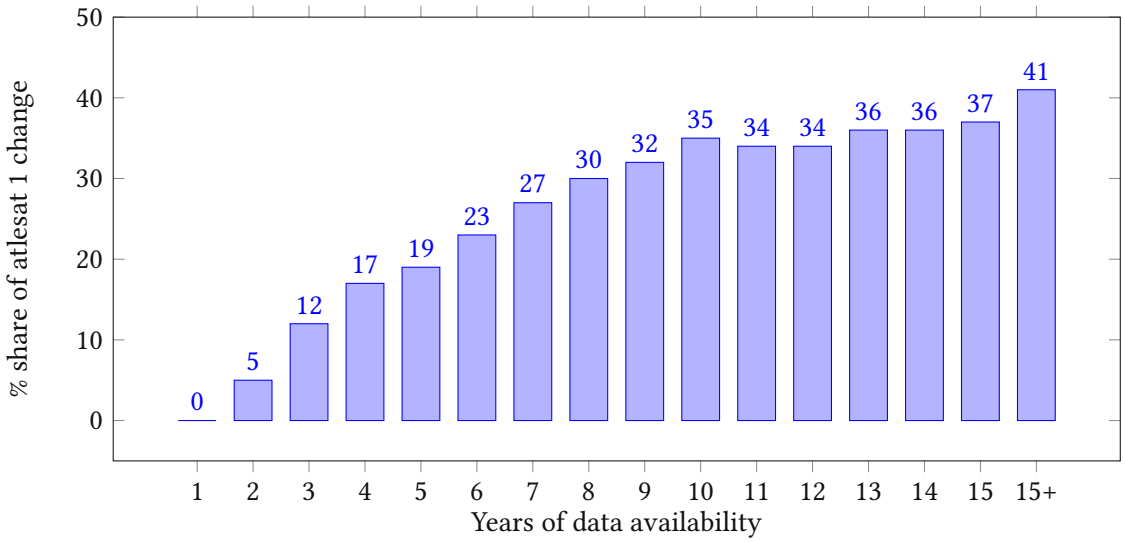

Table 2: Changes in economic activity across industries

	Industry changed to						
Previous Ind.	Constr.	Electr.	Fin. Serv.	Manuf.	Mining	NF.Serv.	Total
Construction	20	0	22	8	0	18	68
Electricity	0	4	9	4	0	1	18
Fin. Serv.	24	3	62	15	3	85	192
Manuf.	7	2	41	68	3	99	220
Mining	0	0	4	0	2	3	9
NF. Serv	8	2	82	57	3	57	209
Total	59	11	220	152	11	263	716

Fin. Serv. is Financial Services, NF.Serv is Non-Financial Services, Manuf. is Manufacturing

Over the years, it is possible for firms to move in and out an activity or return to their original activity after having ventured into multiple industries. If we were to include multiple changes, the task becomes unmanageable as it is difficult to track the economic activity of a firm from its historic data. The situation would be further accentuated in case of a large dataset like the MCA21, which currently may not have a detailed history of classification of firms. Since it is difficult to construct the history of each firm, as a preliminary step, some statistical methods can be used to gain predictability over changing economic activity.

In what follows, we attempt to predict the number of changes in primary economic activity using basic information such as years of data availability, industry classification and size at the firm level.

3 Gaining Predictability Over Changing Economic Activity

One possible approach to predict the number of events is to estimate a Poisson regression on a variable that follows a Poisson distribution. It is well known that the Poisson is a discrete probability distribution that expresses the probability of independent occurrences of a given number of events over a time interval. Rare events, accidents, etc. typically follow a Poisson process and can be modeled using suitable control variables in a regression framework. Similarly, the count of changes in economic activity of firms can be taken as events spread across industries and time.

To formulate the basic distribution, suppose that an event can occur 1, 2, … , m times in a given interval and the average number of events in that interval is denoted by a parameter (λ). Given the average value, the probability of observing m events in the interval is given by the expression;

$$f(m, \lambda) = \Pr(X = m) = \frac{\lambda^m e^{-\lambda}}{m!} \tag{1}$$

The Poisson distribution is characterized by a single parameter (λ), which can also be used to denote the rate or intensity of the occurrence of the event. In the present case, we can denote the events of change (C) in a particular time period as following a Poisson distribution, i.e.

$$C \sim \text{Pois}(\lambda); \;\; \lambda > 0 \tag{2}$$

In order to estimate the count of changes, we re-group the data as follows. First, given the heterogeneity of firms, we classify each firm into a size decile based on the distribution of the initial size of firms. We use the three year average of total income plus total assets as a measure of size. Second, we assign each firm into an industry group based on their initial primary economic activity [See note 1 for details]. Thus, as a starting point of the history, each firm f_{ij} is classified into an initial size (i) and industry (j) group.

The Poisson distribution can be extended to group data, which is relevant in our case. Let C_{ij} denote the total number of changes observed in the ith size decile and the jth industry. Then, extending the Poisson over count in each group, i.e. $C_{ij} \sim \text{Pois}(\lambda_i)$, we can denote the group total as $C_i \sim \text{Pois}(n_i\lambda_i)$ $j = 1, 2, \ldots n$. In other words, if individual counts C_{ij} are a Poisson process with mean λ_i, then the group total C_i is also a Poisson process with the expected count given by $n_i\lambda_i$. With $C \sim \text{Pois}(\lambda)$, the count of changes ($C$) can be expressed as in a Poisson regression as;

$$C = \beta_0 + \beta_1 x_{1,j} + \cdots + \beta_k x_{k,j} \tag{3}$$

where xs are explanatory variables. The response variable can also be expressed as a 'rate' over time or 'incidence' within a given quantity. Thus, (C) can be re-expressed as; C/t or C/n where (n) is a measure of quantity.

The log-linear version can be written as; $\ln(C/n) = \beta_0 + \beta_1 x_{1,j} + \cdots + \beta_k x_{k,j}$, which can be rearranged as;

$$\ln(C) = \beta_0 + \beta_1 x_{1,j} + \cdots + \beta_k x_{k,j} + \ln(n) \tag{4}$$

In the literature, the term $\ln(n)$ is called an *offset* effect (See Casella & Berger (2002), Cameron and Trivedi (2010) for details and related cases). The offset variable is introduced to account for over-dispersion that makes the mean and variance of the Poisson variable unequal. The variable $\ln(n)$ is also similar to the notion of 'exposure' and can be explicitly added as a control variable in the regression. The inclusion of the offset variable makes the count equal to;

$$C = e^{\beta_0 + \beta_1 x_{1,j} + \cdots + \beta_k x_{k,j} + \ln(n)} = (n).e^{\beta_0}.e^{\sum_k \beta_k x_{k,j}} \tag{5}$$

thereby making the mean proportional to (n).

The equation is estimated using the log-link specification with two main variables of interest; years of data availability and dummies for size decile. Findings from 1 show that over time, firms are frequent in changing their economic activity. Thus, number of years of data availability becomes an important predictor which is expected to show a positive relation to number of changes. However, as a caveat, years of data availability is not synonymous to the age of the firm. Nevertheless, it is sufficient to capture a long history of changes in economic activities of the firm and characterize the Poisson distribution.

The formulation rests on two more aspects. The spread of the events is always in context to a certain population or exposure. In this particular case, the number of firms in each size decile and industry serve as a measure of exposure. Second, alternate specifications need to be considered in cases of over-dispersion, i.e. where mean and variance of the Poisson variable are unequal. We use two variants of Poisson and Negative Binomial models as alternate specifications [see note 2 for details]. Using (5), the estimated equation is of the type;

$$C_{ij} = \ln(n_{ij}) + \sum_k \beta_k x_{ij} + \upsilon_{ij} \tag{6}$$

where (C) is the count of changes, (n) is the number of firms and (x)'s include dummies for size decile. Table 3 shows the parameter estimates of four models followed by the predicted counts and descriptive statistics for each group.

Table 3: Parameter estimates of Poisson and Negative Binomial Model

Dep. Var (count)	(1)	(2)	(3)	(4)
Variables	Pois	NB	PoisGLM	NBGLM
ln(Avg. years)	0.347***	0.419***	0.347***	0.419***
	(0.0814)	(0.132)	(0.127)	(0.129)
ln(nfirms)	0.886***	0.915***	0.886***	0.915***
	(0.0247)	(0.0229)	(0.0285)	(0.0200)
Size Dec. (2)	0.0877	0.0940	0.0877	0.0940
	(0.114)	(0.113)	(0.0819)	(0.102)
Size Dec. (3)	0.0225	0.0455	0.0225	0.0455
	(0.114)	(0.117)	(0.0851)	(0.105)
Size Dec. (4)	0.00723	0.0367	0.00723	0.0367
	(0.113)	(0.113)	(0.0879)	(0.107)
Size Dec. (5)	−0.0621	−0.0836	−0.0621	−0.0836
	(0.112)	(0.134)	(0.0928)	(0.110)
Size Dec. (6)	−0.106	−0.0780	−0.106	−0.0780
	(0.112)	(0.126)	(0.0948)	(0.111)
Size Dec. (7)	−0.273**	−0.185	−0.273***	−0.185*
	(0.119)	(0.144)	(0.0982)	(0.110)
Size Dec. (8)	−0.282**	−0.148	−0.282***	−0.148
	(0.128)	(0.144)	(0.0975)	(0.110)
Size Dec. (9)	−0.363***	−0.300**	−0.363***	−0.300***
	(0.118)	(0.136)	(0.0955)	(0.108)
Size Dec. (10)	−0.581***	−0.480***	−0.581***	−0.480***
	(0.144)	(0.156)	(0.100)	(0.109)
Constant	−1.043***	−1.430***	−1.043***	−1.430***
	(0.218)	(0.345)	(0.236)	(0.257)
N	62	62	62	62
exp(Constant)	0.352	0.239	0.352	0.239
ln(α)		0.020***		
LR χ^2	1754	2317	6525	2316
Prob< χ^2	0.00	0.00	0.00	0.00
Pseudo R^2	0.958			
Deviance GoF	241.00		241.00	73.47

Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Model 1 is the basic Poisson regression with a log-link specification. As hypothesized, the (avg.) number of years of data availability shows a positive and significant impact on count of changes. Similarly, the offset variable (number of firms) also suggests that the count of changes increase with an increase in number of firms. To quantify the impact, first, the coefficient on ln(Avg. years) can be interpreted as an Incidence Rate Ratio (IRR), i.e. [exp(0.347) ≈ 1.414] which shows that on average, a year increase in available data increases the probability of change by around 41% [exp(0.347) − 1 × 100 ≈ 41.4] . These estimates are fairly consistent with the result in Table 1 and can further be improved with changes in specifications.

The coefficients on size decile dummies suggests that larger sized firms are more likely to change their activity as the net effect of their coefficients and the constant value is positive, i.e. $[e^c + e^{sd_i}]$ where c is the value of the constant and sd_i are the coefficients on respective size decile dummies.

In Model 2, we formulate a Negative Binomial model to account for over-dispersion in count. The result is similar to the Poisson case but improves the magnitudes of the parameters. The coefficient on ln(Avg. years) shows a much larger impact [$\exp(0.419) - 1 \times 100 \approx 52.0$] which confirms the earlier premise that number of changes increase with a longer history of the firm. The Negative Binomial model also suggests that count shows over dispersion as the parameter (α) = $\exp(0.020) \approx 1.021$ is positive and statistically significant. The counts predicted by these models are tabulated in Tables 4 and 5.

Before we conclude on the results, we can change model specifications to verify if parameter estimates are sensitive to changes over-dispersion. In the Poisson model over-dispersion (α) was set to zero, while in the Negative Binomial case it is equal to the mean. In Models 3 and 4 we allow for the variation to be proportional to the size of the exposure instead of the mean. We use the Generalized Linear Model (GLM) that has a link function to allow the variance of each group to be a function of its predicted value [See Note 3]. The model provides us two additional outcomes, i.e. a scale parameter to understand the difference between the mean and variance and the extent to which standard errors are inflated. The estimated scale parameter for Model 3 is 4.69 suggesting that the variance in count is nearly 4 times the mean and thus the standard errors are inflated nearly 2 times ($\sqrt{4.69} \approx 2.1$). Correcting for the standard error also gives us reliable parameter estimates and thus these models can be used to predict the count and a comparison can be drawn to choose the best fitting model.

Tables 4 and 5 show the predicted values along with the information on variables used in the model. The figures include number of firms, average years of data and the average of initial size. Predictions of Model (3) are the closest to the actual count values (nchanges) and is therefore most suitable for drawing inferences from the results. Across size deciles, the predicted values are much closer for higher deciles and particularly for lower count values. The results also revel interesting facets about firm level data. Across industries, in comparison to sectors such as construction or mining, we expect a larger number of firms in manufacturing or financial services. Thus, in terms of exposure, these sectors are more likely to experience changes in primary economic activity. To delve deeper, in Table 6 we tabulate the descriptive statistics by different size decile.

Descriptive statistics of predicted counts show that the GLM Poisson model predicts closest to actual count in most of the deciles. At the aggregate, the model closely follows the actual distribution and predicts within the range [max–min] of the actual count variable.

Table 4: Predicted count of change for each size decile & industry
using Poisson & Negative Binomial Regression

Size. Dec	First industry	N.firms	Avg. Yrs.	Avg. F.size	E(Change)	nchanges	Pois	NB	PoisGLM	NBGLM
1	Construction and Real estate	569	5.62	0.35	0.35	202	177	164	177	164
1	Electricity	201	6.07	0.34	0.18	37	72	65	72	65
1	Financial services	990	7.15	0.44	0.28	281	314	300	314	300
1	Manufacturing	1668	6.21	0.36	0.36	603	475	457	475	457
1	Mining	26	7.19	0.34	0.38	10	13	11	13	11
1	Non-financial services	1375	6.00	0.42	0.23	312	395	377	395	377
2	Construction and Real estate	311	5.89	3.03	0.32	100	115	105	115	105
2	Electricity	108	6.06	2.75	0.30	32	45	41	45	41
2	Financial services	1512	8.93	2.96	0.38	573	539	534	539	534
2	Manufacturing	1096	7.48	3.08	0.36	395	381	369	381	369
2	Mining	17	7.82	3.07	0.53	9	10	8	10	8
2	Non-financial services	1524	6.65	3.02	0.31	470	490	475	490	475
3	Construction and Real estate	331	5.97	8.88	0.31	102	114	107	114	107
3	Electricity	57	6.26	8.72	0.25	14	24	22	24	22
3	Financial services	1417	8.98	9.04	0.34	487	477	480	477	480
3	Manufacturing	1269	8.49	9.10	0.36	457	425	424	425	424
3	Mining	31	10.45	9.20	0.71	22	17	15	17	15
3	Non-financial services	1503	7.27	9.01	0.29	443	467	464	467	464
4	Construction and Real estate	365	6.33	19.50	0.30	111	125	119	125	119
4	Electricity	55	6.53	19.81	0.36	20	24	21	24	21
4	Financial services	1327	9.76	19.24	0.35	466	457	464	457	464
4	Manufacturing	1442	9.45	19.57	0.36	516	486	494	486	494
4	Mining	33	9.39	20.75	0.61	20	17	16	17	16
4	Non-financial services	1440	7.28	19.51	0.29	419	443	443	443	443
5	Construction and Real estate	341	6.97	36.57	0.32	110	114	103	114	103
5	Electricity	50	6.46	35.72	0.08	4	20	17	20	17
5	Financial services	1017	9.46	35.45	0.33	340	333	319	333	319
5	Manufacturing	1800	10.65	37.03	0.33	599	575	564	575	564
5	Mining	27	11.85	35.32	0.52	14	14	13	14	13
5	Non-financial services	1438	7.92	36.25	0.29	415	425	406	425	406

N.Firms is number of firms, Avg. years is the average of years of data availability, Avg. F.size is the average value of initial size*
E(Change) is the probability of change, i.e. [Nchanges/Nfirms], Nchanges is the total number of changes in primary economic activity

Table 5: Predicted count of change for each size decile & industry using Poisson & Negative Binomial Reg

Size. Dec	First industry	N.firms	Avg. Yrs.	Avg. F.size	E(Change)	nchanges	Pois	NB	PoisGLM	NBGLM
6	Construction and Real estate	321	6.40	65.57	0.36	115	100	95	100	95
6	Electricity	45	7.09	66.48	0.20	9	18	16	18	16
6	Financial services	854	8.95	63.90	0.31	269	268	267	268	267
6	Manufacturing	2082	11.36	65.82	0.31	644	641	666	641	666
6	Mining	24	12.92	63.04	0.54	13	13	12	13	12
6	Non-financial services	1339	7.81	65.37	0.28	370	380	380	380	380
7	Construction and Real estate	307	6.97	117.97	0.39	121	84	85	84	85
7	Electricity	67	8.00	114.38	0.31	21	23	22	23	22
7	Financial services	628	8.34	114.56	0.28	173	169	176	169	176
7	Manufacturing	2330	11.49	116.95	0.25	576	601	667	601	667
7	Mining	40	8.90	119.65	0.30	12	15	15	15	15
7	Non-financial services	1296	7.62	116.10	0.23	299	310	328	310	328
8	Construction and Real estate	342	7.01	224.46	0.31	107	92	97	92	97
8	Electricity	74	6.24	209.65	0.26	19	23	23	23	23
8	Financial services	511	8.37	215.10	0.38	193	139	151	139	151
8	Manufacturing	2392	11.19	216.26	0.23	556	604	701	604	701
8	Mining	35	11.14	221.59	0.63	22	14	15	14	15
8	Non-financial services	1305	7.51	216.57	0.22	283	308	341	308	341
9	Construction and Real estate	398	6.12	482.36	0.27	108	92	91	92	91
9	Electricity	78	6.36	465.53	0.12	9	22	21	22	21
9	Financial services	451	7.64	476.62	0.30	135	111	111	111	111
9	Irrigation	2	8.50	541.95	0.50	1	1	1	1	1
9	Manufacturing	2307	10.53	473.87	0.21	495	528	568	528	568
9	Mining	17	13.76	471.49	0.47	8	7	7	7	7
9	Non-financial services	1414	7.07	482.59	0.22	305	298	307	298	307
10	Construction and Real estate	483	6.35	3286.66	0.27	129	89	92	89	92
10	Electricity	171	9.05	16545.27	0.19	33	40	41	40	41
10	Financial services	636	9.82	7093.53	0.26	167	133	142	133	142
10	Irrigation	6	12.83	13953.82	0.50	3	2	2	2	2
10	Manufacturing	1944	9.30	2971.76	0.15	289	350	385	350	385
10	Mining	47	14.02	9487.74	0.28	13	15	15	15	15
10	Non-financial services	1379	6.83	4300.76	0.16	227	232	247	232	247
	Total	46665	–	–	–	13307	13307	13479	13307	13479

N.Firms is number of firms, Avg. years is the average of years of data availability, Avg. F.size is the average value of initial size*
E(Change) is the probability of change, i.e. [Nchanges/Nfirms], Nchanges is the total number of changes in primary economic activity

Table 6: Descriptive statistics of predicted counts by size decile

Size Dec.	Stat.	Actual	Predicted			
		Count	Pois	NB	PoisGLM	NBGLM
1	Mean	240.83	240.83	228.94	240.83	228.94
1	SD	216.42	183.51	177.50	183.51	177.50
1	Min	10.00	12.53	10.78	12.53	10.78
1	Max	603.00	474.50	456.56	474.50	456.56
2	Mean	263.17	263.17	255.35	263.17	255.35
2	SD	245.28	234.36	231.65	234.36	231.65
2	Min	9.00	9.67	8.32	9.67	8.32
2	Max	573.00	538.66	533.74	538.66	533.74
3	Mean	254.17	254.17	252.07	254.17	252.07
3	SD	230.54	224.86	226.49	224.86	226.49
3	Min	14.00	17.05	15.50	17.05	15.50
3	Max	487.00	477.26	480.13	477.26	480.13
4	Mean	258.67	258.67	259.43	258.67	259.43
4	SD	232.66	226.42	230.86	226.42	230.86
4	Min	20.00	17.10	15.56	17.10	15.56
4	Max	516.00	485.97	494.14	485.97	494.14
5	Mean	247.00	247.00	236.93	247.00	236.93
5	SD	242.03	232.45	227.51	232.45	227.51
5	Min	4.00	14.48	12.65	14.48	12.65
5	Max	599.00	575.25	564.32	575.25	564.32
6	Mean	236.67	236.67	239.28	236.67	239.28
6	SD	245.52	245.41	254.96	245.41	254.96
6	Min	9.00	12.87	11.84	12.87	11.84
6	Max	644.00	640.53	666.07	640.53	666.07
7	Mean	200.33	200.33	215.28	200.33	215.28
7	SD	212.38	225.06	250.02	225.06	250.02
7	Min	12.00	15.04	14.53	15.04	14.53
7	Max	576.00	601.38	666.52	601.38	666.52
8	Mean	196.67	196.67	221.21	196.67	221.21
8	SD	203.32	226.42	263.30	226.42	263.30
8	Min	19.00	14.32	14.66	14.32	14.66
8	Max	556.00	604.34	700.86	604.34	700.86
9	Mean	151.57	151.57	157.85	151.57	157.85
9	SD	185.86	195.32	209.32	195.32	209.32
9	Min	1.00	0.95	0.82	0.95	0.82
9	Max	495.00	528.47	567.51	528.47	567.51
10	Mean	123.00	123.00	132.00	123.00	132.00
10	SD	111.76	127.55	140.05	127.55	140.05
10	Min	3.00	2.34	2.22	2.34	2.22
10	Max	289.00	349.82	384.96	349.82	384.96
Total	Mean	214.63	214.63	217.42	214.63	217.42
Total	SD	203.19	202.35	209.64	202.35	209.64
Total	Min	1.00	0.95	0.82	0.95	0.82
Total	Max	644.00	640.53	700.86	640.53	700.86

Within the descriptive statistics the primary interest is in comparing the mean of the Poisson variable. Since the essence of the model is in this single parameter, accurately modeling the mean is of a greater priority. Empirically, over dispersion can be modeled using various specifications, but it cannot substitute the basic component of the Poisson model. Thus, if we compare the results jointly on mean and variance after allowing for over dispersion in the variance, the GLM Poisson model gives us the most desired result.

Based on the results of the empirical model, what have we understood about changes in primary economic activity of firms? The results show that statistically, one can gain predictability over changes in economic activity with a high degree of accuracy. A Poisson model with modifications to allow for over dispersion sufficiently characterizes the actual distribution of changes in economic activity across industry and size deciles. The results suggests that firms across industries are frequent in changing the nature of their business which is likely to pose a problem if re-classification is not done on a routine basis. Also, the number of changes increase as we consider a longer time span, i.e. firms that have been in existence for a several years are more likely to show changes in their nature of business.

Qualitatively, the results have implications for firm and industry level analysis. First, changes in primary economic activity are driven by the maximum revenue contribution criteria. If frequent changes in sources of major revenue are not captured on a year-to-year basis, it can lead to classification problems and hence present an inaccurate picture at the industry level. The analysis also shows that firms with longer history and larger size are more likely to face a classification problem. While this exercise is limited to a statistical analysis of changes in economic activity, the reasons behind why and when firms change their business need to be explored both qualitatively and quantitatively.

4 Conclusion

Identification of economic activity of firms is a prerequisite for building several aggregates at the industry level. In the Indian context, systems such as the National Industrial Classification (NIC), ITC-HS codes for trade in products, among other are used for assigning and classifying firms into various economic activities. While such systems serve solely as a means for identification, the use of classification of economic activity based on such systems can have unintended consequences for building industry level aggregates. In practice, firms are known to diversify and move across industries as part of their usual business strategy. In addition, multiple and frequent changes in economic activity over time makes it difficult for any statistical information system to have an error-free list of firms are currently and have always been in one economic activity.

The impact of changes in economic activity on any estimation process cannot be assumed to be negligible. There are two main consequences that cannot be ignored; (i) misclassification introduces spurious volatility in levels and growth rates as it does not represent actual movements

in value addition in a sector, and (ii) it distorts the GVA-to-output (GVA/GVO) ratio which is significantly different for manufacturing and services.

To identify changes in economic activity, we use data from CMIE for over 46000 firms spread over a period of 20 years. Based on the history of the economic activity of firms, we count the number of times firms change the nature of their business and thus move in and out of different industries. We show that firms with longer history are more likely to have changed their business more than once and such changes are not apparent while dealing with industry level aggregates. Considering a period of over 25 years, findings also show that the share of firms in total that have changed their activity at least once is close to 29%. The history of classification also reveals interesting facets about firms' businesses. Firms are frequent in changing in their nature of business and thus a classification system has to be dynamic to capture such frequent changes. Sources of revenue keeps changing for a firm, which from a data point of view alters the classification of a firm, say from manufacturing to service or vice-versa.

Based on the frequency of changes in economic activity, we use Poisson and Negative Binomial regression models to gain predictability over number of changes. We show that a Poisson model after controlling for over dispersion in variance is nearly accurate in predicting the count of changes across different size deciles of firms. Results also show that changes higher size deciles are more likely than others which can eventually lead to a greater distortion in industry level estimates. Identification of economic activity remains amongst the finer aspects of measurement and accuracy of macro aggregates. The case of the manufacturing or service sector is no different. Identifying primary activity of firms continues to be a challenge and may well require new systemic thinking in addition to current efforts and practices. Historic data of firms can reveal changes in economic activity, but the task of deciphering information from a large dataset like the MCA21 is challenging.

Notes

1. Data and Variable description:

- Data were sourced from CMIE Prowessdx database vintage as of March 2018. The information used covered the History of Classifications which provides company-wise time-series details on the industry in which the company operates. In turn the information on the product/industry was sourced from various company documents such as Annual report, websites, Red herring prospectus etc.
- Years of data availability: Number of years for which data is available for a firm.
- Size: Three year average of the total income and assets of a firm and should be greater than zero.
- Industry: A firm is classified as belonging to a particular industry, if the industry accounts for more than half the sales of the company. The industry classification of companies is facilitated by disclosure requirements in Section 3(i), (ii) and 4(D) of Part II of Section VI of the Companies Act, 1956. The Act mandates that all companies disclose, in their Annual Reports, quantitative details of all products purchased, produced or traded by them. [See CMIE Knowledge Base for details]

2. In cases of over dispersion, we can add a random term to the Poisson mean, say θ to capture the heterogeneity across the entities. If we assume that the conditional distribution of C, given a value of θ follows a Poisson distribution, with mean and variance equal to $\theta\lambda$, we can then express the relation as;

$$C \mid \theta = \text{Pois}(\lambda_\theta)$$

The θ parameter is assumed to capture the unobserved factors that affect the mean and variance of the response variable. Empirically, we could estimate the Poisson model, provided we obtain a measure of θ. Instead, in absence of any knowledge of θ, if we assume that θ follows a Gamma distribution, with its two parameters α and β, we can re-characterize our original response variable as follows. First, the mean and variance of the Gamma distribution are α/β and α/β^2. Let $\alpha = \beta = 1/\sigma^2$. Then, the mean of the distribution becomes unity, and variance σ^2. Incorporating this change in the Poisson variable, we get the well known result of a Negative Binomial distribution. (See Casella & Berger (2002), Cameron and Trivedi (2010) for details)

3. The method provides a measure of over dispersion by estimating a scale parameter $\hat{\phi} = (\frac{\chi^2_P}{n-k})$ where χ^2_P is the Pearson goodness of fit statistics, and (n, k) are the number of observations and parameters. The corrected standard errors are obtained as $\sqrt{\hat{\phi}}$. The GLM in the Negative Binomial model uses the estimated shape parameter ($\alpha = 0.020$) in the variance function to account for over dispersion.

References

Cameron, Colin A. and Trivedi, Pravin K. (2010) *Microeconometrics using Stata*, Stata Press, StataCorp, Texas: US

Casella, George and Berger, Roger (2002) *Statistical Inference*, Second Ed: Cengage Learning: US

CSO (2015a) "No room for doubts on new GDP numbers", *Economic and Political Weekly*, Vol. L, No. 16, April 18th, pp. 86-89

CSO (2015b) *Final Report of the Sub-Committee on Private Corporate Sector including PPPs*, National Accounts Division, Ministry of Statistics and Programme Implementation, Government of India, New Delhi

Manna, GC (2017) "An investigation into some contentious issues of GDP estimation", *Journal of Indian School of Political Economy*, Vol. XXIX, No. 1 & 2, January-June, 2017

MCA (2018) *Monthly Information Bulletin on Corporate Sector*, Ministry of Corporate Affairs, Govt. of India, January, 2018

McCullagh, P and Nelder, J (1983) *Generalized Linear Models*, Springer Science, Springer: Verlag

Nagaraj, R (2015a) "Seeds of doubt on new GDP numbers Private corporate sector overestimated?", *Economic and Political Weekly*, Vol. L, No. 13, March 28th, pp. 14-17

Nagaraj, R (2015b) "Seeds of doubt remain: A reply to CSO's rejoinder", *Economic and Political Weekly*, Vol-L, No. 18, May 8th , pp. 64-66

Sapre, Amey and Sinha, Pramod (2016) "Some areas of concern about Indian manufacturing sector GDP estimation", National Institute of Public Finance and Policy (NIPFP) *Working Paper*, No. 172/2016

* * * * * * *

MORE IN THE SERIES

- Patnaik Ila., Mittal, S., and Pandey, R. (2019). Examining the trade-off between price and financial stability in India WP No. 248 (January).

- Datta, P. (2018). Value destruction and wealth transfer under the Insolvency and Bankruptcy Code, 2016 WP No. 247 (December).

- Bailey, R., Parsheera, S., Rahman, F., and Sane, R. (2018). Disclosures in privacy policies: Does "notice and consent" work? WP No. 246 (December).

Radhika Pandey, is Consultant, NIPFP
Email: radhika.pandey@nipfp.org.in

Amey Sapre, is Assistant Professor, NIPFP
Email: amey.sapre@nipfp.org.in

Pramod Sinha, is Consultant, NIPFP
Email: pramod.sinha@nipfp.org.in

National Institute of Public Finance and Policy,
18/2, Satsang Vihar Marg,
Special Institutional Area (Near JNU),
New Delhi 110067
Tel. No. 26569303, 26569780, 26569784
Fax: 91-11-26852548
www.nipfp.org.in