Annals of "Dunarea de Jos" University of Galati
Fascicle I. Economics and Applied Informatics
Years XXV – nº3/2019
ISSN-L 1584-0409 ISSN-Online 2344-441X
www.eia.feaa.ugal.ro

DOI https://doi.org/10.35219/eai1584040957

An Exercise in Cross-Validation of Data on Employment and Unemployment

Inamdar NIRAD*

ARTICLE INFO

Article history:
Accepted September 2019
Available online December 2019
JEL Classification
J82, J21, J16, O15, O12

Keywords:
Cross-validation of data, Women's employment, Unemployment, NSSO, IHDS-II

ABSTRACT

This paper compares two well-known resources of socioeconomic data of India – one from National Sample Survey Office (NSSO) 68th round and one from India Human Development Survey 2nd round (IHDS-II). Academic research hinges on validation of underlying data, especially in the socioeconomic field, since it can have Government policy implications. This study compares the two data sources from the same time frame (2011 – 2012) on relevant parameters pertaining to employment and unemployment. In particular, for the greater part of the study, we focus on women. First, we identify some factors influencing the labour force participation rate (LFPR) and stock of persons' unemployment rate (SPUR) of women – sector, marital status, education, relation to household head, religion, social group and household size. We perform a series of Probit analyses these separately to determine which factors are significant for LFPR and SPUR. Then, we combine and include their interaction effects. At every step, we compare the results from NSSO and IHDS-II to cross-validate them. We find that for LFPR, both data sources give similar results that all variables are significant, but for SPUR, they give different results. Finally, we perform Kolmogorov-Smirnov (K-S) tests and observe that both samples are normally distributed. However, given the disparity of results in a majority of analyses, one cannot say that the two data sources are similar. Therefore, in the absence of cross-validation of data from NSSO & IHDS-II, the scope for further research is to include a third data source and derive results from similar analyses. Finally, a comprehensive study would entail proposing a consolidated model, based on three sources.

1. Introduction

Since 1950, the National Sample Survey Office (NSSO) has provided demographic and socio-economic data on India. This has been the basis for academic research in economics, sociology and political science, which in turn, has influenced administrative policy formulation. Considering the widespread implications of such policies, it is vital that the underlying research data are accurate. Hence, academicians have paid a lot of attention towards designing frameworks to test for validity of data. In particular, a number of papers provide methodologies to compare and cross-validate two data sources. While all these papers present detailed findings, we observe that in the available literature, no researcher has compared NSSO & IHDS with respect to employment and unemployment. We recognise that this is a vital socioeconomic issue, affecting nearly 1 billion citizens of India. Therefore, this paper compares the two sources in a similar time period (2011 – 2012), to bring out the similarities & differences on the data. In particular, we will examine whether the prime factors affecting employment & unemployment are the same.

2. Literature review

Minhas (1988) compares the NSSO estimates of household expenditure & the independent data on private consumption from the National Accounting System (NAS). It attempts to – (a) investigate the scientific rationale behind these comparisons, (b) indicate the weaknesses and strengths of the external validator data set as a touchstone to assess the reliability of the NSSO estimates, (c) assess the degree of cross-validation, or the lack of it, between the two sets relating to two years, and (d) draw some lessons for improving the two data sets as well as caution against their incorrect uses.

On similar lines, Bhattacharya (2003) provides a methodology to compare two different datasets from the banking industry in India. In particular, it examines the validity of the BankScope database by comparing results based on it to those obtained from the population-level data for India disseminated by the Reserve Bank of India (RBI). The paper finds that despite good coverage and minor reporting errors, there is strong evidence of selectivity bias in the BankScope data. Choudhury (2002) attempts to identify the extent and nature of potential selectivity biases in samples drawn Prowess, which is a firm-level database on Indian

*Indian Institute of Management, Lucknow, India. E-mail address: efpm@iiml.ac.in (I. Nirad – Corresponding author)

Industry. Centre for Monitoring Indian Economy (CMIE), a private organisation, compiles this database. The research shows that due to non-random selection of samples, such data could be prone to selectivity biases.

Hirway (2002) analyses the validity of the NSSO data. It says that, the concepts and methods of NSSO are not able to capture work of the poor, particularly of women, satisfactorily. The workforce in certain 'difficult to measure sectors', such as subsistence work, home-based work or informal work, can be better captured through time use surveys (TUS). Using data from the pilot time use survey (1998-99), this paper shows that (a) this survey technique is of getting more realistic estimates of workforce and (b) some of the work not captured in NSSO surveys but captured in the time use surveys is likely to explain the changes in the employment situation in the nineties to a considerable extent. So, in the last two decades, there has been an ongoing debate about the validity of NSSO as a credible source. Scholars have pointed its shortcomings out.

Although NSSO is the most established data collection organisation in India, doubts persist about its level of accuracy. In a bid to overcome some of these questions, other institutions also conduct their own data collection exercises. One such is India Human Development Survey (IHDS), a collective effort of researchers from the University of Maryland and the National Council of Applied Economic Research (NCAER), New Delhi. In a similar effort as the papers mentioned above, Oldiges (2012) specifically compares NSSO & IHDS on the parameter of per capita cereal consumption (PCCC). It divides the population into numerous brackets, representing different ranges of percentiles as per mean per capita expenditure (MPCE). It notes the difference in the methodologies that whereas NSSO defines MPCE for a monthly period, IHDS-II defines for a mixed recall period. Chancel & Piketty (2017) compares these two datasets with respect to income equality.

3. Description of data

The data are in two groups. One group is from the official Government of India source, National Sample Survey Office (NSSO) 68th round (2011 – 2012). Particularly, we refer to Schedule 10 on 'Employment and Unemployment' and in particular, we focus on two files, which are called Block 4 (Demographic particulars of household members) and Block 5.1 (Usual Principal Status). The second group is from India Human Development Survey-II (IHDS-II), a nationally representative, a multi-topic survey of 42,152 households in 1,420 villages and 1,042 urban neighbourhoods across India. The goal of IHDS is to document changes in the daily lives of Indian households in an era of rapid transformation.

Both groups have two vital variables, one to identify a household and one to identify a member within a household. A combination of these gives us information on each individual. For both groups, we create a proxy variable for level of education, to shrink the number of categories to 3. This will make the analysis concise. The 3 categories are – '*niraakshar*' (literally meaning 'illiterate or equivalent'), '*shaala*' (literally meaning 'school or equivalent') and '*dip/grad*' (which is equivalent to 'a diploma or higher certificate'). This proxy variable is the same as in Inamdar (2018).

Regarding the dependent variables, both NSSO & IHDS-II contain information on the usual status of each individual. From this, we derive two dependent variables – *lfp & unemp*, which stand for labour force participation (LFP) and unemployed respectively. If usual_status is 'employed' or 'unemployed', it implies that LFP is 1 (indicates 'participation'), else if the value is 'out of labour force', LFP is 0 (indicates 'non participation'). Among those who participate, some may be unemployed. Only in those cases, unemp is 1. The following table shows the corresponding coding.

Table 1. Derivation of values of dependent variables from usual status

usual_status	Derived value of LFP	Derived value of unemp
Employed	1	0
Unemployed	1	1
Out of labour force	0	0

Considering the main motivation of this paper to compare and cross-validate data sources, we have taken care to see that both sets have the same context in terms of the time period (2011 – 2012) and similar coverage (all states of India). Since our area of interest is employment and economic activities, we restrict our study to potential earning individuals i.e. those in the age bracket of 15 – 64. The following table compares the salient elements of the designs of the two surveys.

Table 2. Comparison of salient elements of survey designs

Design element	Parameter	NSSO	IHDS-II
Respondents	Rural households	59,700	27,579
Respondents	Urban households	42,024	14,573
Respondents	Total households	**101,724**	**42,152**
Respondents	Women	280,763	135,118
Respondents	Men	176,236	69,450
Respondents	Total respondents	**456,999**	**204,569**

Design element	Parameter	NSSO	IHDS-II
Time period	Year	2011 – 2012	2011 – 2012
Coverage	Areas / locations	All Indian states	All Indian states
Methodology	Sampling	Stratified random sampling	Stratified random sampling

4. Comparative analysis

We bifurcate our comparative study, which makes it better to present the results. The first part contains descriptive statistics in the form of graphs. It helps to set the context of the comparison to follow. Here, we include both, female and male genders. The second part contains a list of tables, which show calculations of certain parameters based on statistical analysis. Here, we focus mainly on females within the sample sets and conduct deeper studies on two important problems – labour force participation (LFP) and unemployment.

Sub-section 4A – Graphs of basic parameters

1. Coverage – households

Table 3. Coverage of households (rural and urban)

Households	NSSO		IHDS-II	
Rural	59,700	58.7%	27,579	65.4%
Urban	42,024	41.3%	14,573	34.6%
Total	**101,724**	**100%**	**42,152**	**100%**

With a rather high (nearly 7%) difference in the split between rural and urban households, we expect differences from the two in the results on various parameters. One must state that although the absolute number of households in one survey is more than double that in the other survey, this difference is immaterial. As Singh (2018) says, the effect of size is marginal at most. The smaller survey also has a sizeable sample size, which allows for testing of statistical significance.

Coverage – individuals

Table 4. Coverage of individuals (rural and urban)

Individuals	NSSO		IHDS-II	
Rural	280,763	61.4%	135,118	66.1%
Urban	176,236	38.6%	69,450	33.9%
Total	**456,999**	**100%**	**204,569**	**100%**

In terms of individuals also, NSSO has a sample size of more than twice that of IHDS-II. Here, the difference in the split between rural and urban is smaller (5%). We observe that for IHDS-II, the split between rural and urban is similar at an individual level and at a household level. But, in NSSO, there is a 3% difference in the split between rural and urban at an individual level and at a household level. In this sense, the smaller survey seems more reliable.

2. Potential earners (ages 15 – 64)

Table 5. Number of potential earners (rural and urban)

Individuals	NSSO		IHDS-II	
Rural	182,683	60.2%	84,703	63.9%
Urban	120,682	39.8%	47,821	36.1%
Total	**303,365**	**100%**	**132,524**	**100%**

Since this study is related to economic activities, one must consider only those individuals in the age bracket of 15 years – 64 years. Just like the terminology in Inamdar (2018), this study shall use the terminology of 'potential earners'. The difference in the split between rural and urban narrows down further to about 3%. Here, we see that both NSSO and IHDS-II are consistent in terms of the rural-urban split in the general sample (all age groups) and the restricted sample (age groups 15 – 64). So, we feel that the two datasets are comparable.

3. Employment status (ages 15 – 64)

Table 6. Split of potential earners as per '*usual status*'

Individuals	NSSO		IHDS-II	
Employed	151,901	50.1%	64,893	49.0%
Unemployed	6,185	2.0%	2,776	2.1%
Out of labour force	145,279	47.9%	64,855	48.9%
Total	**303,365**	**100%**	**132,524**	

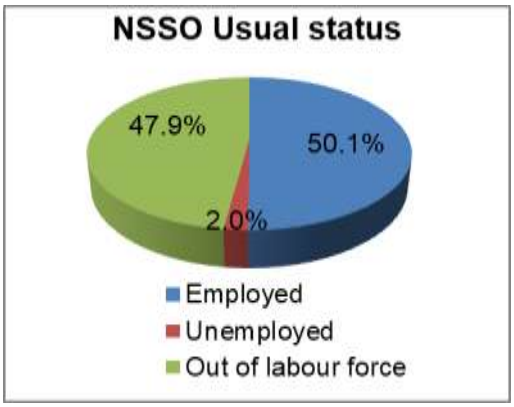

Figure 1. Usual status as per NSSO

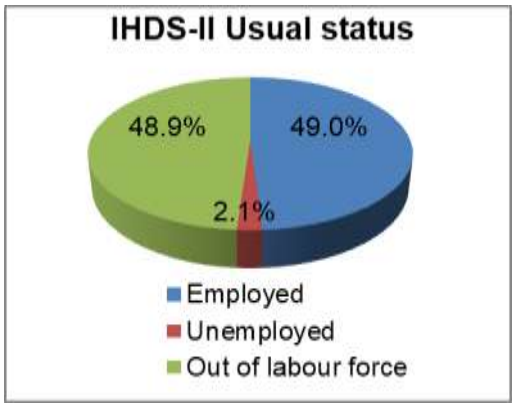

Figure 2. Usual status as per IHDS-II

For the restricted sample of potential earners, the proportions of employed, unemployed and out of labour force are very similar, with less than 1% difference. Thus, with respect to 'employment status', both the surveys are comparable.

4. LFPR & SPUR calculations (ages 15 – 64)

Table 7. Labour Force Participation Rate and Stock of Persons Unemployment Rate

Individuals	NSSO	IHDS-II
LFPR	52.1%	51.1%
SPUR	3.9%	4.1%

Labour Force Participation Rate (LFPR) = (Number of employed + Number of unemployed) / Total population

Stock of Persons Unemployment Rate (SPUR) = (Number of unemployed) / (Number of employed + Number of unemployed)

Since both datasets have very similar values of LFPR and SPUR, with a negligible difference, one can say that there is validation of the two.

5. Employment statuses of women & men (ages 15 – 64)

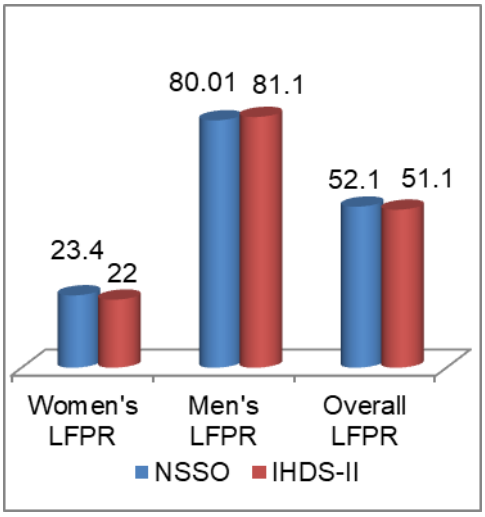

Figure 3. LFPR (percent)

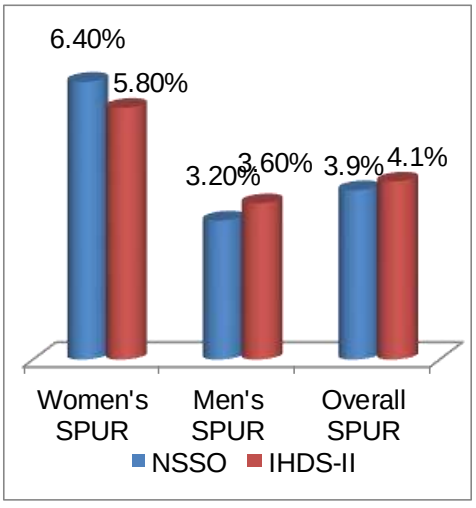

Figure 4. SPUR (percent)

This corroborates numerous studies highlighting the stark difference between LFPRs of women and men. Similarly, there is a stark difference between SPURs of women and men. Hence, it establishes a strong reason to study the economic activities of women separately. So, we restrict our study to women.

6. Income estimate

Next, we look at comparisons of estimates of income and compare three different parameters in this. First is the estimate of all 'potential earners' (ages 15 – 64) at a national level. Next is all potential earners at a state level. In particular, we look at the Indian state of Maharashtra, which is at the forefront when it comes to contribution to the national gross domestic product (GDP). Lastly, we compare the estimates exclusively for women at an all-India level, since that is the focus of this study. Note that NSSO and IHDS-II estimates of income follow different definitions and methodologies. So, the comparison between the two is not direct.

NSSO reports total earnings for a particular week for individuals. In as many as 265,678 (i.e. about 87%) cases, there are null values, hence, we ignore these cases for our calculations. Despite that, we have 37,787 cases with values, which are sufficient for statistical analysis. NSSO reports wages (cash plus kind) on a weekly basis and also mentions that on an average, each person is employed for 10 months in a year. On this basis, we multiply the weekly wages by 40 (equivalent to 40 weeks per year) to get the average annual income.

NSSO annual income (per individual) = NSSO weekly income * 40

IHDS-II gives the annual income of the entire household. It does not report individual income estimates. So, we adjust the per capita income of the household by taking into account only potential earners. The calculation is below.

1. Potential earners = Number of citizens in the age group of 15 – 64
2. Effective individual income = Household income / Potential earners

Simply dividing the household income by the household size (number of members) gives misleading results, since some members of every household are outside the earning age range.

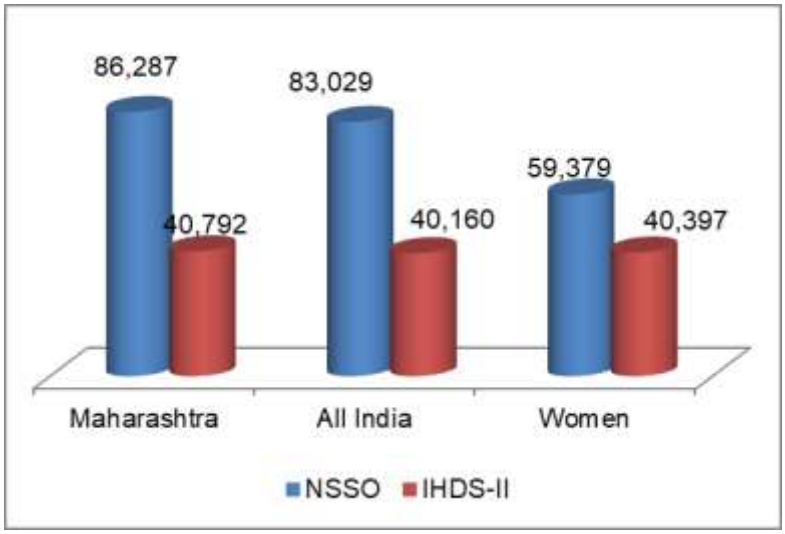

Figure 5. Estimates of average annual income (Rs.)

There is significant variation in the figures from the two sources, in absolute terms, with the NSSO estimates being more than double those of IHDS-II. Both say that the Maharashtra average income is higher than the national average income. The ratios of the former to the latter is 104% in the case of NSSO and 102% in the case of IHDS-II, which are very similar. However, the estimates of women's average differ vastly on two counts. Firstly, the women's average is lower in NSSO, while it is slightly higher in IHDS-II. Secondly, the ratio of women's average to the overall average is 78% for NSSO and 102% for IHDS-II. So, with respect to income, the datasets are different. Another factor which could partly account for this disparity is that the latter source has a much higher concentration of rural households (65%) than the former (58%). The average income in bound to be lower in this sector. There is also the distinct possibility of under-reporting of earnings on the part of respondents, to maintain secrecy.

Table 8. Income estimates and ratios with respect to national average

	NSSO		IHDS-II	
	Income	Ratio	Income	Ratio
Maharashtra avg.	86,287	104%	40,792	102%
National avg.	83,029	100%	40,160	100%
Women avg.	59,379	72%	40,397	101%

Sub-section 4B – Tables for LFPR statistical analysis

Having compared various parameters at an aggregate level, now, we restrict our study on women. From the income table and from previous literature, we know that from a social and economic perspective, women are vastly dissimilar to men. That is why in this sub-section, we focus on comparing parameters pertaining to the economic activities of women. The two important metrics to consider are LFPR and SPUR and the key factors which affect these two are **– sector, marital status, education level, relation to the household head, religion, social group and household size.** Of these, only the last one is an integer (parametric) variable, the others are all categorical (non-parametric).

We perform Probit analysis on LFPR & SPUR separately. In each case, we compare the results of NSSO & IHDS-II.

1. Determination of overall significance of factors for LFPR

First, we take each of the 7 factors, one at a time and run separate Probit tests, with LFPR as the dependent variable.

Comparison of Probit results using individual variables

Table 9. Results of Probit analysis to test for variables significant for LFPR

LFPR	NSSO		IHDS-II	
	X^2	Signif	X^2	Signif
Sector	841.726	0.000	437.383	0.000
Marital	3006.943	0.000	3872.388	0.000
Education	3345.774	0.000	704.892	0.000
Relation to head	5166.559	0.000	2521.498	0.000
Religion	2310.034	0.000	168.472	0.000
Social group	1812.402	0.000	754.981	0.000
Household size	59.707	0.000	242.033	0.000

All factors are significant, but there is a lot of variation in the X^2 values from the two datasets. Only for 'marital status', the values are similar to each other. Since all variable are significant, we can say that both datasets give similar results.

2. Identification of categories significant for LFPR

The earlier test assessed the relative importance of each factor at an overall level. Next, we perform t-tests using individual variables, by breaking each categorical variable into its various categories. This will help to determine respective coefficients.

Table 10. t-test for LFPR using values of the sector variable

a) Decomposition of variable categories – Sector

LFPR	NSSO			IHDS-II		
	t stat	std. err	Signif	t stat	std err	Signif
sector = rural	29.054	0.007	0.000	20.862	0.0116	0.000
sector = urban	n/a	n/a	n/a	n/a	n/a	n/a

With 'urban' as the reference, both NSSO and IHDS-II show that rural is significant for LFPR and their coefficients are very similar.

Table 11. t-test for LFPR using values of the marital status variable

b) Decomposition of variable categories – Marital status

LFPR	NSSO			IHDS-II		
	t stat	std. err	Signif	t stat	std err	signif
marital = divorced/ separated	10.329	0.0426	0.000	-9.390	0.0459	0.000
marital = married	-43.185	0.0135	0.000	-11.000	0.017	0.000
marital = never married	-48.421	0.0152	0.000	14.382	0.0178	0.000

LFPR	NSSO			IHDS-II		
	t stat	std. err	Signif	t stat	std err	signif
marital = widow(er)ed	n/a	n/a	n/a	n/a	n/a	n/a

With 'widowed/ widowered' as the reference, both NSSO and IHDS-II show that the categories '*married*', '*divorced/ separated*' and '*never married*' are significant for LFPR. However, their respective coefficients have opposite signs from the two datasets. Thus, we can say that the two sources are not similar.

Table 12. t-test for LFPR using values of the education level variable

c) Decomposition of variable categories – Education level

LFPR	NSSO			IHDS-II		
	t stat	std. err	Signif	t stat	std err	signif
edulevel = dipgrad	47.563	0.0119	0.000	21.415	0.021	0.000
edulevel = niraakshar	45.195	0.0077	0.000	-20.431	0.023	0.000
edulevel = shaala	n/a	n/a	n/a	n/a	n/a	n/a

With '*shaala*' as the reference, both NSSO and IHDS-II show that the categories '*dipgrad*' and '*niraakshar*' are significant for LFPR. The coefficients of '*dipgrad*' are similar. However, the coefficients of '*niraakshar*' have opposite signs from the two datasets. Thus, we can say that the two sources are not similar.

Table 13. t-test for LFPR using values of the relation to head variable

d) Decomposition of variable categories – Relation to household head

LFPR	NSSO			IHDS-II		
	t stat	std. err	Signif	t stat	std err	signif
relation = self	7.101	0.0138	0.000	34.000	0.02	0.000
relation = daughter-in-law	-78.440	0.0109	0.000	-25.062	0.0162	0.000
relation = grandmother	-32.160	0.0375	0.000	-9.980	0.051	0.000
relation = mother-in-law	-37.542	0.0236	0.000	-0.280	0.1252	0.000
relation = servant	0.000	0.0752	0.000	-0.573	0.1938	0.000
relation = daughter	-87.379	0.0103	0.000	-10.878	0.0148	0.685
relation = wife	n/a	n/a	n/a	n/a	n/a	n/a

With 'wife' as the reference, both NSSO and IHDS-II show that the categories '*self*', '*daughter-in-law*', '*mother-in-law*', '*servant*' and '*daughter*' are significant for LFPR. The coefficients have the same signs from the two datasets or are too small to make a difference. Thus, we can say that the two sources are similar.

Table 14. t-test for LFPR using values of the religion variable

e) Decomposition of variable categories – Religion

LFPR	NSSO			IHDS-II		
	t stat	std. err	Signif	t stat	std err	signif
religion = buddhism	10.967	0.042	0.000	-1.161	0.062	0.000
religion = christianity	16.771	0.029	0.000	-1.232	0.050	0.000
religion = islam	-12.261	0.028	0.000	2.680	0.046	0.000
religion = Jainism	-2.309	0.060	0.000	0.800	0.081	0.000
relation = Hinduism	n/a	n/a	n/a	n/a	n/a	n/a

With '*Hinduism*' as the reference, both NSSO and IHDS-II show that the categories '*Buddhism*', '*Christianity*', '*Islam*' and '*Jainism*' are significant for LFPR. However, their respective coefficients have opposite signs from the two datasets. Thus, we can say that the two sources are not similar.

Table 15. t-test for LFPR using values of the social group variable

f) Decomposition of variable categories – Social group

LFPR	NSSO			IHDS-II		
	t stat	std. err	Signif	t stat	std err	signif
social = obc	-1.901	0.0242	0.000	12.184	0.009	0.000
social = sc/st	-3.034	0.0356	0.000	23.608	0.010	0.000
social = general	n/a	n/a	n/a	n/a	n/a	n/a

With '*General*' as the reference, both NSSO and IHDS-II show that the categories '*OBC*' and '*SC/ST*' are significant for LFPR. However, their respective coefficients have opposite signs from the two datasets. Thus, we can say that the two sources are not similar.

Table 16. t-test for LFPR using values of the household size variable

g) Decomposition of variable categories – Household size

LFPR	NSSO			IHDS-II		
	t stat	std. err	Signif	t stat	std err	Signif
hh_size	-7.857	0.001	0.000	1785.714	0.003	0.000

As an integer variable, this does not require a reference category. Both NSSO and IHDS-II show that it is significant for LFPR. However, their coefficients have opposite signs from the two datasets. Thus, we can say that the two sources are not similar.

3. Combination of all variables

Next, we consider a combination of all variables together. In the Probit model, along with the individual variables, we include the interaction effects as well. The table below shows the results of the individual variables and their interaction effects.

Table 17. Probit analysis for LFPR using a combination of all variables

LFPR	NSSO		IHDS-II	
Variable	X^2	Signif	X^2	Signif
Sector	235.467	.000	26.239	.000
Marital	108.356	.000	141.697	.000
Education	614.639	.000	246.527	.000
Relation to head	456.015	.000	178.986	.000
Religion	270.778	.000	21.216	.000
Social	62.097	.000	131.145	.000
Household size	.005	.944	.044	0.833
sector * marital * edulevel * relation * religion * social * hh_size	2858.637	0.000	1294.890	.254

For LFPR, '*household size*' is not significant in either NSSO or IHDS-II. All other variables are significant in both. The interaction is significant as per NSSO and not significant as per IHDS-II. Considering this discrepancy, one can say that the two sources are not analogous to each other.

Sub-section 4C – Tables for SPUR statistical analysis

1. Determination of overall significance of factors for SPUR

When we run separate Probit tests, with SPUR ('*unemp*') as the dependent variable, we get the following results.

Table 18. Results of Probit analysis to test for variables significant for SPUR

SPUR	NSSO		IHDS-II	
	X^2	Signif	X^2	Signif
Sector	71.175	0.000	32.292	0.000
Marital	1778.664	0.000	3167.527	0.000
Education	2534.209	0.000	1100.663	0.000
Relation to head	1954.019	0.000	1151.176	0.000
Religion	374.693	0.000	112.665	0.000
Social group	146.911	0.000	6273953.453	0.000
Household size	100.998	0.000	2.525	0.112

Here too, we see that the X^2 values are very different from one another. More importantly, while we see that in NSSO, all variables are significant, in IHDS-II, the variable 'household size' is not significant. Considering this difference, one can say that the two datasets give dissimilar results for '*unemployment*'.

2. Identification of categories significant for SPUR

The earlier test assessed the relative importance of each factor at an overall level. Next, we perform t-tests using individual variables, by breaking each categorical variable into its various categories. This will help to determine respective coefficients.

Table 19. t-test for SPUR using values of the sector variable

a) Decomposition of variable categories – Sector – urban (reference), rural

SPUR	NSSO			IHDS-II		
	t stat	std. err	Signif	t stat	std err	Signif
sector = rural	-8.443	0.0167	0.000	-5.677	0.0266	0.000
sector = urban	n/a	n/a	n/a	n/a	n/a	n/a

With 'urban' as the reference, both NSSO and IHDS-II show that rural is significant for SPUR and their coefficients are very similar. The negative sign indicates that belonging to the rural sector decreases the likelihood of unemployment vis-a-vis the urban sector.

Table 20. t-test for SPUR using values of the marital status variable

b) Decomposition of variable categories – Marital status

SPUR	NSSO			IHDS-II		
	t stat	std. err	Signif	t stat	std err	signif
marital = divorced/ separated	9.172	0.1039	0.000	3.629	0.1174	0.000
marital = married	5.876	0.0679	0.000	-0.713	0.0631	0.000
marital = never married	16.465	0.0679	0.000	16.483	0.0617	0.000
marital = widow(er)ed	n/a	n/a	n/a	n/a	n/a	n/a

With 'widowed/ widowered' as the reference, both NSSO and IHDS-II show that the categories '*married*', '*divorced/ separated*' and '*never married*' are significant for SPUR. The coefficients of 'divorced/ separated' and 'never married' are similar. However, the coefficients of '*married*' have opposite signs from the two datasets. Thus, we can say that the two sources are not similar.

Table 21. t-test for SPUR using values of the education level variable

c) Decomposition of variable categories – Education level

SPUR	NSSO			IHDS-II		
	t stat	std. err	Signif	t stat	std err	signif
edulevel = dipgrad	40.408	0.0196	0.000	25.814	0.013	0.000

SPUR	NSSO			IHDS-II		
	t stat	std. err	Signif	t stat	std err	signif
edulevel = niraakshar	-19.515	0.0309	0	11.200	0.008	0.000
edulevel = shaala	n/a	n/a	n/a	n/a	n/a	0.000

With '*shaala*' as the reference, both NSSO and IHDS-II show that the categories '*dipgrad*' and '*niraakshar*' are significant for SPUR. The coefficients of '*dipgrad*' are similar. However, the coefficients of '*niraakshar*' have opposite signs from the two datasets. Thus, we can say that the two sources are not similar.

Table 22. t-test for SPUR using values of the relation to head variable

d) Decomposition of variable categories – Relation to household head

SPUR	NSSO			IHDS-II		
	t stat	std. err	Signif	t stat	std err	signif
relation = self	-11.540	0.0526	0.085	-0.559	0.111	0.085
relation = daughter-in-law	12.302	0.0278	0.000	1.583	0.0638	0.000
relation = grandmother	6.978	0.0642	0.000	11.511	0.0834	0.000
relation = mother-in-law	-3.960	0.1192	0.021	1.143	0.3657	0.000
relation = servant	0.000	0.2545	0.436	-0.001	3752.135	0.021
relation = daughter	32.987	0.0231	0.000	28.398	0.0412	0.000
relation = wife	n/a	n/a	n/a	n/a	n/a	n/a

With '*wife*' as the reference, both NSSO and IHDS-II show that the categories '*daughter-in-law*', '*mother-in-law*' and '*daughter*' are significant for SPUR. The coefficients have the same signs from the two datasets. However, in both cases, the category '*self*' is not significant. Also, in NSSO, '*servant*' is not significant whereas in IHDS-II, it is significant. Thus, we can say that the two sources are similar.

Table 23. t-test for SPUR using values of the religion variable

e) Decomposition of variable categories – Religion

SPUR	NSSO			IHDS-II		
	t stat	std. err	Signif	t stat	std err	signif
religion = Buddhism	1.473	0.104	0.000	0.179	0.134	0.000
religion = christianity	9.440	0.068	0.000	-3.987	0.095	0.000
religion = islam	4.142	0.069	0.000	-0.788	0.090	0.000
religion = Jainism	-0.185	0.173	0.000	-0.732	0.167	0.000
religion = Hinduism	n/a	n/a	0.000	n/a	n/a	n/a

With '*Hinduism*' as the reference, both NSSO and IHDS-II show that the categories '*Buddhism*', '*Christianity*', '*Islam*' and '*Jainism*' are significant for SPUR. The coefficients of '*Buddhism*' & '*Jainism*' are similar. However, the coefficients of the other two categories have opposite signs. Thus, we can say that the two sources are not similar.

Table 24. t-test for SPUR using values of the social group variable

f) Decomposition of variable categories – Social group

SPUR	NSSO			IHDS-II		
	t stat	std. err	Signif	t stat	std err	signif
social = obc	11.364	0.011	0.000	-4.167	0.012	0.000
social = sc/st	6.376	0.015	0.000	n/a	n/a	n/a

SPUR	NSSO			IHDS-II		
	t stat	std. err	Signif	t stat	std err	signif
social = general	n/a	n/a	n/a	n/a	n/a	n/a

With '*General*' as the reference, both NSSO and IHDS-II show that the categories '*OBC*' and '*SC/ST*' are significant for SPUR. However, their respective coefficients have opposite signs from the two datasets. Thus, we can say that the two sources are not similar.

Table 25. t-test for SPUR using values of the household size variable

g) Decomposition of variable categories – Household size

SPUR	NSSO			IHDS-II		
	t stat	std. err	Signif	t stat	std err	Signif
hh_size	-10.278	0.004	0.000	-28.462	0.001	0.000

As an integer variable, this does not require a reference category. Both NSSO and IHDS-II show that it is significant for SPUR. Their coefficients have the same sign from the two datasets. Thus, we can say that the two sources are similar.

3. Combination of all variables

Next, we consider a combination of all variables together. In the Probit model, along with the individual variables, we include the interaction effects as well. The table below shows the results of the individual variables and their interaction effects.

Table 26. Probit analysis for SPUR using a combination of all variables

SPUR	NSSO		IHDS-II	
Variable	X^2	Signif	X^2	Signif
Sector	6.961	.008	.455	.500
Marital	5.738	.220	2.813	.590
Education	251.190	.000	28.366	.000
Relation to head	72.655	.000	5.875	.882
Religion	7.412	.284	9.053	.107
Social	30.268	.000	3.560	.614
Household size	0.276	0.599	.088	.767
sector * marital * edulevel * relation * religion * social * hh_size	578.426	1.000	175.917	1.000

For SPUR, only '*education*' is significant in IHDS-II. Only '*sector*', '*education*' and '*relation to head*' are significant as per NSSO. Considering this discrepancy, one can say that the two sources are not analogous to each other.

Sub-section 4D – Tests for normality of distributions

Finally, we conduct the Kolmogorov-Smirnov (K-S) test on the annual income variable to test whether its distribution is normal. We do it separately on the two data sources. After these 1-sample tests, we combine the two sources and perform a 2-sample test.

1. Separate 1-sample K-S tests on annual income

Table 27. Results of independent 1-sample K-S tests on annual income

Variable	NSSO	IHDS-II
Sample size (N)	52,456	40,950
Mean (Avg. value)	59,379	40,397
Standard deviation	159,801	55,232
Kolmogorov-Smirnov Z	66.945	65.658
Asymptotic significance (2-tailed)	.000	.000

Here, both NSSO & IHDS-II have almost identical values of the Z statistic. Both values are significant. NSSO has a much greater standard deviation than IHDS-II. We infer that the income distributions of both data sources follow a normal distribution and are comparable.

2. Combined 2-sample K-S test

For the two-sample test, we take the 'income' values from both sources. Since this test requires a grouping variable, we code NSSO as '1' and IHDS-II as '2'. From the table, we see that the K-S statistic is significant. So, we infer that the two samples are from the same distribution. That suggests that the income distribution of both, NSSO and IHDS-II is normal.

Table 28. Results of 2-sample K-S tests on annual income

Variable	2-sample results
Sample size (N)	93,406
Mann-Whitney U	390,990,666
Kolmogorov-Smirnov Z	78.645
Asymptotic significance (2-tailed)	.000

5. Limitations and conclusions

This study has two important limitations. One, it considers only two sources of data, both of which have been the subject of criticism over the lack of reliability of the data. For a stronger comparison, one must take 3 data sources into account and compare them. That will bring greater confidence to the data validation exercise. Singh (2017) is a good example of a comparative study, which takes many major states of India to carry out benchmarking. Two, while the study points out all aspects where the two datasets corroborate or don't corroborate each other, it offers no explanation in cases where the results are dissimilar. This is especially true of the income variable. Both, NSSO and IHDS-II have different methods to estimate and report incomes. This paper is an attempt to adjust them to a common metric. However, such an approach involves its own assumptions, as we have pointed out earlier. Given the sensitive nature of data related to economic status, establishment of those assumptions is outside the scope of this study.

In conclusion, one can state that both NSSO and IHDS-II have fairly large datasets. Their coverages are slightly different in the sense that the former has a smaller proportion of rural households (58%) than the latter (65%). On other demographic factors, both sources are largely similar. Their results related to employment status are almost identical, which suggests overall congruence between them. Hence, the overall LFPR and SPUR from the two sources are almost identical. With respect to LFPR, all factors – sector, marital status, education, relation to head, religion, social group and household size – are significant in both datasets. This result is common to both data sources. However, with respect to SPUR, there is an important difference. In IHDS-II, the variable '*household size*' is not significant whereas it is significant in NSSO. In this sense, the two sources are not comparable. One of the reasons for this could be multicollinearity i.e. interaction among the variables. So, we perform one more test, taking all variables and also including their interaction effect. What we observe is that for LFPR, both the data sources are similar, such that 6 out of 7 variables are significant (with the 7th being household size). For SPUR, one observes starkly different results. Except for '*education*', all variables are insignificant in IHDS-II. NSSO is not much better with only 'sector' and '*relation to head*' being significant.

When we conduct the 1-sample K-S test on the income variable individually, we see that both data sources have a normal distribution. When we conduct the 2-sample K-S test together on the income variable, we see that both samples come from the same distribution. So, this result is valid.

Considering the series of tests above and their corresponding results, the summary is that NSSO and IHDS-II datasets are more dissimilar than similar. This implies that studies based on the two are likely to give different results. The future scope for research based on this is to study the factors driving '*unemployment*' more closely. It will necessitate standardisation of incomes, without which statistical analyses will not

produce reliable results. That area remains unexplored in the Indian context. Only after we address this can we build the necessary confidence in a comparison exercise.

References

1. *Minhas, B (1988), 'Validation of Large Scale Sample Survey Data Case of NSS Estimates of Household Consumption Expenditure', Sankhyā: The Indian Journal of Statistics, Series B, Vol. 50, No. 3, pp. 49 – 63.*
2. *Sorensen, H, Sabroe, S & Olsen, J (1996), 'A Framework for Evaluation of Secondary Data Sources for Epidemiological Research',* International Journal of Epidemiology, *Vol. 25, No. 2, pp. 435 – 442.*
3. *Sundaram, K & Tendulkar, S (2001), 'NAS-NSS Estimates of Private Consumption for Poverty Estimation: A Disaggregated Comparison for 1993-94', Economic and Political Weekly, Vol. 36, No. 2, pp. 119 – 129.*
4. *Choudhury, M (2002), 'Potential Selectivity Bias in Data: Evaluation of a Firm-Level Database on Indian Industry', Economic and Political Weekly, Series B, Vol. 37, No. 8, pp. 757 – 766.*
5. *Hirway, I (2002), 'Employment and Unemployment Situation in 1990s: How Good Are NSS Data?', Economic and Political Weekly, Vol. 37, No. 21, pp. 2027 – 2036.*
6. *Bhattacharya, K (2007), 'How Good is the Bankscope Database? A Cross-Validation Exercise with Correction Factors for Market Concentration Measures', Bureau of Indian Standards, No. 133, 2003.*
7. *Berkhout, J & Lowery, D (2008), 'Counting organized interests in the European Union: a comparison of data sources', Journal of European Public Policy, Vol. 15, No. 4, pp. 489 – 513.*
8. *Müller, K (2009), 'How robust are simulated employment effects of a legal minimum wage in Germany? A comparison of different data sources and assumptions', German Institute for Economic Research, No. 900, 2009.*
9. *Oldiges, C (2012), 'Cereal Consumption and Per Capita Income in India', Economic and Political Weekly, Vol. 47, No. 6, pp. 63 – 71.*
10. *Saikia, N & Kulkarni, P (2012), 'Data for Research into Health Inequality in India Do We Have Enough?', Economic and Political Weekly, Vol. 51, No. 26, pp. 111 – 116.*
11. *Singh, A. (2016), 'Do Technology Spillovers Accelerate Performance of Firms? Unravelling a Puzzle from Indian Manufacturing Industry'. Annals of the University Dunarea de Jos of Galati: Fascicle: XVII, Medicine, Vol. 22, No. 3.*
12. *Singh, A. (2016), 'R&D spillovers & productivity growth: evidence from Indian manufacturing', Indian Journal of Industrial Relations, Vol. 51, No. 4, pp. 563 – 579.*
13. *Singh, A. (2017), 'Merge or Acquire-A Strategic Framework', Annals of the University Dunarea de Jos of Galati: Fascicle: XVII, Medicine, (3).*
14. *Singh, A. (2017), 'Does FDI Promote Productivity? A Deep Dive', Indian Journal of Industrial Relations, Vol. 52, No. 3.*
15. *Singh, A. (2017), 'Keeping It Simple: Comparative Analysis of TFP across Manufacturing Industries and Major States of India', Theoretical Economics Letters, Vol. 7, No. 06, pg. 1821.*
16. *Desai, S, Vanneman, R (2018), 'India Human Development Survey (IHDS), 2005', University of Maryland, National Council of Advanced Economic Research (NCAER).*
17. *Singh, A (2018), 'Does Size Matter? – The Effect of Size of Production Workers, Management Staff and Proprietors on Productivity', Theoretical Economics Letters, Vol. 8, No. 11, pg. 2290.*
18. *Sarkar, S, Sahoo, S & Klasen, S (2018), 'Employment transitions of women in India: A panel analysis', World Development, Vol. 115, No. 3, pg. 291 – 309.*