Social Science Spectrum ISSN 2454-2806
Vol. 1, No. 4, December 2015, pp. 239-248

Variance Decomposition of Macro Level Data with Application to Unemployment in Northeast India

Laishram Ladusingh*

Abstract

Analysis based on aggregated data is referred to as ecological analysis and the discrepancy arising when the analysis based on area level gives conclusion very different from expected relationship at unit level is termed as ecological fallacy. It is for this reason that census data are under-utilized for the exploration of cause and effect relationship. The uses of aggregate data from the census are confined to projection, estimation of rates, trend analysis and compositional changes. The three sources of ecologic fallacy pertain to situations where there is some form of group effect. This includes situations where there is a failure to distinguish constructs at different levels (e.g., mean group X is assumed to measure the same thing and individual-level X), where something about the groups is associated with individual-level predictors of the outcomes (mean group X is associated with other individual-level factors related to Y), or where some social process results in the grouping of persons by the dependent variable. This paper describes statistical technique which can be adopted for macro level analysis by decomposing variance under multilevel regression model.

Key words: Macro data, multilevel regression, variance decomposition.

Introduction

Indian census is the main source of macro level data on socio-demographic, occupation, migration and employment status pertaining to the people and places. However, census data are under-utilized for the exploration of cause and effect relationship and the uses of aggregate data from the census are confined to projection, estimation of rates, trend analysis and compositional changes. There are reasons for not using macro level aggregate data for the analysis of cause and effect relationship. One of them is little progress in developing methods which can be applied to group level data to estimate unit level relationships (Steel & Holt, 1996). Analysis based on aggregated data is referred to as *ecological analysis* and the discrepancy arising when the analysis based on area level means conclusion very different from expected relationship at unit level is termed as *ecological fallacy*. The three sources of the ecologic fallacy pertain to situations where there is some form of group effect. This includes situations where there is a failure to distinguish constructs at different levels (e.g., mean group *X* is assumed to measure the same thing and individual-level *X*), where something about the groups is associated with individual-level predictors of the outcomes (mean group *X* is associated with other individual-level factors related to *Y*), or where some social process results in the grouping of persons by the dependent variable (Firebaugh, 1978; Hammond, 1973). The focus of this paper is to understand statistical methods appropriate for macro level unemployment data in Census of India 2001 without cropping in *ecological fallacy*.

An accompanying objective of the present project is to find compositional and contextual backgrounds of unemployment for Northeast India. Compositional factors relate to individual items such as caste, sex, educational status, etc., while on the other hand the place of residence in rural-urban areas, state of residence etc., are the contextual background. We intend to answer the question whether unemployment in Northeast India is compositional or contextual or both? The underlying notion is that if unemployment is compositional, similar type of people will face similar unemployment problems regardless of where they live. But if it is contextual, unemployment status

* Laishram Ladusingh, Professor of Demography & Statistics, International Institute for Population Sciences, Govandi Station Road, Deonar, Mumbai – 400088. Email: lslaishram@iips.net

would be the same for people living in similar types of places. This conceptualization of compositional and contextual background is based on Subramanian et al. (2000). In case contextual background is significant, similar type of people would experience unemployment unevenly, depending on where they are, thus inducing spatial variation in prevalence of unemployment.

Review of literature

One reason for the frequent use of aggregate data is that individual-level data may be limited, while aggregate data containing the relevant information may be readily available from administrative sources or population census and may offer valuable clues about individual behaviour. Sometimes because of confidentiality or ethical issues on making unit level data public, researchers wishing to investigate individual level relationship are compelled to use aggregated macro level data. The aggregate data are usually in the form of means or percentages for a set of groups into which the population has been partitioned. It is well known that statistical analysis based on aggregated data, such as area or group means, may be invalid because of the ecological fallacy. This fallacy occurs when analyses based on area level means give conclusions very different from those that would be obtained from an analysis of unit level data, if they were available. For example, Robinson (1950) obtained a correlation coefficient of 0.11 between illiteracy and being foreign born from personal level data but when he calculated the correlation between percentage of illiterate and percentage of foreign born, at the state level he obtained a correlation of -0.53. He clarified the ecological correlation problem by stating mathematically the exact relationship between ecological correlation and individual correlation, and by showing the bearing of that relation upon the practice of using ecological correlations as substitute for individual correlations. Menzel (1950) on the basis of Robinson's own examples went on to show that ecological correlations may retain their validity even after it has been shown that the ecological and individual correlations clearly differ. Ecological correlations may tell something about territorial units, which can be used as contextual properties explaining the variations in the correlated variables.

Empirical demonstration of discrepancy between statistics calculated from macro and individual level date back to Gekhle & Biehl (1934) and also include Yule & Kendall (1950), Blalock (1964), Clark & Avery (1976), Openshaw & Taylor (1979), Erson & Lane (1983), Arbia (1989) and Fotheringham & Wong (1991). Whilst the size of such effects has been investigated, little progress has been made in developing methods which can be applied to group level data to estimate unit level relationships. For the first time Goodman (1959) considered a model under the regression framework for aggregated data analysis which could validly be used to draw inferences regarding relationships at the individual level. Subsequent models proposed by Dogan & Rokkan (1969), Hannan & Burstein (1974), Litchman (1974), Smith (1977) and Blalock (1979, 1985), take into account the group formation or the interaction between group and individual level variables either implicitly or explicitly to deal with ecological analysis and ecological fallacy.

In a recent development, Steel and Holt (1996) proposed methods of adjusting group level analyses to produce estimates of unit level parameters and validity of the results were discussed at length. They showed that the differences between unit and group level analyses may be explained through a model which incorporates the effect of variables which characterize to which area unit individuals belong, together with area level characteristics which produce correlations between individuals in the same areal unit or group. Methods are suggested which, together with some auxiliary information, can be used to adjust area level analyses to provide less biased estimates of unit level parameters. In addition, the identification of a small and convenient set of grouping variables is an important aspect of the proposed methodology and procedures are suggested to achieve it. These grouping variables may be important in providing a substantive explanation of the population structure. The result is a strategy for the analysis and adjustment of aggregation effects in multivariate statistical analysis.

Goodman (1963) had presented some techniques for estimating individual correlation on the basis of the ecological data, using both qualitative and quantitative variables. In most of the models treated by him, it is assumed that the ecological relationship is only a reflection of an individual-level relationship, in which case the ecological slope can be used as an estimate of the individual slope. Subramanian et al. (2000) under the structure of multilevel regression suggested an approach for analysing aggregate data on illiteracy of Census of India, 1991 to examine whether variations in illiteracy relate to the type of people in particular places or to the characteristics of places. Brown et al. (2005) improve the work of Subramanian's co-workers with provision for controlling over dispersion when the response variable is a proportion where the binomial assumption is no longer valid.

It is hypothesized that multilevel research that attempts to describe ecological effects in themselves is at risk of reaching beyond an epidemiological understanding of what constitutes an ecological effect, and what sources of error may be influencing an observed ecological effect. A direct cross level effect, cross level effect modification and an indirect cross level effect are the three basic types of ecological effects. Sources of error and weaknesses in study design that may affect estimates of ecological effects include a lack of variation in the ecological exposure in the available data not allowing for intra-class correlation, selection bias confounding at both the ecological and individual levels, misclassification of variables, misclassification of units of analysis and assignment of individuals to those units, model mis-specification and multi-collinearity. Identification of ecological effects requires the minimization of these sources of error and a study design that captures sufficient variation in the ecological exposure of interest (Blakely and Woodward, 2000).

Aggregated data are also frequently used in every field of economics to explain individual behaviour. But data aggregation can result in misleading conclusions regarding the economic behaviour of individuals. Garrett (2002) developed a simple framework to show how coefficient estimates and their statistical significance can differ using aggregated versus less aggregated data. Only when we make an assumption that the hypothesized relationship between the economic variables in question is homogenous across all individuals, aggregated data can be used to explain individual behaviour.

The use of aggregate data alone to make an inference about individual-level relationships can introduce bias, leading to the ecological fallacy. The ecological fallacy arises from confounding of the individual-level relationship due to heterogeneity in exposure variable of interest and other covariates within groups (Rothman & Greenland, 1998). One approach to address concerns regarding ecological fallacy has been the development of multilevel modeling (Goldstein, 2003). Research involving a combination of both individual-level and aggregate data has a long history in epidemiology (Chambers & Skinner, 2003). Alterman et al. (2001) examined the adjusted association between job characteristics and select causes of death applying multilevel modeling to a combination of individual–level and aggregate data. The results were compared to those obtained from logistic regression modeling which clearly demonstrated the potential of drawing incorrect conclusions when multilevel modeling is not used.

In multilevel studies, multilevel models allow simultaneous examination of the effects of group-level and individual- level variables on individual-level outcomes while accounting for non-independence of observations within groups. They also allow the examination of both between group and within-group variability, as well as how group level and individual-level variables are related to between group, within-group, and total inter-individual variability in the outcome. Thus, multilevel models can be used to draw inferences regarding the causes of inter-individual variation (or the relation of group-level and individual-level variables to individual-level outcomes), but inferences can also be made regarding intergroup variation, whether it exists in the data and to what extent it is accounted for by group-level and individual-level characteristics (Roux, 2004).

Mattei & Derivry (1988) showed that it is possible to demonstrate the impact of social context on individual behaviour and at the same time avoid the risk of ecological fallacy by conducting simulated experiments which involves analysing aggregate data in statistically constructed social contexts.

Need for the study

It is well known that statistics calculated from data aggregated to group level from the macro level data in Census can be very different from those calculated from individual level data. In order to enhance use of Census macro level data beyond trend analysis, estimation of rates, projection and spatial distribution, there is an urgent need to acquaint and demonstrate effective use of recent methodological advances to explore the published macro data available from Indian Census. Keeping this in view, the present project is devoted to analysis of district level data on unemployment from Census of India, 2001 to disentangle compositional and contextual aspects of unemployment in Northeast India as demonstration.

Data sources and methodology

The source of data for this project is the Census of India, 2001 on working status. We have used information on seeking/available for work and these macro level data are available at the district level by sex, caste, residence background and literacy status. Age group of 15-60 years' population is being considered in obtaining the proportion of unemployed population. Individual level data collected in the census are aggregated into count data at district level and are further aggregated into count data at state level. As such, census data clearly exhibit hierarchy in data structure, individuals within districts clustered within states. Besides, district level count of unemployment is an aggregate of binary numbers 1s for unemployed individuals and 0s for employed individuals. This clearly suggests hierarchical interdependence in census data and it is important to adopt a statistical method which recognized dependence in data structure. Multilevel regression models (Goldstein, 2003; Bryk & Raudenbush, 1992), can account for the inter-dependence of observations by partitioning the total variance into different components of variation due to various cluster levels in the data. Though partitioning variances do not pose any methodological challenge in dealing with a continuous dependent variable with a normal error distribution at each level, the extension to models where the response is a proportion is less obvious. However, from the census, as mentioned above, data routinely available are merged into proportion (e.g., proportion of unemployed) at aggregate (district, state, etc.,) level. Brown et al. (2005) described a generalized binomial response model to deal with proportion response and ecological fallacy by incorporating an additional layer in the multilevel logistic regression model due to Goldstein et al. (2002).

We now describe the procedure for preparation of inputs including response from Census data. Unemployed counts at district level are available by sex, literacy status, caste, residence background and state. Total population in the 15-60 years group and population not working and seeking work at the district level are categorized into 2x2x3x2 cells, considering all possible combination of sex, literacy status, caste and residence background. For each of the possible combination which shall be referred as cell, proportion of unemployed is computed and treated as the response variable, while sex, literacy status, caste and residence background are considered as covariates. Besides, we have imported net state domestic product (nsdp), percentage of agricultural workers and percentage of population below the poverty line as state level factors and district literacy rates into the present study.

For each cell (i) within districts (j), within (k), the counts of individuals (y_{ijk}) can be model as binomial distribution, that is,

$$y_{ijk} \sim Binomial(n_{ijk}, p_{ijk})$$

where, p_{ijk} is the probability of being unemployed and n_{ijk} is the number of unemployed persons. Assuming n_{ijk} observations are independent and conditional on the estimates of probabilities (p_{ijk}) for the i^{th} cell, which correspond to combinations of compositional and contextual characteristics. We model p_{ijk} as

$$logit(p_{ijk}) = X'_{ijk}\beta + v_k + u_{jk} + e_{ijk}$$

$$v_k \sim N(0, \sigma_v^2)$$

$$u_{jk} \sim N(0, \sigma_u^2)$$

$$e_{ijk} \sim N(0, \sigma_e^2)$$

Proportion of variance at state level=$\sigma_v^2/(\sigma_v^2 + \sigma_u^2 + \sigma_e^2)$

Proportion of variance at district level=$\sigma_u^2/(\sigma_v^2 + \sigma_u^2 + \sigma_e^2)$

Sources of data on unemployment and employment

Statistics on employment and unemployment can be collected either through censuses of population and economic establishments and national level sample surveys (NSS), or through returns under various Acts relating to regulation of economic establishments regarding workers, etc. Employment Exchanges recording job seekers, etc., can be another source. The two most important sources of data on employment/unemployment are Census and NSS. It has been the tradition in the population census of India to collect information on the economic activity of the people. The Census data provide an inventory of human resources of the country showing their number, characteristics, occupation and distribution among various branches of economy. The measurement of economic activity has been attempted in every census of the country even though there has been variation in the concepts adopted from time to time.

The NSSO collects data through sample surveys based on the scientific technique of random sampling through household enquiry both in rural and urban areas. In a number of earlier rounds, it experimented with various concepts and methodologies in trying out and standardizing a proper framework to estimate in quantitative terms the characteristics of labour force, employment, unemployment and under-employment. The experimental surveys were followed by regular annual sample surveys till the late sixties. Afterwards, the quinquennial surveys started in 27th round (1972-73), with the follow-up surveys in 32nd round (1977-78), 38th round (1983), 43rd round (1987-88), 50th Round (1993-94), 55th round (1999-2000) and 61st round (2004-05). The NSSO also gives annual estimates of employment and unemployment on the basis of thin sample in each round since its 45th round (1989-1990) in its annual series.

The NSS is considered to be the only reliable source of information on unemployment by many researchers (Krishnamurthy, 1988). However, the collection of data on any item through the national census has its own advantages in terms of wide coverage, level of regional disaggregation of the data and feasibility of cross-classification by a fairly large number of characteristics (Kulkarni & Kumar, 1989).

Census on the whole can be considered to be a more reliable source of information with regard to the 'work force participation rate' in the country than the NSS employment surveys since it has a much larger and comprehensive coverage of the population in relation to the NSS sample based estimates and, therefore, is closer to the actual picture. The NSS data, moreover, depend on the census data to calculate appropriate multipliers to inflate samples so as to be representative of the overall central and state populations. However, the NSS figures score much higher on their conceptual precision and depth of information on various aspects of work status of individuals and related variables.

Unemployment scenario in Northeast India

Despite the fact that literacy rate in the north-eastern region in India (68.77) is above the national average (65.38), the employability is low resulting in high rate of unemployment and underemployment. The region, according to the Ministry of the Development of the Northeast Region (DONER), has a net unemployment rate of 12 per cent. It is a startling revelation that more than half a million young and educated people in India's north-eastern states are jobless. At least 600,000 educated young people in Manipur are unemployed, which accounts for a quarter of Manipur's 2.5 million people, according to the Chief Minister of Manipur. In desperation they are resorting to low-income jobs and at times even militancy. There is a high demand for government jobs in northeast even if they are "high-risk jobs" (Datta, 2009).

The rate of employment generation in Assam is much lower compared with the national average. In 1999-2000, the national unemployment rate was 2.3per cent while the same was a staggering 4.6per cent in the case of Assam. The unequal distribution of population and the rural-urban divide have also led to a disparity in the employment opportunities. While schemes like NREGA (National Rural Employment Guarantee Scheme) have managed to increase employment opportunities significantly in Assam, the impact has been more pronounced in the rural areas and some of the most backward districts of Assam have witnessed a reduction in the number of people migrating to towns to earn their livelihood.

In the absence of major industrial establishment and other employment opportunities in the region, unemployment rate, particularly of urban educated youths, is not only high but also increasing rapidly. According to the Current Daily Status (CDS), as shown in Figures 1 & 2, unemployment level is the highest in urban areas of Tripura at both the points of time, followed by Assam which is also higher than the national level.

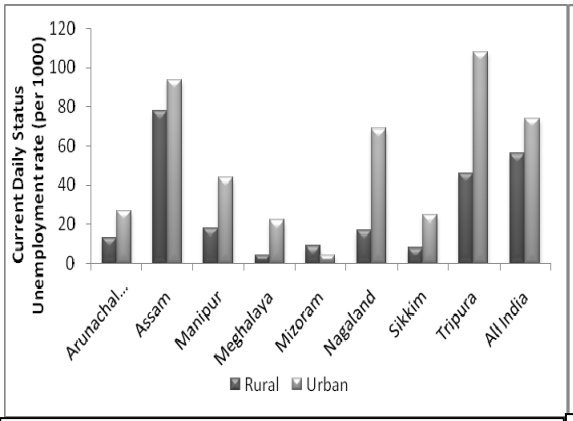

Figure1: Current Daily Status Unemployment Rates (per 1000) by Place of Residence in Northeast States, 1993-94

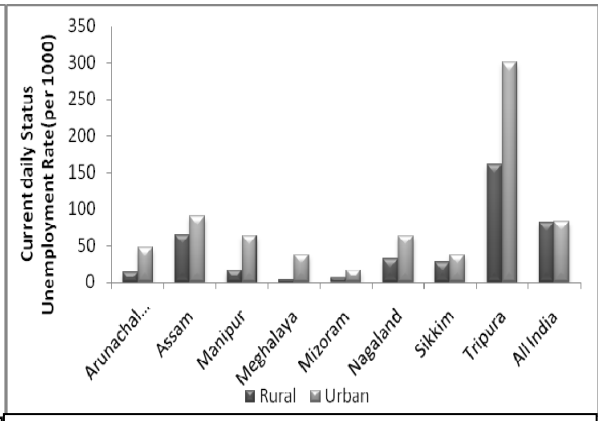

Figure2: Current Daily Status Unemployment Rates (per 1000) by Place of Residence in Northeast States, 2004-05

Sources: NSSO (1997): Employment and Unemployment in India, 1993-94, 50th Round, Report No. 409.
NSSO (2006): Employment and Unemployment Situation in India, 2004-05, 61st Round, Report No. 515.

In the remaining states, though lower than the national average, what is more intriguing is that the volume of unemployment in absolute terms is growing in the post-globalization period. It is this section of the society which becomes easy prey to negative elements like insurgency and drug abuse. Both rural and urban unemployment rates have increased drastically from 1993-94 to 2004-05 in Tripura, but the increase in urban unemployment rate has been comparatively more. In spite of the fact that unemployment has increased at the national level, Assam has witnessed a decline in both rural and urban unemployment rates in 2004-05. However, the decline has been more for the rural areas.

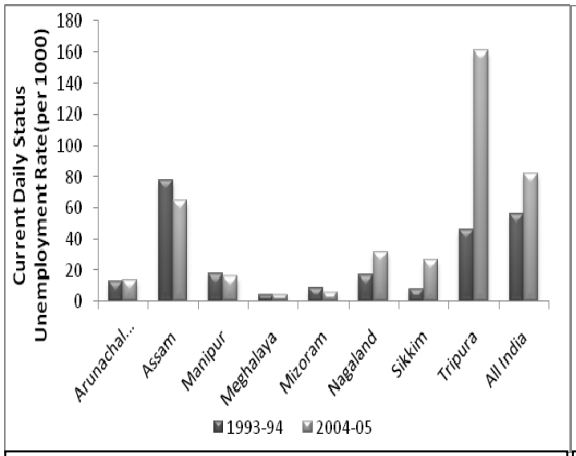

Figure 3: Current Daily Status Unemployment Rate (per 1000) in rural areas of Northeast States in 1993-94 & 2004-

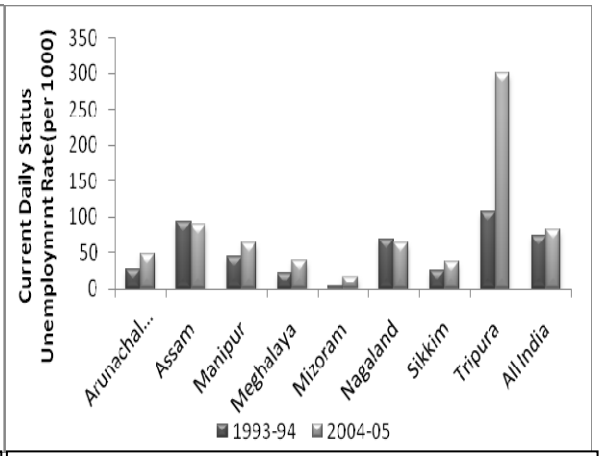

Figure 4: Current Daily Status Unemployment Rate (per 1000) in urban areas of Northeast States in 1993-94 & 2004-05

Sources: NSSO (1997): Employment and Unemployment in India, 1993-94, 50th Round, Report No. 409.
NSSO (2006): Employment and Unemployment Situation in India, 2004-05, 61st Round, Report No. 515.

Limited job opportunities available in the private sector, negligible investment by large business houses and industrial groups in the State and the high-level of qualifications required for the limited jobs on offer means that the rural youth would almost always be at a disadvantageous position. The need of the hour is to create jobs that would address the unemployment woes of the educated unemployed in the rural areas where unemployment has assumed alarming proportions and is the root cause of social evils. Realizing the urgent need to generate employment and hasten the rate of industrialization, Assam Chief Minister Tarun Gogoi announced the new State Industrial policy on the 15th of July, 2009 in which there is provision of giving special incentives for projects or industries generating a minimum of 1000 regular employment.

Results

The procedure of disaggregating district level macro data on unemployment into cells on the basis of all combinations of sex, literacy status, caste, residence background and district of residence described in the methodology section, we expect 2x2x3x2x76 (number of districts) = 1824 cells. But in some districts there is no population belonging to certain combinations of compositional and contextual background. Consequently, we were left with 1515 cells and these are the cases for application of multilevel regression model described in the preceding section. Corresponding to each cell proportion of unemployed persons is computed (WHAT) and taken as response variable.

Parameter estimates of multilevel regression outline in the methodology section are shown in Table 1 with corresponding standard errors and odds ratios. The estimated values of co-efficient of multilevel logistic regression convey useful findings as regards the compositional and contextual factors of unemployment in Northeast India. One striking result of contextual nature is that unemployment associates positively with percentage of population below the poverty line at the state level and this relation is significant statistically at $p<0.05$. The more is the population engaged at the state level, the proportion of unemployed seeking work tends to be lower, though the association is statistically not pronounced to be significant. The odds ratios of being unemployed among urban residents, literates and females are respectively 4.8, 5.2 and 5.3 per cent higher relative to their rural residents, illiterates and male counterparts and all these differentials are statistically significant at 5 per cent level of significance. Compared with the general caste category, those who are of scheduled castes (SC) and scheduled tribes (ST) have higher odds of unemployed but only the relation of SC is significant at $p<0.05$. We have noted from the present analysis that in- between Northeast India the likelihood of being unemployed varies significantly but within states among the districts the variation is not significant. However, unemployment rates

of type of people characterized by the different cells also vary significantly at 5 per cent level. Sixty-nine per cent of the variation in unemployment rate in Northeast India is due to a significant differential between the eight states in the region and variations among districts within states contribute just 6.2 per cent of the total variation.

Table 1: Parameter estimates of multilevel logit model

Fixed effects	B	t-value
Intercept	-0.075	-2.44
Log_e(nsdp)	0.005	0.34
Percentage of agricultural workers	-0.002	-0.67
Percentage below poverty line	0.005	2.50
District literacy rate	0.001	1.03
Residence background		
Rural(R)		
Urban	0.047	15.67
Educational status		
Illiterate(R)		
Literate	0.051	17.00
Sex		
Male(R)		
Female	0.052	17.33
Caste		
General & others(R)		
SC	0.018	4.5
ST	0.001	0.25
Random effect		
Between states	0.011	3.27
Between districts	0.001	1.49
Between cells	0.004	4.10

The normal probability plot of standardized level-three (cell) residuals shown in Figure 5 confirmed the validity of normality assumption of error at cell level. We have also validated normality assumptions at state and district level error components in the multilevel logistic regression. The validation assured the fitness of good in the proposed model for this study.

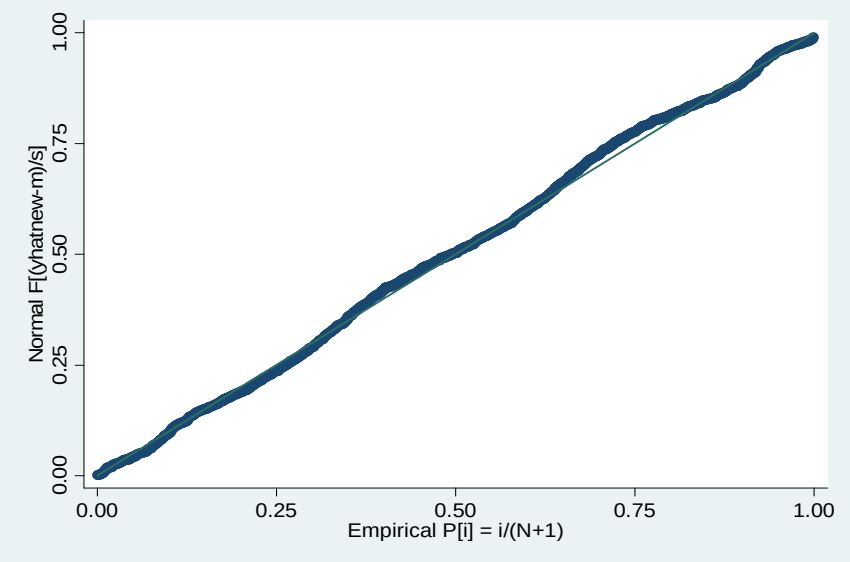

Figure 5: Normal probability plot of standardized level-three (cell) residuals

We have obtained predicted probabilities (p_{ijk}) of being unemployed for the types of people characterized by different combinations of residence background, sex, literacy status and caste and those are shown as bar chart in Figure 6.

It is noted that by caste unemployment is higher among the SCs followed by the STs, while people belonging to general and other castes fair better than them. In terms of gender differential unemployment in Northeast India, it is also observed that educated females who reside in urban areas have the highest employment rates ranging from 19.7 to 17.3 per cent, while uneducated males who live in rural areas are the least unemployed with predicted probabilities of being unemployed varying from 5.4 to 2.4 per cent depending on their castes.

To make out proper interpretation of these findings, we have to realize that more educated people normally live in urban localities and the type of limited employment opportunities available there are generally in public sector as industries and corporate own firms nearly non-exist in the region. Uneducated males irrespective of caste in rural areas are categorized as the types of people facing least unemployment, the reason being that uneducated males would look for mostly agriculture related works and there are opportunities for them. It may be that uneducated people in rural areas irrespective of caste have no choice but to take up any type of work which is available for their survival.

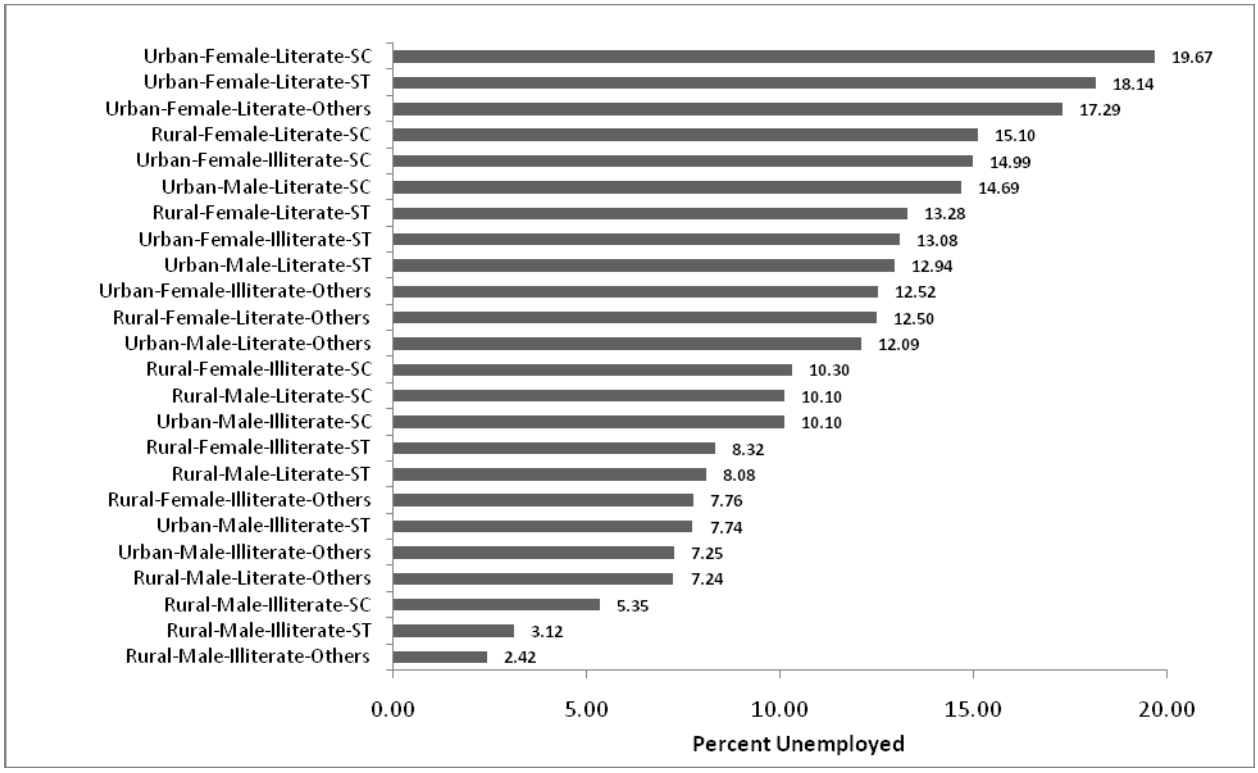

Figure 6: Predicted probabilities (p_{ijk}) of unemployment of types of people characterized by different combinations of compositional and contextual backgrounds.

References

Arbia G. (1989). *Spatial data configuration in statistical analysis of regional and economic and related problems*. Dordrecht: Kluman.

Blalock H. M. (1979). Measurement and conceptualization problems: The major obstacle to integrating theory and research. *American Sociological Review*, 44, 881-894.

Blalock H. M. (1985). Cross level analysis in the collection and analysis of community data. In J. B. Casterlin (Ed.). *ISI: World Fertility Survey*.

Blalock H. M. (1964). *Causal inferences in non-experimental research*. Chapel Hill NC: University of North California Press.

Chambers, R. L. & Skinner, C. J. (2003). *Analysis of survey data*. New York: John Wiley.

Doggan, Mattie & Rokkan S. (1969). *Quantitative ecological analysis in social sciences*. Cambridge, Mass: MIT Press.

Mattei, D. & Daniel, D. (1988). France in ten slices: An analysis of aggregate data. *Electoral Studies*, 7(3), 251-267.

Erson, S. & Lane, J. (1983). The ecological approach versus the survey approach. *European Political Data Newsletter*, 11-24.

Fotheringham, A. S. & Wong, D. W. S. (1991). A modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning*, A, 23, 1025-1044.

Firebaugh, G. (1978). A rule for inferring individual-level relationships from aggregate data. A*merican Sociological Review*, 43, 557–72.

Gekhle, C. E. & Biehl, K. (1934). Certain effects of grouping upon the size of correlation coefficient in census tract materials. *Journal of American Statistical Association*, 29 (Supplement), 169-170.

Goldstein, H. (2003). *Multilevel statistical models.* New York: John Wiley.

Garrett Thomas A. (2002). *Aggregated vs. disaggregated data in regression analysis: Implications for inference.* Federal Reserve Bank of St. Louis.

Goodman (1963). Ecological regression and behaviour of individuals. *American Sociological Review,* 18, 663-664.

Hannan, M. T. & Burstein, L. (1974). Estimation and grouped observation. *American Sociological Review,* 39, 374-392.

Hammond, J. (1973). Two sources of error in ecological correlations. *American Sociological Review*, 38, 764–77.

Krishnamurty, J, (1988). Unemployment in India: The broad magnitudes and characteristics. in T. N. Srinivasan & P. K. Bardhan (eds.), *Rural poverty in South Asia.* Delhi: Oxford University Press.

Kulkarni, S. & Kumar Santosh, V. (1989). Regional dimensions of unemployment in India: An analysis of 1981 census data. In K. Srinivasan & K. B. Pathak (eds.), *Dynamics of population and family welfare.* Bombay: Himalaya Publishing House.

Litchman, A. J. (1974). Correlation, regression and ecological fallacy: A critique. *Journal of Interdisciplinary History*, 4, 417-433.

Menzel Herbert (1950). Comment on ecological correlations and the behaviour of individuals. *American Sociological Review*, 15, 674

Openshaw, S. & Taylor, P. J. (1979). A million or so correlation coefficient: Three experiments on modifiable areal unit problem. In N. Wrigley (ed.), *Statistical Applications in the Spatial Sciences*. London: Pion. 127-144.

Rothman, K. J. & Greenland, S. (1998). *Modern epidemiology,* Philadelphia: Lippincott-Raven.

Roux Ana V. Diez (2004). The Study of group-level factors in epidemiology: Rethinking variables, study designs, and analytical approaches. *Epidemiologic Reviews*, 26, 104-111.

Robinson William, S. (1950). Ecological correlations and the behaviour of individuals. *American Sociological Review*, 15, 351-357.

Smith K.W. (1977). Another look at the clustering perspective on aggregate problems. *Sociological Methods and Research*, 5: 289-316.

Steel, D. & Holt, D. (1996). Analyzing and adjusting aggregation effects: The ecological fallacy revisited. *International Statistical Review,* 64, 39-60.

Yule, U. & Kendall, M. S. (1950). *An introduction to the theory of statistics*. London: Charles Griffin.