Demography India
A Journal of Indian Association of Study of Population
Journal Homepage: https://demographyindia.iasp.ac.in/

Indian Age Data Quality in the Light of Socio-Economic and Developmental Factors: A Multivariate Study

Jayanta Datta[1]*

Abstract

In this paper, we examined possible determinants of age data quality using census 2001 and 2011 data for major states, employing multiple linear regression. Data quality measured by the total modified Whipple's index was taken as the dependent variable, level of birth registration, average household size, infant mortality rate, literacy rate, percent BPL, percent elderly (more than 60 years above) population, percent Hindu, Muslim, schedule caste, schedule tribe and percent Urban were considered as independent variables. We removed multi-collinearity followed by stepwise backward regression and finally fitted the significant explanatory variables and analysed data using commonality analysis. Literacy rate and percent urban population for the census-2001, and literacy rate, percent BPL, percent urban, percent schedule tribe population for census-2011 were significant determinants of data quality. We observed literacy explaining 75% in the census-2001 and 50% in census-2011, which was a predominant factor in determining the data quality. There were no significant joint effects explaining the variation in the quality of data for census-2001, whereas a few commonality factors played some role, particularly for the census year 2011.

Keywords

Census, Commonality analysis, India, Stepwise backward regression.

* *Corresponding Author*
[1] *Statistical Officer and Assistant Registrar (i/c), Tripura University, Agartala, India. Email-Id: jayantatuso@gmail.com*

Introduction

The study of age reporting errors in population censuses demands systematic attention, as accurate age distribution is fundamental to demographic analysis. In developing nations like India, abnormalities such as digit preference and avoidance frequently misstate this distribution. In 2026, as India transitions toward digital census frameworks and a modernised data ecosystem, establishing a rigorous multivariate baseline from past census cohorts is critical for tuning current population projection models and actuarial life tables. While previous Indian studies have attempted to link errors in age data to various socio-economic, cultural, and developmental factors, they have been largely speculative or have focused narrowly on literacy alone. This paper addresses this crucial research gap by using multiple linear regression and commonality analysis to assess how the shifting dynamics of these multivariate determinants affected age data quality across the 2001 and 2011 Indian Censuses.

Age structures constitute the highest priority variable in demography, census age distributions are typically investigated more intensively than other metrics, with prominent guidelines advising the adjustment of raw age data before its empirical use (Kerr, 2003; Shryock et al., 1976; Srinivasan, 1998). Despite its critical importance, age tracking in developing nations frequently suffers from systematic distortions like digit preference and avoidance—particularly age heaping on terminal digits ending in "0" or "5"—which stem from the fact that conscious chronological awareness is heavily behavioral, normative, and culturally controlled (Ewbank, 1981; Mukherjee & Mukhopadhyay, 1988; Nagi et al., 1973). Early observations by Saxena et al. (1986) speculatively linked these inaccuracies to subjective factors like health, work status, old-age prestige, and marital milestones. To move beyond mere speculation, recent literature has successfully connected data improvements to measurable socio-economic shifts; for instance, Agrawal and Khanduja (2015) identified advancements in institutional literacy as a primary driver for reducing absolute age ignorance and field-level calculation errors (Ambkanavar & Visaria, 1975), while Ansary and Arif (2018) verified that simultaneously upgrading literacy, urbanization, and civil birth registration effectively secures reporting accuracy. Nonetheless, a lingering operational hurdle remains: the structural reliance on a single household head to act as a proxy informant (Jain, 1980). In communities where precise age tracking holds low social utility, this proxy method heavily compounds rounding errors and deliberate misstatements for other household members (Ewbank, 1981), creating an intricately tangled web of data errors that single-variable assessments cannot fully unravel.

Structural Determinants of Age Data Quality

When viewed structurally, previous demographic research suggests that age reporting errors are not random but are deeply rooted in a complex matrix of socio-economic, developmental, and cultural determinants.

The first major thematic pillar identified in the literature encompasses socio-demographic and institutional development. A population's

educational level and literacy (Agrawal & Khanduja, 2015; Ambkanavar & Visaria, 1975; Ansary & Arif, 2018; Byerlee & Terera, 1981; Dechter & Preston, 1991; Fayehun et al., 2020; Jain, 1980; Mukherjee & Mukhopadhyay, 1988; Nagi et al., 1973; Pal et al., 2015; Saxena et al., 1986; Singh, 2017; Tekpli, 2010) have long been established as dominant factors in shaping cognitive capacity, reducing basic ignorance or neglect regarding exact age returns. This institutional awareness is heavily reinforced by the completeness of civil infrastructure, where the level of birth registration (Ansary & Arif, 2018; Mukherjee & Mukhopadhyay, 1988) serves as a critical structural predictor for securing accurate, legally verified age tracking over time. Furthermore, these tracking systems interact closely with geographic and household environments; exposure to the percentage of urban population or place of residence (Ansary & Arif, 2018, 2018; Borkotoky & Unisa, 2014; Fayehun et al., 2020; Mukhopadhyay & Majumdar, 2009; Nagi et al., 1973; Singh, 2017; Yazdanparast et al., 2012) significantly alters spatial data precision, while immediate structural configurations, such as mean household size and the number of children (Byerlee & Terera, 1981; Mukhopadhyay & Majumdar, 2009; Saxena et al., 1986) directly compound the logistical complexity of managing individual household data returns. The second dimension involves social stratification and cultural frameworks, which frequently govern individual reporting behaviors and cognitive biases. Empirical studies indicate that age data quality often varies systematically across societal lines, notably driven by an individual's caste (Mukherjee & Mukhopadhyay, 1988; Mukhopadhyay & Majumdar, 2009) and religion (Mukherjee & Mukhopadhyay, 1988). These societal structures manifest as culturally controlled phenomena, giving rise to widespread cultural preference or avoidance (Mukherjee & Mukhopadhyay, 1988; Nagi et al., 1973; Scott & Sabagh, 1970). These cultural tendencies are further cross-cut by basic demographic attributes such as gender (Byerlee & Terera, 1981; Fayehun et al., 2020) and the biological age (Byerlee & Terera, 1981; Dechter & Preston, 1991; Saxena et al., 1986) of the respondent, which introduce distinct structural distortions like old-age prestige or middle-age underestimation.

Finally, the literature highlights economic indicators and fieldwork logistics as critical constraints on data precision. Sectoral economic transitions, measured through the proportion of the population economically active or work status (Nagi et al., 1973; Saxena et al., 1986) and the percentage of the non-agricultural labour force (Mukhopadhyay & Majumdar, 2009; Nagi et al., 1973), typically correlate with heightened age awareness due to formal employment records. Conversely, data reliability is frequently undermined by immediate behavioral and logistical barriers during field enumeration. These include high rates of population migration (Byerlee & Terera, 1981) that disrupt stable tracking, and milestones like marriage (Saxena et al., 1986) used as proxy reference points. These field dynamics are heavily exacerbated by regional variations in the women's response rate (Mukhopadhyay & Majumdar, 2009), natural cognitive limitations like a recall lapse (Yusuf et al., 2014), and a widespread systemic reliance on proxy information by third parties (Byerlee & Terera, 1981; Caldwell & Igun,

1971) who may lack precise knowledge regarding other household members.

However, there is indeed a lack of extensive studies on this field for the Indian census data for the reference study period. Almost all previous studies were focused on relating age data quality with education only. Besides education, other socio-economic, cultural and developmental factors may also influence the patterns of age reporting. In this paper, we attempted to fulfil this research gap to some extent by incorporating the maximum possible number of determinants of age data quality, subject to data availability by states/UTs. We tried to investigate the data quality by measuring heaping in age data vis-à-vis some selected socio-economic, developmental and cultural factors influencing the reporting of age in Indian census data using multiple linear regression and commonality analysis in the light of multicollinearity.

The objective of the study is to examine the impact of different factors on the quality of age reporting. This paper is confined to some of the available socio-economic and developmental factors in relation to the literature review, on which data pertaining to the last two censuses of India by major states are available. A multivariate study of age data quality in the light of available socio-economic variables (independent variables) for census-2001 and census-2011 is considered in this study.

Initially, a number of variables were included in the study, but due to collinearity among the variables and guided by the literature review, a few variables were excluded in the final analysis. All the independent variables considered were tested for collinearity using VIF. We have further analyzed the data employing backward regression and commonality analysis.

Data Sources

1. Single-year age returns for both genders classified by place of residence for the census year-2001, published by RGI, Govt. of India.
2. Single-year age returns for both genders classified by place of residence for the census year-2011, published by RGI, Govt. of India.
3. Sample Registration System Bulletin (SRS bulletin) Vol-36, No-2, October 2002.
4. Sample Registration System (SRS bulletin) Vol-47, No-2, October 2012.
5. Number and percentage of population below poverty line (2004-05, based on Tendulkar method on Mixed Reference Period) published by RBI.
6. Number and percentage of population below poverty line (2011-12, based on Tendulkar method on Mixed Reference Period) published by RBI.
7. Vital Statistics of India based on the civil registration system-2011, published by Office of the Registrar General, India, Ministry of Home Affairs (Statement 15: Level of registration of births, of States/UTs, column-10, page no 29)
8. Vital Statistics of India based on the civil registration system-2010, published by Office of the Registrar General, India, Ministry of Home Affairs (Statement 13: Level of

registration of births, of States/UTs, column-01, page no 28)

9. Chapter-3, Literacy and education (Table 3.3 State-wise Literacy Rates in Last 3 Decades) published by the Office of Registrar General, India.
10. Percent elderly 2016, by Central Statistics Office, Ministry of Statistics and Programme Implementation, Government of India, Table 1.6. Size of elderly population (aged 60+) and their share in total population in States and Union Territories, p-27.
11. Table C-01: Population by religious community, published by RGI, Govt. of India–2001.
12. Table C-01: Population by religious community, published by RGI, Govt. of India–2011.
13. Population Figures of India and States, 2001.Table (PCA-TOT) Primary Census Abstract Total, published by RGI, Govt. of India.
14. Population Figures of India and States, 2011.Table (PCA-TOT) Primary Census Abstract Total, published by RGI, Govt. of India.

Methodology

The study is primarily focused on finding out significant determinants and their joint effects influencing the errors in age reporting. Andhra Pradesh, Assam, Bihar, Chandigarh, Gujarat, Haryana, Himachal, Jammu & Kashmir, Jharkhand, Karnataka, Kerala, Madhya Pradesh, Maharashtra, Odisha, Punjab, Rajasthan, Tamil Nadu, Uttar Pradesh, Uttaranchal[2], and West Bengal were the big states[3] that have been considered in this study. There are different indices in demography to measure the quality of data in age reporting. In this analysis, errors in age reporting are calculated using the total modified Whipple's index (W_{tot}), a summary index computed using all the digit specific modified Whipple's Index.

Computation of Digit Specific Modified Whipple's Index: -

Age heaping is calculated for all the digits (0-9). The degree of preference or avoidance for digits (0-9) can be computed as follows:

$W_0 = 5(P_{30} + P_{40}+P_{50}+P_{60})/({}_5P_{28} + {}_5P_{38}+{}_5P_{48}+{}_5P_{58})$;
$W_1 = 5(P_{31}+P_{41}+P_{51}+P_{61})/({}_5P_{29} + 5P_{39}+{}_5P_{49}+{}_5P_{59})$;
and so on...;
$W_9 = 5(P_{29}+P_{39}+P_{49}+P_{59})/({}_5P_{27} + {}_5P_{37}+{}_5P_{47}+{}_5P_{57})$;

Where P_x is the population of completed age x and ${}_5P_X$ the population of the age range (x, x + 4).

In the absence of digit preference or avoidance, these "digit-specific modified Whipple's index" values are equal to 1. An index above or below 1 signifies, respectively, a choice for or avoidance of the digit in question.

Total modified Whipple's index (W_{tot} an overall summary index) is given by:

$$W_{tot} = \sum_{i=0}^{9} (|W_i - 1|)$$

[2]*Due to the non-availability of data on the level of birth registration for the year 2001 of Uttaranchal, it was excluded.*
[3]*Because of the availability of some of the socio-economic variables for big states only, we included these selected big states in our analysis.*

Where $W_{i's}$ are digit-specific modified Whipple's index (Spoorenberg & Dutreuilh, 2007).

Data frames 'D_01' and 'D_02' (using RStudio Version 1.3.1073) were prepared with state names, total modified Whipple's index (W_{tot}) values, and different socio-economic variables for the census year 2001 and census year 2011, respectively. We fitted multiple linear regression models taking total modified Whipple's index (W_{tot}) values as dependent variable and per cent population below poverty line[4] (*BPL*), per cent level of birth registration (*Breg*), per cent elderly (more than 60 years old) population with respect to state populations (*Elderly*), average household size (*HHSize*), infant mortality rates (*IMR*), literacy rates (*Literacy*), per cent Hindu with respect to total state population (*Hindu*), per cent Muslim with respect to total state population (*Muslim*), per cent schedule caste with respect to total state population (*SC*), per cent schedule tribe with respect to total state population (*ST*), per cent urban population with respect to total state population as explanatory variables (*Urban*).

Multiple Regression Models: -

Model 1 (Census-2001)

$$Wtot_i = \beta_0 + \beta_1 BPL_i + \beta_2 Breg_i + \beta_3 Elderly_i + \beta_4 HHSize_i + \beta_5 IMR_i + \beta_6 Literacy_i + \beta_7 Hindu_i + \beta_8 Muslim_i + \beta_9 SC_i + \beta_{10} ST_i + \beta_{11} Urban_i + \varepsilon_i$$

Model 2 (Census-2011)

$$Wtot_i = \beta_0 + \beta_1 BPL_i + \beta_2 Breg_i + \beta_3 Elderly_i + \beta_4 HHSize_i + \beta_5 IMR_i + \beta_6 Literacy_i + \beta_7 Hindu_i + \beta_8 Muslim_i + \beta_9 SC_i + \beta_{10} ST_i + \beta_{11} Urban_i + \varepsilon_i$$

Where, 'i' represents respective states/UTs; β_0 = Intercept of the equation; β_1 to β_{11} = Coefficients; ε_i = Error term.

Data on independent variables for model-1 and model-2 correspond to the respective census years 2001 and 2011.

Result and Discussion

Many independent variables were included in the regression models at the initial stage, and multicollinearity diagnostics were performed. We performed Variance Inflation Factor (VIF) test to detect the multicollinearity problem among the explanatory variables. We removed several explanatory variables with high Variance Inflation Factor (VIF) and high bi-variate correlation among themselves. The VIF of all the included independent variables considered is less than 10 (Table 1).

The tolerance value statistics are greater than the common threshold of 0.10 for all these explanatory variables. This situation confirmed the absence of multicollinearity among the response variables. All the analyses were done using software RStudio Version 1.3.1073.

[4]*Percentage of population below poverty line computed as per Tendulkar method on Mixed Reference Period (MRP) 2004-05 is considered for the data frame D_01 and (MPR) 2011-12 is considered for the data frame D_02.*

Table 1 Codes and output summary for VIF and tolerance of the fitted model-1 and model-2

```
> # Performing VIF test for model1
> model1 <- lm(Wtot~. - State ,data=
D_01)
>ols_vif_tol(model1)
```

	Variables	Tolerance	VIF
1	BPL	0.2737622	3.652806
2	Breg	0.1390758	7.190323
3	Elderly	0.2441645	4.095600
4	HHSize	0.3037619	3.292052
5	Hindu	0.2513651	3.978277
6	IMR	0.1797651	5.562815
7	Literacy	0.1120747	8.922619
8	Muslim	0.1756247	5.693960
9	SC	0.2730215	3.662715
10	ST	0.3280213	3.048583
11	Urban	0.2323250	4.304316

```
> # Performing VIF test for model2
> model2 <- lm(Wtot~. - State ,data=
D_02)
>ols_vif_tol(model2)
```

	Variables	Tolerance	VIF
1	BPL	0.3122970	3.202080
2	Breg	0.2601661	3.843699
3	Elderly	0.1991887	5.020364
4	HHSize	0.2899454	3.448925
5	Hindu	0.2470408	4.047914
6	IMR	0.1921875	5.203252
7	Literacy	0.1802801	5.546923
8	Muslim	0.1731839	5.774210
9	SC	0.3290465	3.039084
10	ST	0.3732843	2.678924
11	Urban	0.2366922	4.224896

Source: Author's computation using census data.

We first performed multivariate regression analysis, eliminating a few unimportant explanatory variables using the backward regression process based on p-value (up to a p-value of 0.1). Secondly, we performed a commonality analysis on the remaining significant variables (Ray-Mukherjee et al., 2014) and executed stepwise backward regression. For the census-2001, out of 11 explanatory variables, nine variables were eliminated by stepwise backward regression (Table 2).

Table 2 Codes and output summary of stepwise backward regression (Model 1)

```
> # stepwise backward regression CENSUS 2001
>ols_step_backward_p(model1, prem = 0.1, progress = FALSE,
+                     details = FALSE,)->p
> p
                         Elimination Summary
-------------------------------------------------------------------------
        Variable                    Adj.
Step    Removed     R-Square    R-Square     C(p)        AIC        RMSE
-------------------------------------------------------------------------
   1    Muslim        0.8923      0.7576    10.0521    39.5388     0.5613
   2    SC             0.887      0.7739     8.4001    38.4538     0.5421
   3    Breg          0.8817      0.7871     6.7415    37.3108     0.5260
   4    Elderly       0.8726      0.7915     5.3399    36.7255     0.5205
   5    IMR            0.853      0.7795     4.6211    37.4408     0.5353
   6    ST            0.8437      0.7836     3.2329    36.6121     0.5304
   7    Hindu         0.8255      0.7757     2.4215    36.7000     0.5400
   8    HHSize        0.8152      0.7782     1.0980    35.7932     0.5369
   9    BPL           0.7872      0.7607     0.9270    36.4683     0.5577
-------------------------------------------------------------------------
```

Source: Author's computation using census data.

Stepwise backward regression in the case of census-2011 data removed seven variables (Table 3). Eliminations were done based on a p-value of 0.1.

Table 3 Codes and output summary of stepwise backward regression (Model 2)

```
> # stepwise backward regression CENSUS 2011
>ols_step_backward_p(model2, prem = 0.1, progress = FALSE,
+                     details = FALSE,)->q
> q
                          Elimination Summary
-------------------------------------------------------------------------
        Variable                    Adj.
Step    Removed     R-Square    R-Square     C(p)        AIC        RMSE
-------------------------------------------------------------------------
   1    Muslim        0.9498      0.8941    10.0084    12.7486     0.2723
   2    Elderly       0.9496      0.9043     8.0400    10.8271     0.2588
   3    SC            0.9487      0.9113     6.1911     9.1996     0.2491
   4    IMR           0.9456      0.9138     4.6833     8.3668     0.2455
   5    Hindu         0.9418      0.9149     3.2890     7.7153     0.2440
   6    Breg           0.935      0.9118     2.3763     7.9291     0.2485
   7    HHSize        0.9252      0.9052     1.9420     8.7398     0.2575
-------------------------------------------------------------------------
```

Source: Author's computation using census data.

After removing all the insignificant variables in predicting age data quality (W_{tot}) for Census-2001 by backward elimination, we regress total modified Whipple's index (W_{tot}) on the significant explanatory variables '*Literacy*' and '*Urban*'. The summary output is given in Table 4.

Table 0 Codes and output summary of final model after elimination (Census 2001)

```
> # FINAL MODEL AFTER REMOVING VARIABLES
> final_1 <- lm(Wtot ~ Literacy+Urban,data= D_01)
> summary(final_1)
Call:
lm(formula = Wtot ~ Literacy + Urban, data = D_01)
Residuals:
     Min       1Q   Median       3Q      Max
-1.34738 -0.27370  0.00477  0.29393  0.91981

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.015413   0.874234  13.744 2.81e-10 ***
Literacy    -0.108650   0.014489  -7.499 1.27e-06 ***
Urban        0.020032   0.008669   2.311   0.0345 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5577 on 16 degrees of freedom
Multiple R-squared:  0.7872,    Adjusted R-squared:  0.7607
F-statistic:  29.6 on 2 and 16 DF,  p-value: 4.198e-06
```

Source: Author's computation using census data.

In predicting age data quality for the census year 2001, independent variables '*Literacy*' and '*Urban*' were significant explanatory variables at 1 per cent and 5 per cent level respectively.

For the census-2011, after eliminating insignificant variables in predicting age data quality (W_{tot}) by backward regression, we fitted a multiple linear regression equation of 'age data quality' on dependent variables '*Literacy*', '*BPL*', '*Urban*' and '*ST*'. The summary output is given in Table 5.

Table 4 Codes and output summary of final model after elimination (Census 2011)

```
> # FINAL MODEL AFTER REMOVING VARIABLES
> final_2 <- lm(Wtot ~ Literacy+BPL+Urban+ST,data = D_02)
> summary(final_2)

Call:
lm(formula = Wtot ~ Literacy + BPL + Urban + ST, data = D_02)

Residuals:
     Min       1Q   Median       3Q      Max
-0.35393 -0.15329 -0.05003  0.13509  0.42440

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.541450   0.775344  12.306 3.06e-09 ***
Literacy    -0.102081   0.010197 -10.011 4.92e-08 ***
BPL          0.028925   0.007515   3.849  0.00158 **
Urban        0.013837   0.003924   3.527  0.00305 **
ST          -0.024096   0.008871  -2.716  0.01593 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2575 on 15 degrees of freedom
Multiple R-squared:  0.9252,    Adjusted R-squared:  0.9052
F-statistic: 46.36 on 4 and 15 DF,  p-value: 2.854e-08
```

Source: Author's computation using census data.

In predicting age data quality for the census year 2011, all the independent variables '*Literacy*', '*BPL*', and '*Urban*' were significant at 1 per cent level, and the explanatory variable '*ST*' was significant at 5 per cent level.

We performed commonality analysis (Table 6), taking W_{tot} as the dependent variable and the significant variables (*Literacy, Urban*) as independent variables for the census year 2001.

For the census year 2001, the total variation in the age data quality explained by '*Literacy*' uniquely is about 75 per cent. The explanatory variable '*Urban*' uniquely explains only 7 per cent variation of the data quality. All these explanatory variables explained 82 per cent of the variation in the age data quality uniquely. Also, independent variables '*Literacy* and *Urban*' collectively explained only three per cent variation in the data quality, but the direction was negative.

Table 5 Codes and output summary of commonality analysis (Census-2001)

```
> #Commonality analysis
>library(MBESS)
> library(yhat)
> #CENSUS 2001
> CC_01<-commonalityCoefficients(D_01, "Wtot", list("Literacy", "Urban"),"F")
>print(CC_01)
$CC
                              Coefficient     % Total
Unique to Literacy                 0.7477       94.98
Unique to Urban                    0.0710        9.02
Common to Literacy, and Urban     -0.0315       -4.00
Total                              0.7872      100.00

$CCTotalbyVar
Unique  Common  Total
Literacy 0.7477 -0.0314 0.7163
Urban    0.0710 -0.0315 0.0395
```

Source: Author's computation using census data

Table 6 Codes and output summary of commonality analysis (Census-2011)

```
> #CENSUS 2011
> CC_11<-commonalityCoefficients(D_02, "Wtot", list("Literacy","BPL"
,"Urban","ST"),"F")
>print(CC_11)
$CC
                                       Coefficient     % Total
Unique to Literacy                          0.5000       54.05
Unique to BPL                               0.0739        7.99
Unique to Urban                             0.0621        6.71
Unique to ST                                0.0368        3.98
Common to Literacy, and BPL                 0.2639       28.53
Common to Literacy, and Urban              -0.0424       -4.58
Common to BPL, and Urban                    0.0169        1.83
Common to Literacy, and ST                 -0.0082       -0.89
Common to BPL, and ST                      -0.0301       -3.26
Common to Urban, and ST                     0.0239        2.58
Common to Literacy, BPL, and Urban          0.0052        0.57
Common to Literacy, BPL, and ST             0.0179        1.94
Common to Literacy, Urban, and ST          -0.0334       -3.61
Common to BPL, Urban, and ST               -0.0134       -1.45
Common to Literacy, BPL, Urban, and ST      0.0521        5.63
Total                                       0.9252      100.00

$CCTotalbyVar
         Unique Common  Total
Literacy 0.5000 0.2552 0.7552
BPL      0.0739 0.3125 0.3864
Urban    0.0621 0.0089 0.0710
ST       0.0368 0.0087 0.0455
```

Source: Author's computation using census data.

We performed commonality analysis taking W_{tot} as the dependent variable and the significant variables (*Literacy*, *BPL*, *Urban*, and *ST*) as independent variables for the census year 2011 (Table 7).

For the census year 2011, the total variation in the age data quality explained by 'Literacy' uniquely is about 50 per cent. The explanatory variable 'BPL' uniquely explained 7 per cent, 'Urban' explained 6 per cent, and 'ST' explained about 4 per cent of the variation in the value of total modified Whipple's index (W_{tot}). All these explanatory variables explained about 67 per cent of the variation in the age data quality uniquely. Independent variables 'BPL' in common explained about 31 per cent of the total variation in W_{tot} values. Also, the independent variable 'Literacy' in common explained about 26 per cent of the total variation in W_{tot} values. The joint effect of 'Literacy and BPL' was the highest, explaining 26 per cent of the variation in the age data quality. The joint effect of 'Literacy, BPL, Urban and ST' was the second-highest, explaining about 5 per cent variation in the value of the total modified Whipple's index (W_{tot}). All other joint effects were less than 5 per cent.

The empirical shifts observed between the 2001 and 2011 models reveal a fundamental structural transition in the drivers of age data quality in India. In 2001, the overwhelming dominance of the literacy rate, which uniquely accounted for 75% of the explained variance with virtually no significant joint effects, confirms that baseline educational exposure was historically the singular barrier to accurate age reporting. By 2011, however, the contraction of literacy's unique explanatory power to 50%, alongside the emergence of per cent BPL, per cent urban, and per cent Scheduled Tribe (ST) population as significant, unique, and joint determinants, highlights a more complex socio-demographic landscape. This evolutionary matrix indicates that as macro-level literacy expanded across India, it ceased to be the lone differentiator of data precision. Instead, localised pockets of economic deprivation and social marginalisation have crystallised into distinct structural barriers that independently perpetuate digit preference and age heaping. Consequently, these findings demonstrate that untangling modern age reporting errors requires a multivariate approach, as traditional single-variable frameworks can no longer capture how spatial and economic inequalities jointly distort national demographic counts.

Policy Recommendation

In light of the full transition toward a modernized, digital census framework, the Registrar General and Census Commissioner of India (RGI) may adopt a strategically targeted, socio-spatially nuanced approach to fieldwork and data validation rather than relying on generalized, literacy-centric interventions. First, the RGI may deploy intensive data-awareness campaigns and specialized surveyor training within economically marginalized (*BPL*) and Scheduled Tribe (*ST*) pockets, as the empirical evidence demonstrates that these sub-populations continue to exhibit distinct structural vulnerabilities in age-reporting accuracy. Second, utilizing the digital capabilities of the new enumeration platform, the RGI may integrate the Civil Registration

System (CRS) directly with the digital census portals to allow real-time cross-referencing and automatic electronic validation of oral doorstep responses against legally verified birth records. Finally, to eliminate the profound rounding errors associated with single-informant proxy responses in areas where chronological age tracking holds low social utility, the RGI may introduce simplified, culturally relevant electronic historical event calendars within the enumeration application to assist households in accurately reconstructing individual chronological ages.

This study provides a rigorous, multivariate evaluation of the shifting socio-economic, developmental, and cultural determinants of census age data quality in India across the 2001 and 2011 cohorts. By employing multiple linear regression paired with systematic multi-collinearity diagnostics and commonality analysis, the research moves beyond historically speculative narratives to isolate quantified structural drivers.

The empirical findings reveal a dynamic evolutionary shift: while institutional literacy stood out as the overwhelmingly dominant predictor in 2001—uniquely explaining three-quarters of the data variance—its isolated explanatory power diminished significantly by 2011. In the latter census, urbanization, economic deprivation, and social marginalization emerged as vital joint and unique determinants. This changing matrix indicates that as basic literacy expands, structural inequalities and sub-population characteristics increasingly dictate the spatial configuration of digit preference and age heaping. Ultimately, as India navigates a modernized, digital data ecosystem, establishing this empirical baseline serves an essential demographic purpose. It provides the precise statistical insights necessary to calibrate contemporary population projection models, correct actuarial life tables, and design targeted field-level interventions to guarantee the mathematical integrity of future national counts.

Limitation

While examining the various effects of socio-economic and developmental factors on age data reporting, we found a serious lag in the availability of data with respect to time and space. This limitation resulted in the selection of only a few variables, with the condition of the availability of data by state for the reference time period.

Acknowledgement

The author acknowledges the Office of the Registrar General & Census Commissioner, Government of India, for downloading data of the census-2001 and census-2011. The author declares no ethical issues. There is no conflict of interest to declare, and the study is not funded by any organization.

Declaration of Conflicting Interests

The Author declares that there is no conflict of interest.

Funding

Any organization does not fund the study.

References

Agrawal, G., & Khanduja, P. (2015). Influence of literacy on India's tendency for age misreporting: Evidence from Census 2011. *Journal of Population and Social Studies*, 23(1), 47-56

Ambkanavar, J. P., & Visaria, P. (1975). Influence of literacy and education on the quality of age returns. *Demography India*, 4(1), 11–15.

Ansary, R., & Arif, M. (2018). Quality of age statistics in India: An insight of changing course of reliability. *Journal of the Geographical Institute "Jovan Cvijic" SASA*, 68(3), 345–361. https://doi.org/10.2298/IJGI180218006A

Borkotoky, K., & Unisa, S. (2014). Indicators to examine quality of large scale survey data: An example through district level household and facility survey. *PLoS ONE*, 9(3), e90113. https://doi.org/10.1371/journal.pone.0090113

Byerlee, D., & Terera, G. (1981). Factors affecting reliability in age estimation in rural West Africa: A statistical analysis. *Population Studies*, 35(3), 455–465. https://doi.org/10.2307/2174666

Caldwell, J. C., & Igun, A. A. (1971). An explement with census-type age enumeration in Nigeria. *Population Studies*, 25(2), 287–302. https://doi.org/10.1080/00324728.1971.10405804

Dechter, A. R., & Preston, S. H. (1991). Age misreporting and its effects on adult mortality estimates in Latin America. *Population Bulletin of the United Nations*, 31–32, 1–16.

Fayehun, O., Ajayi, A. I., Onuegbu, C., & Egerson, D. (2020). Age heaping among adults in Nigeria: Evidence from the Nigeria Demographic and Health Surveys 2003–2013. *Journal of Biosocial Science*, 52(1), 132–139. https://doi.org/10.1017/S0021932019000348

Jain, S. P. (1980). Census single year age returns and informant bias. *Demography India*, 9(1–2), 286–296.

Mukherjee, B. N., & Mukhopadhyay, B. K. (1988). A study of digit preference and quality of age data in Turkish censuses. *Genus*, 44(1–2), 201–227.

Mukhopadhyay, B. K., & Majumdar, P. K. (2009). A multivariate statistical analysis of reporting error in age data of India. *Journal of Social Sciences*, 19(1), 57–61. https://doi.org/10.1080/09718923.2009.11892691

Nagi, M. H., Stockwell, E. G., & Snavley, L. M. (1973). Digit preference and avoidance in the age statistics of some recent African censuses: Some patterns and correlates. *International Statistical Review*, 41(2), 165–174. https://doi.org/10.2307/1402833

Pal, J. K., Mukhopadhyay, B. K., & Tewari, H. R. (2015). Age reporting error and effect of education: A village study. *American Journal of Social Science Research*, 1(3), 158–162.

Ray-Mukherjee, J., Nimon, K., Mukherjee, S., Morris, D. W., Slotow, R., & Hamer, M. (2014). Using commonality analysis in multiple regressions: A tool to decompose regression effects in the face of multicollinearity. *Methods in Ecology and Evolution*, 5(4), 320–328. https://doi.org/10.1111/2041-210X.12166

Saxena, P. C., Verma, K. R., & Sharma, K. A. (1986). Errors in age reporting in India, a socio-cultural and psychological explanation. *Indian Journal of Social Work*, 47(2), 127–135.

Scott, C., & Sabagh, G. (1970). The historical calendar as a method of estimating age: The experience of the Moroccan multi-purpose sample survey of 1961-1963. *Population Studies*, 24(1), 93–109. https://doi.org/10.2307/2173265

Shryock, H. S., Siegel, J. S., & Stockwell, E. G. (1976). *The methods and materials of*

demography (Condensed ed.). Academic Press.

Singh, M. (2017). *Understanding digit preferences in India using modified Whipple index: An analysis of 640 districts of India* [Unpublished manuscript]. International Institute for Population Sciences.

Spoorenberg, T., & Dutreuilh, C. (2007). Quality of age reporting: Extension and application of the modified Whipple's index. *Population*, 62(4), 729–741.

Tekpli, G. D. (2010). *Evaluation and adjustment of age sex data of population and housing census of Ghana 2000 and 2010* [Master's thesis, University of Ghana]. University of Ghana Digital Library.

Yazdanparast, A., Pourhoseingholi, M. A., & Abadi, A. (2012). Digit preference in Iranian age data. *Italian Journal of Public Health*, 9(4), 64–70. https://doi.org/10.2427/5630

Yusuf, F., Martins, J. M., & Swanson, D. A. (2014). *Methods of demographic analysis*. Springer. https://doi.org/10.1007/978-94-007-6784-3