Data Quality in Indian Demographic Surveys

Manish Singh*

Background: The most important demographic variable like age data collected from censuses and other demographic sample surveys being affected by several age reporting errors. Ignorance, inability to reckon one's age correctly, misunderstanding of the concepts are some of the important causes for the bias especially in developing countries like India, particularly in certain population groups (Registrar General of India, 2008).

Usually age data suffers from problems like the 'age under-enumeration' and 'age distortions' due to liking for certain ages e.g. digits like 0 and 5 as preferred more compared to digits like 1 or 9 in societies having low literacy rate. People while reporting age data does not realize 'age' as an important factor as age misreporting in fact, influence to a great extent all types of demographic analysis that considers 'age' as a variable (Registrar General of India, 2008).

Such a tendency of digit preference is known as heaping and is usually seen in single-year age returns of censuses and sample surveys. Respondent's unawareness about his/her 'data of birth' is said to be a major cause for incorrect reporting of their age.

Thus there is a great need to measure and adjust the age data before using it in any rigorous exercise like the 'projecting the population by age and sex into the future years." Several measures have already been developed namely "Whipple's index and 'Myers' Index (See, Ewbank, 1981; Shyrock and Siegal, 2004) and still developing to improve the quality of age data. Modified Whipple's index (W_{tot}) developed by Spoorenberg in 2007 is an outcome of such an attempt made in the recent past, and also an improvement over previous ones.

Data and Methods

The single year age-sex distribution of population of India, its states and districts for the District Level Household and Facility Surveys-3 (DLHS-3). A detailed description of the Whipple's Index and its modifications have been presented based on the studies by Spoorenberg and Dutreuilh (2007), and Spoorenberg (2009) as follows:

2.1 Whipple's Index and its modifications

Whipple's original index of digit preference popularly known as the index of concentration is calculated in two steps using the single year age data of each sex separately of the ages 23 to 62.

***International Institute for Population Sciences, Mumbai, India**

It is defined as

$$\mathbf{WI = 5\ (P_{25} + P_{30} + P_{35} + \ldots + P_{60})/(P_{23}+P_{24}+\ldots+P_{61}+P_{62})\ \ -----(1)}$$

Where P_x (x=23,……62) is the population of the completed age x

The above index is developed assuming that a "continuous and linear decrease in the number of persons of each age within the age range of 23 and 62." It implies that the above linearity assumption cannot be applied to other ages of 0-22 years and 63 and above years.

However the choice of the age limits fixed here is made very arbitrarily but found to be effective.

The calculated values of WI fall between a minimum value of 100 to a maximum value of 500. Maximum is expected only when "no returns are recorded with any digits other than the two 0 and 5." The quality of the data based on its value is highly accurate (WI< 105), fairly accurate ((105<WI< 110), approximate (110<WI<125), rough (125<WI<175), and very rough (WI > 175) (Registrar General India, 2008).

In a recent study Spoorenberg (2009) states that "the original Whipple's index does not account well for the quality of age reporting once it is improving and reaches better levels." In addition he states "the total modified Whipple's Index (WI_{tot}) offers however a simple alternative which fully accounts for the changes in the attraction/repulsion on all age-digits."

Roger et, al (1981) and Shryock and Siegel et al., (2004) proposed the following two formulae as first modification for WI as

$$\mathbf{W_0 = 10\ (P_{30} + P_{40} + P_{50} + P_{60})/(P_{23}+P_{24}+\ldots+P_{61}+P_{62})\ \ -----(2)}$$

and

$$\mathbf{W_5 = 10\ (P_{25} + P_{35} + P_{45} + P_{55})/(P_{23}+P_{24}+\ldots+P_{61}+P_{62})\ \ -----(3)}$$

From the above two WI, they obtained $\mathbf{WI_R = (W_0 + W_5)\ /2\ \ ---------(4)}$

WI_R allows one to distinguish age preferences in favor of 0 or 5, assuming a linearity over a ten-year age range, which is unrealistic (Spoorenberg, 2007,).

Noumbissi (1992) unlike WI_R proposed a modification to WI based on the linearity assumption over an age range of 5 years and introduced ten formulae allowing age heaping for all 10 digits of 0, 1,2,..,9 has given below :

$$\mathbf{W_0 = 5\ (P_{30} + P_{40} + P_{50} + P_{60})/({}_5P_{28}+{}_5P_{38}+{}_5P_{48}+{}_5P_{58})\ \ -----(5)}$$

$$\mathbf{W_5 = 5\ (P_{25} + P_{35} + P_{45} + P_{55})/({}_5P_{23}+{}_5P_{33}+{}_5P_{43}+{}_5P_{53})\ \ -----(6)}$$

$$\mathbf{W_1 = 5\ (P_{31} + P_{41} + P_{51} + P_{61})/({}_5P_{29}+{}_5P_{39}+{}_5P_{49}+{}_5P_{59})\ \ -----(7)}$$

$$\mathbf{W_2 = 5\ (P_{32} + P_{42} + P_{52} + P_{62})/({}_5P_{30}+{}_5P_{40}+{}_5P_{50}+{}_5P_{60})\ \ -----(8)}$$

$$W_3 = 5\,(P_{23} + P_{33} + P_{43} + P_{53})/({}_5P_{21}+{}_5P_{31}+{}_5P_{41}+{}_5P_{51}) \quad \text{-----(9)}$$

$$W_4 = 5\,(P_{24} + P_{34} + P_{44} + P_{54})/({}_5P_{22}+{}_5P_{32}+{}_5P_{42}+{}_5P_{52}) \quad \text{-----(10)}$$

$$W_6 = 5\,(P_{26} + P_{36} + P_{46} + P_{56})/({}_5P_{24}+{}_5P_{34}+{}_5P_{44}+{}_5P_{54}) \quad \text{-----(11)}$$

$$W_7 = 5\,(P_{27} + P_{37} + P_{47} + P_{57})/({}_5P_{25}+{}_5P_{35}+{}_5P_{45}+{}_5P_{55}) \quad \text{-----(12)}$$

$$W_8 = 5\,(P_{28} + P_{38} + P_{48} + P_{58})/({}_5P_{26}+{}_5P_{36}+{}_5P_{46}+{}_5P_{56}) \quad \text{-----(13)}$$

$$W_9 = 5\,(P_{29} + P_{39} + P_{49} + P_{59})/({}_5P_{27}+{}_5P_{37}+{}_5P_{47}+{}_5P_{57}) \quad \text{-----(14)}$$

Where ,

${}_nP_x$ denote the population of age range (x to x+n-1)

$W_i = 1$ (indicates no digit preference or avoidance)

$W_i > 1$ **or** $W_i < 1$ **(**indicates a digit preference or avoidance for digit in question)

Realizing the fact that the above 10 digit-specific modified Whipple's indexes are very cumbersome to handle when studying the spatial or temporal etc. aspects , Spoorenberg (2007) suggested the following total modified Whipple's Index (W_{tot}) as a summary measure that summarizes all age performance and avoidance effects:

$$W_{tot} = \sum_{i=0}^{9}(|W_i - 1|) \quad \text{---- (15)}$$

Here,

W_i = digit-specific modified Whipple's index for each of the ten digits (0-9) developed by Noumbissi (1992)

$W_{tot} = 0$, for no digit preference

$W_{tot} = 16$, for complete digit preference at 0 and 5 i.e $W_0 = W_5 = 5$ and $W_i = 0$

Superiority of W_{tot} has been found by Spoorenberg (2009) by applying it to various single-year age data sets of various regions and time periods in the world (compared with that of Myers' blended index and the original Whipple's index). The results based on W_{tot} produces practically more or less the same results as Myer's blended index. (Favoring W_{tot}, and dismissing the original WI, Spoorenberg (2009) in his conclusion states that " if one wants to assess with more precision the quality of age reporting and its change through time the original Whipple's index is not a completely fair and reliable measure.")

Thus W_{tot} may be used as a fitting alternative in all future analysis of the age sex data. W_{tot} is easy to calculate and easy to understand.

Thus in the present paper using W_i and W_{tot} it has been tried to see the extent of digit preference in District Level Household and Facility Surveys-3 (DLHS-3).

Conclusions:

An analysis of 601 districts and 27 states of India in District Level Household and Facility Surveys-3 (DLHS-3) is done by using GIS software and is shown by means of four maps of India. This map shows the value of W_{tot} by sex in districts and states of India. Map1 shows that quality of data is good (i.e. the value of W_{tot} is less than or equal to 2.0) in Tamil Nadu and poor (i.e. the value of W_{tot} is above 6.0) in Uttar Pradesh, Gujarat, Chhattisgarh and Jharkhand states of India in case of males. Similarly in case of females Map2 shows data quality is good in Sikkim only and poor in Rajasthan, Madhya Pradesh, Bihar and Jharkhand.

Map3 shows an analysis of 601 districts of India and it shows that quality of data is good (i.e. the value of W_{tot} is less than or equal to 2.0) in Thirunelveli, Kanniyakumari, Theni, Madurai, Kollam, Alappuzha, Kottayam and Thrissur districts and poor (i.e. the value of W_{tot} is above 6.0) in most of the districts of India. Similarly in case of females Map4 shows data quality is good in most of the districts of Tamil Nadu and Kerala. In general it is observe that data quality is good in females as compared to males.

It is observe that both males and females tend to misreport their ages and quality of age data varies from state to state and also from district to district.

References

Choudhary, C.R. (2006) 'A study of quality of Single year age data in India', Seminar paper submitted for the partial fulfilment for the Master of population studies, International Institute for population Sciences, Mumbai.

International Institute for Population Sciences (IIPS), 2010. District Level Household and Facility Survey (DLHS-3), 2007-08: India. Mumbai: IIPS.

NOUMBISSI A., (1992), "L'indice de Whipple modifie: une application aux donnees du Cameroun, De la Suede et de la Belgique>>, Population, 47(4), pp. 1038-1041.

Prakasam, C.P. (1984) 'On quality of age data for population count-1981, in Indian states', Paper submitted to the Annual Conference of Indian Association for the Study of the population, held at Indian institute for Management, 24th December to 27th December, 1984, Bangalore.

REGISTRAR GENERAL & CENSUS COMMISSIONER, INDIA, (1976), Census of India 1971, series 1, India, Social and Cultural Tables, Part II-C(ii), New Delhi.

REGISTRAR GENERAL & CENSUS COMMISSIONER, INDIA, (1987), Census of India 1981, Social and Cultural Tables, Tables C-1 to C-6, New Delhi.

REGISTRAR GENERAL & CENSUS COMMISSIONER, INDIA, (1997), Census of India 1991, series 1, India, Part VIA-C Series, Social-Cultural Tables, Volume 2, Tables C-3 Part A and B, C-4,C-5 and C-6, India, states and Union Territories, New Delhi.

REGISTRAR GENERAL & CENSUS COMMISSIONER, INDIA, (2005), Census of India, 2001,C series: Social and Cultural Tables, Table C-13: Single Year Age Returns by Residence and sex, New Delhi, Available at: http://www.censusindia.net/results/C_Series/c13_India.pdf(accessed 27 February 2007).

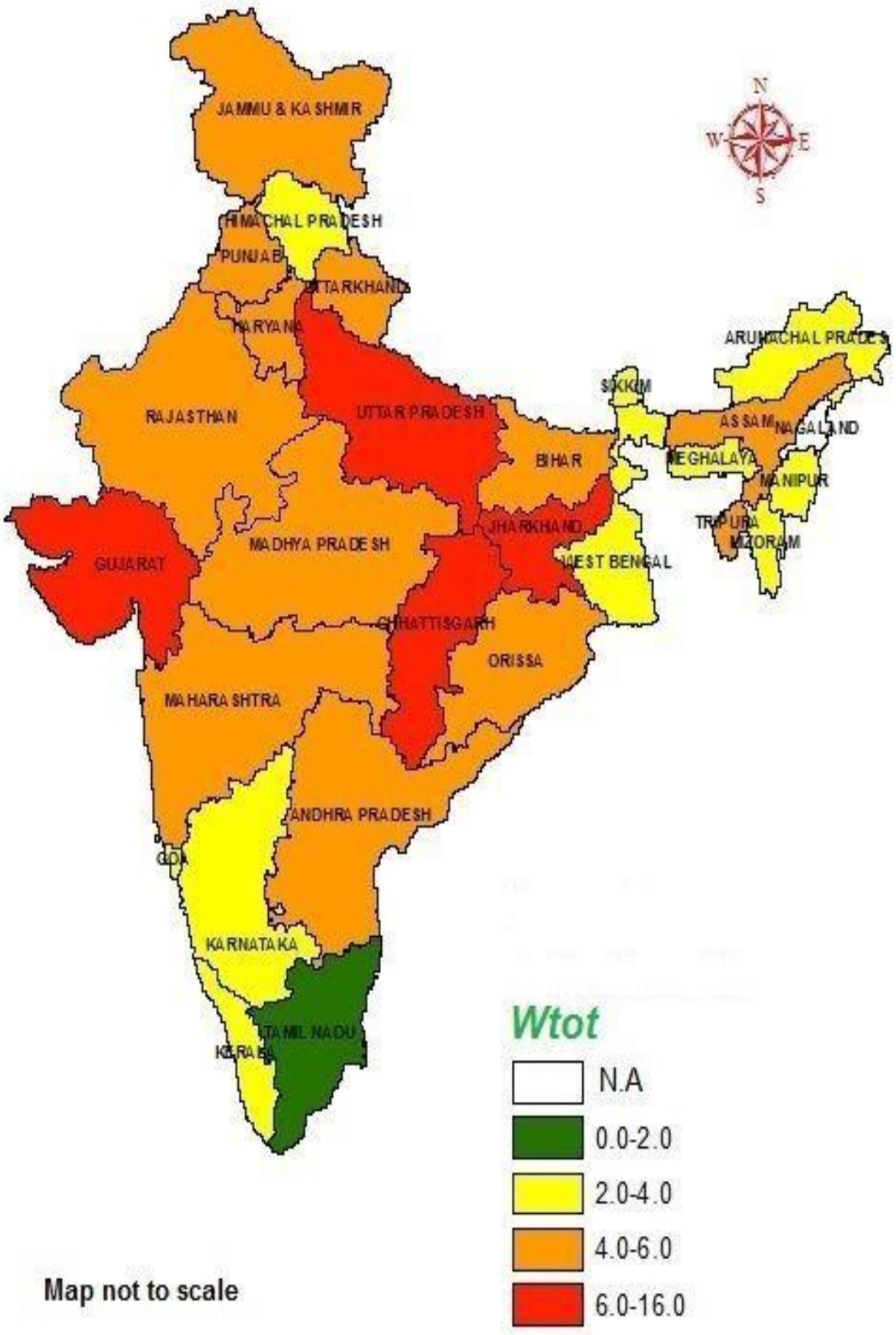
Suong, Y . (1995) 'Quality of Age Data by Sex in Censuses of Some selected Asian Countries', Seminar Paper Submitted as a part of Requirements for Diploma course in Population Studies, International Institute for population Sciences, Mumbai.

Spoorenberg, Thomas and C.Dutreuilh (2007) Quality of Age Reporting: Extension and Application of the Modified Whipple's Index, *Population (English Edition),Vol 62,No.4,P.729-741*

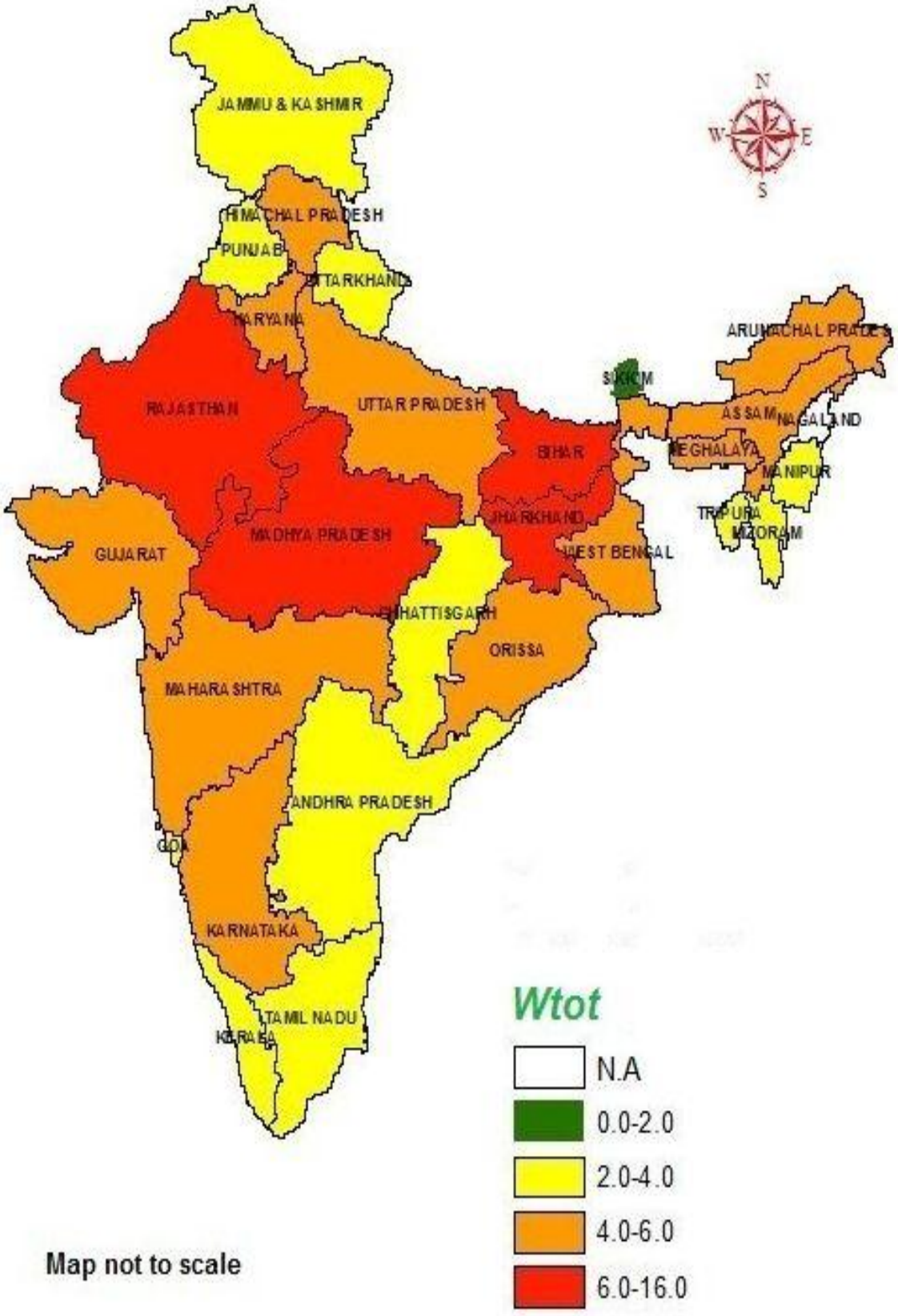
Spoorenberg, Thomas (2009) Assessing the quality of age reporting at a time of general data quality improvement: going beyond the original Whipple's index. *XXVI IUSSP International Population Conference*, Morocco 27 September, 2009, Session P-5. Morocco.

Unisa S, Dwivedi LK, Reshmi RS, Kumar K(2009).Age reporting in Indian census: An insight. Paper presented at the 26th IUSSP International Population Conference. Morocco.

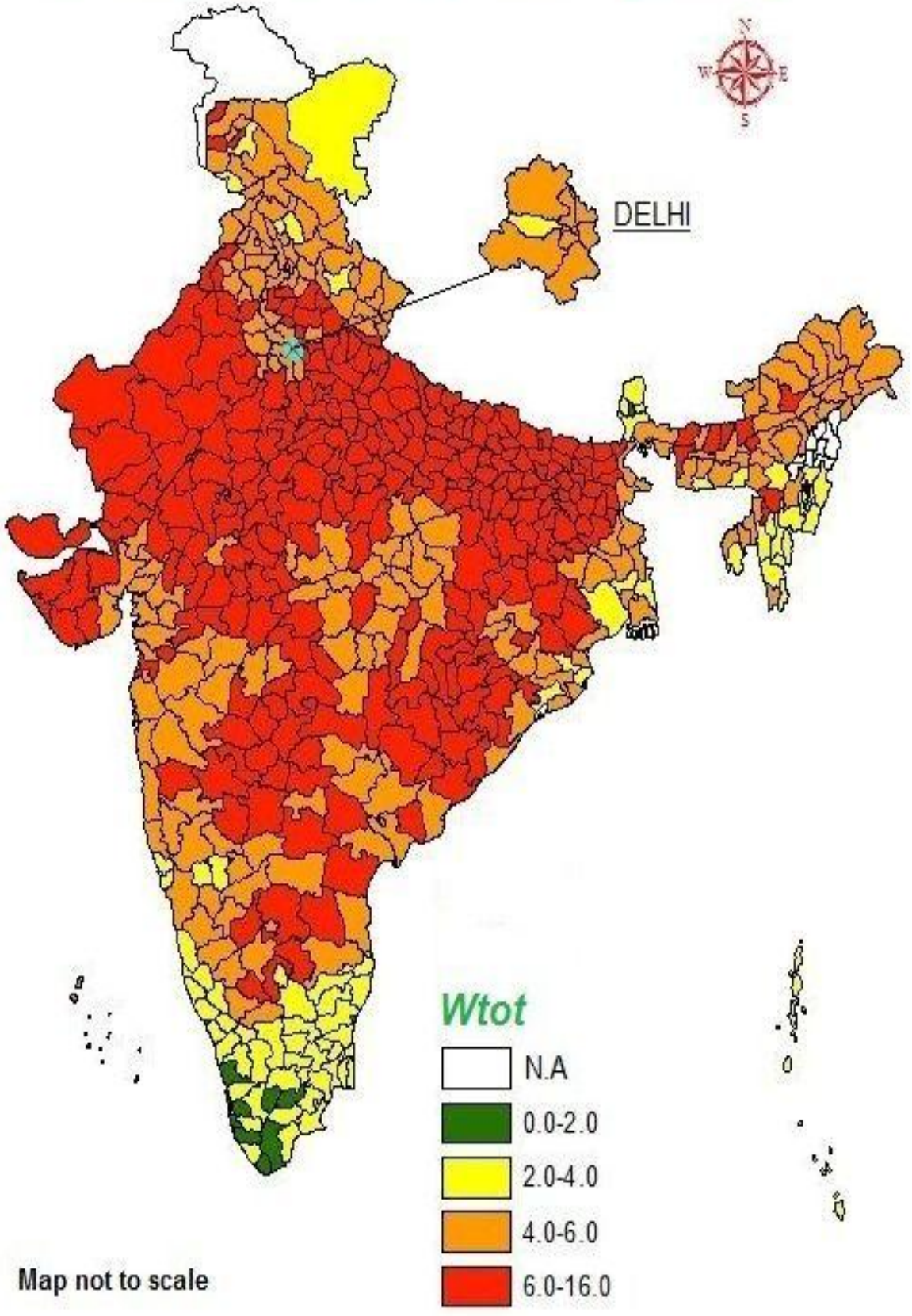
MAP1:27 STATES OF INDIA (MALES)



MAP2:27 STATES OF INDIA (FEMALES)



MAP3:601 DISTRICTS OF INDIA (MALES)



MAP4:601 DISTRICTS OF INDIA (FEMALES)