केवल कार्यालयी प्रयोग के लिए
For official use only

सत्यमेव जयते

सांख्यिकीय प्रकटन नियंत्रण (एसडीसी) के लिए मानक उद्विकसित करने हेतु विशेषज्ञ समिति की रिपोर्ट

Report of the Expert Committee to evolve standards for Statistical Disclosure Control (SDC)

राष्ट्रीय प्रतिदर्श सर्वेक्षण कार्यालय
National Sample Survey Office
राष्ट्रीय सांख्यिकीय संगठन
National Statistical Organisation
सांख्यिकी और कार्यक्रम कार्यान्वयन मंत्रालय
Ministry of Statistics and Programme Implementation
भारत सरकार
Government of India
दिसम्बर 2016
December 2016

For official use only

Report of the Expert Committee to evolve standards for Statistical Disclosure Control (SDC)

National Sample Survey Office
National Statistical Organisation
Ministry of Statistics and Programme Implementation
Government of India

December 2016

भाकृअनुप - भारतीय कृषि सांख्यिकी अनुसंधान संस्थान
लाइब्रेरी एवेन्यू, पूसा, नई दिल्ली-110 012

ICAR – Indian Agricultural Statistics Research Institute
Library Avenue, Pusa, New Delhi-110 012

डॉ. उमेश चन्दर सूद
निदेशक
Dr. Umesh Chander Sud
Director

Dated 19 December 2016

Dear Dr. Manna,

The act of disclosing information collected by statistical agencies from various respondents' to other users of data is an integral part of survey data processing and release of results. However, sometimes the form and extent of such disclosure may have the risk of divulging the identification of the responding unit along with the information provided. As such one of the prime objectives of Statistical Disclosure Control (SDC) is to safeguard the confidentiality of the information furnished by a responding unit while disclosing such data as part of the data dissemination by statistical agencies. Apart from protecting confidentiality of the respondents, the SDC mechanism is also expected to provide guidance to the users of the data about reliability of the estimates while also maximizing the availability of survey data.

Considering the importance of the different aspects of data dissemination, it was apt on the part of the Ministry of Statistics and Programme Implementation, Government of India which is also one of the prime producers of Official Statistics in the country to consider SDC as of great importance and constitute an Expert Committee to evolve standards for Statistical Disclosure Control for sample surveys conducted by NSSO.

The Committee during its tenure held detailed deliberation on all the aspects outlined in the Terms of Reference to formulate policies for SDC through in-depth examination of the available literature, conceptual issues as well as constraints of the data producers in taking pragmatic decisions for withholding or disseminating results.

The report of the Committee, in six chapters, provides overview of different aspects of SDC, analysis of results and makes specific recommendations pertaining to SDC.

While submitting the Report, I, on behalf of all the members of the Committee would like to thank the Ministry of Statistics and Programme Implementation for entrusting this task to us and express hope that this Report will facilitate the producers of data in taking objective decisions on dissemination. I would like to thank all the members of the Committee and the special invitees, without whose valuable contributions the Report would not have taken the present shape.

With regards

Yours Sincerely

(U. C. Sud)

Dr. G. C. Manna
Director General, Central Statistical Office
Ministry of Statistics and Programme Implementation
Government of India

दूरभाष (कार्यालय) Phone (O): 91-011-25841479
ई-मेल: director@iasri.res.in; dr.ucsud@gmail.com
E-mail: director.iasri@icar.gov.in; uc.sud@icar.gov.in
फैक्स Fax: 91-011-25841564
वेबसाइट Website: http://www.iasri.res.in
तार Grams: AGRIRESTA

Executive Summary

Statistical disclosure refers to the act of disclosing information collected by statistical agencies from various respondents' to other users of data. Sometimes the form and extent of such disclosure have the potential of divulging the identification of the responding unit along with the information provided. One of the prime objectives of Statistical Disclosure Control (SDC) is to safeguard the confidentiality of the information furnished by a responding unit while disclosing such data as part of the data dissemination by statistical agencies. Apart from protecting confidentiality of the respondents, the SDC mechanism is also expected to provide guidance to the users of the data about reliability of the estimates while also maximizing the availability of survey data.

No nationally accepted standard for suppression of data due to potential breaches of confidentiality or statistical reliability is available at present. The lack of a national standard, perhaps, reflects the inherent problems in prescribing a standard and the need for data analysts to use judgment in determining whether aggregated data available to the public protect its confidentiality and are reliable enough to allow users to draw reasonable conclusions.

The Committee reviewed the existing system of dissemination of microdata (unit level data) and release of estimates, results, reports and documents based on the sample surveys conducted by the National Sample Survey Office (NSSO) on different socio-economic subjects. The Committee also deliberated the dissemination policy along with the policy practised to protect the confidentiality of the respondents of the survey data implemented by other Ministries/Departments, like Ministry of Human Resource Development, Ministry of Labour and Employment, Ministry of Health and Family Welfare and Ministry of Housing & Urban Poverty Alleviation. The Committee was of the view that a data dissemination policy conforming to the SDC may be developed for MoSPI first so that various aspects and consequences of SDC are fully understood and analysed before such a policy is extended to other Ministries/Departments outside MoSPI.

The Committee members analysed RSEs of the estimates of selected characteristics from NSS data keeping in view the occurrence of rare events in published results. The study of Relative Standard Error (RSE) undertaken for the specific subjects of survey (*Employment-Unemployment and Household Consumer Expenditure*) revealed that the estimates having RSE≤30% and based on sample number of observations > 40 can be accepted for release only in certain situations. However, the Committee felt that such a rule cannot be applied uniformly across all the Tables in a Tabulation Plan even for a single subject, and perhaps, less so for other subjects of survey conducted by NSSO.

While exploring the feasibility of using R Package 'sdcMicro' to implement SDC methods to evaluate and anonymize confidential microdata sets, the Committee observed that the procedure adopted in 'sdcMicro' causes loss of information by modifying unit level data which will not generate same estimates as the original unit level data. The Committee felt that such loss of information is not desirable and the loss of information may also not be uniform across all units.

2 Data analysis for Relative Standard Error and Sample Size

The Committee discussed the findings of the analysis of the relative standard error of the estimates published by NSS surveys through its two important surveys, viz., Employment and Unemployment survey and Consumer Expenditure survey. From these two surveys, some important indicators were selected for the study.

For Employment and Unemployment survey estimates and RSEs in respect of per 1000 distribution of persons in usual status were considered (Table 19 of the Tabulation Plan of NSS 68th round). For the study of reliability of wage rate per day, estimates and RSEs of average wage/salary earnings per day for regular wage salaried work in current daily status of persons of age 15 to 59 years were considered (Table 44 of the approved Tabulation Plan of NSS 68th round). *[para. 2.2]*

For the Survey on Consumer Expenditure, reliability of the estimates of per capita value of consumption of different items was studied (Table 8.2 of Report 2 of the Tabulation Plan, Schedule 1.0). For the purpose of the study, the items of consumption were grouped into five groups namely, (a) food group, (b) energy, (c) clothing and footwear, (d) education and medicine (institutional), (e) miscellaneous goods and services, including medical (non-institutional), rents and taxes, and (f) durable goods. *[para. 2.3]*

For both the surveys, analysis was done considering detailed results at the level of States/UTs/all-India. The Committee critically assessed reliability vis-à-vis the number of sample observations involved in the generation of estimates. *[para. 2.5]*

3 Data analysis for Relative Standard Error of Rare Events

Analysis of RSEs vis-à-vis sample size revealed that for some parameters, though the sample sizes are small ($1 \leq N \leq 40$), the estimates are reliable (RSE$\leq$30%) while in some other cases of small sizes, the estimates are not reliable (RSE>30%). Such observations, led to the necessity to critically analyse characteristics of rare events. *[para. 3.1]*

Analysis was done in respect of Employment and Unemployment surveys, where such rare events may be proportion unemployed, proportion of regular wage/salaried persons in the agricultural sector, proportion of workers in age groups say, 5-9 years, 10-14 years, etc. For rare events the relationship between RSE and the value of the denominator of the proportion were studied and analysed. *[para. 3.2 and para. 3.3]*

4 Statistical Disclosure Control for Microdata Using the R Package sdcMicro

The feasibility of use of SDC for microdata using the R-package 'sdcMicro' was explored. Prof. Sarat Kumar Chettri presented different aspects of 'sdcMicro' namely Anonymization Methods, Micro-aggregation, Adding Noise, Shuffling etc., for use in SDC. The package provides multiple options for reducing the statistical disclosure risk in categorical or continuous variables. Applying the package to a microdata file results in information loss, thereby reducing the utility of the data for research. As potential uses of microdata files are vast, it is difficult to assess the information loss. In many cases the research output has direct relevance to policy making, since such research is undertaken at the behest of Government Departments by third party. *[para. 4.2]*

5 SDC in other Ministries/Departments

Information was collected on availability of unit level data and other related aspects for some of the Ministries. Specific essential information were collected from different Ministries, such as (a) Survey data collected during the last 10 years, (b) System of processing and maintaining the data, (c) Dissemination policy and (d) Policy practised to protect the confidentiality of the respondents.

Replies received from four Ministries of the Government of India namely, Ministry of Human Resource Development, Ministry of Labour and Employment, Ministry of Health and Family Welfare and Ministry of Housing & Urban Poverty Alleviation were examined by the Committee in detail. *[para. 5.1]*

6 Recommendations

Considering that Statistical Disclosure Control (SDC) primarily addresses two issues, viz., protecting the identification of the individual units and release of reliable results, the Committee has made specific recommendations covering these two aspects.

The data disseminating agency has to strike a proper balance between information loss and reduced disclosure risk. Since potential uses of microdata files are vast, it is simply impossible to undertake an exhaustive assessment of information loss. Any data dissemination policy has to strike a balance between data confidentiality and the information necessary for meaningful data analysis. Promoting data sharing is one of the key aspects of the modern day information revolution. Data are usually collected at huge cost and proper sharing of data can reduce overall cost besides reducing respondent burden. *[para. 6.2]*

The Committee recommends for constituting a *data review panel* consisting of producers and users of data which will function on a continuous basis. This panel will examine the microdata that MoSPI plans to release and if necessary, suggest techniques of SDC before their release to users. The concerned data disseminating agencies should thereafter apply the techniques to ensure compliance with the recommended SDC norms. [*para 6.3*]

The Committee recommends that the MoSPI may review the existing undertaking obtained from data users to ensure that the responsibility of maintaining microdata confidentiality of the respondents also lies with the data users. [*para. 6.4*]

The present system of anonymization practised in MoSPI is considered adequate to suppress the identity of the particular respondent. However, possibility of indirectly revealing the identity of the respondent from data specific attributes still remains. The Committee is of the view that complete elimination of such possibilities cannot be built into the disclosure control policies of MoSPI for the time being, in view of the effort required and the possibility of reducing the utility of data. No further anonymization is recommended for the time being. *[para. 6.5]*

The Committee recommends that release of reliable estimates based on a rule of RSE cannot be applied uniformly across all the Tables in a Tabulation Plan even for a single subject, and far less applicable to other subjects of survey conducted by NSSO. Therefore, there is a need to develop separate rules for different surveys after intense data analysis.

[para. 6.10]

The Committee is of the view that it is not feasible to formulate a uniform rule based on RSE and sample size to comment on the reliability of the estimates for data dissemination purpose. The Committee recognises the immense use of the released results of NSSO in the form of reports/documents etc. Therefore, the Committee recommends to release all the estimates generated on the basis of the Tabulation Plan, even if such estimates are based on small sample sizes, with an appropriate caution that such estimates are likely to be less reliable.

[para. 6.11]

To avoid publishing of less reliable estimates in Reports, based on small sample sizes, necessary abridgement of the Tabulation Plan may be done to reduce the level of disaggregation. Level of disaggregation may be decided using the experience gained from the study of results of similar surveys. *[para. 6.12]*

The committee observed that inputs received from other Ministries / Departments in respect of results/reports published, process of anonymization, etc., lack detailed information. As such the Committee feels that it difficult to recommend a uniform policy for Statistical Disclosure Control (SDC) across all Ministries. [*para.6.13*]

6 Acknowledgement

The Group takes this opportunity to express its thanks to all the SDRD and DPD officers who contributed directly or indirectly in the work of the Committee. The Group also records its appreciation for commendable services of Shri P. S. Bose, Deputy Director General and Coordinator, SDRD, NSSO, Shri D. Mukhopadhyay, Deputy Director General, SDRD, Shri Salil Kumar Mukhopadhyay, Director, SDRD, Sri Avishek Podder, Deputy Director, SDRD, and Sri Maharnab Manna, Assistant Director, SDRD, who worked incessantly and with utmost dedication during the entire tenure of the Committee.

Dr. Manisha Pal (Member)	Dr. Sarat Kumar Chettri (Member)	Dr. Diganta Mukherjee (Member)
Shri P. C. Mohanan (Member)	Additional Director General NSSO (DPD) (Member)	Additional Director General NSSO (SDRD) (Member Secretary)

CONTENTS

CONTENTS

Abbreviation	Full form
AHS	Annual Health Survey
CAB	Clinical, Anthropometric & Bio-chemical
COS Act	Collection of Statistics Act
DES	Directorate of Economics and Statistics
DLD	Distance Linkage Disclosure Risk
DLHS	District Level Household Survey
DPD	Data Processing Division
GUI	Graphical User Interface
HMIS	Health Management Information System
HUPA	Housing and Urban Poverty Alleviation
IASRI	Indian Agricultural Statistics Research Institute
IHSN	International Household Survey Network
IIPS	International Institute for Population Sciences
MDAV	Maximum Distance to Average Vector
MoSPI	Ministry of Statistics and Programme Implementation
MPCE	Monthly Per Capita Consumer Expenditure
NDSAP	National Data Sharing and Access Policy
NFHS	National Family Health Survey
NRHM	National Rural Health Mission
NSC	National Statistical Commission
NSSO	National Sample Survey Office
NUEPA	National University for Educational Planning
OGD	Open Government Data
PS+SS	Principal Status + Subsidiary Status
RSE	Relative Standard Error
RTI	Right to Information Act
SDC	Statistical Disclosure Control
SDRD	Survey Design and Research Division
SE	Standard Error
SECC	Socio-Economic Caste Census
SESDSL	Socio-Economic Surveys of Different Segments of Labour
TOR	Terms of Reference
U-DISE	Unified District Information System for Education
UT	Union Territory

Chapter One : Introduction

1.1 Conceptual framework of Statistical Disclosure Control (SDC)

1.1.1 Statistical disclosure refers to the act of the disclosing information collected by statistical agencies from various respondents' to other users of data. Sometimes the form and extent of such disclosure has the potential of divulging the identification of the responding unit along with the information provided. Statistical Disclosure Control (SDC) is concerned with the safeguarding of the confidentiality of the information furnished by a responding unit while disclosing such data as part of the data dissemination by statistical agencies.

1.1.2 *Why Statistical Disclosure Control*: Disclosure of data can occur though access to the physical records maintained by the statistical agencies. However, this is not a matter of statistical disclosure. Statistical disclosure occurs when the information is disclosed from statistical outputs including microdata released as a statistical output and persons other than the authorized statistical investigator who collected the data or the agency responsible for maintaining the data gets access to data that can be ascribed to the original respondent. Again such disclosures can happen either directly or indirectly. Direct disclosure takes place when the data disseminated contain the identity particulars of the respondent like name and addresses, telephone numbers or other direct contact information. Indirect disclosure takes place when the user can indirectly identify the respondents from the contents of the data attributes due to special features that lead to the identity of the respondent.

1.1.3 *Legal requirement for SDC*: Disclosure of information furnished by a responding unit to others is usually subject to several considerations. The most important of which is the legal aspect. Statistical legislations in most countries have explicit clauses demanding the protection of privacy of the respondents and maintaining confidentiality of the information furnished by them. The Collection of Statistics (COS) Act 2008 enacted by the Indian Parliament has very strict provisions of data confidentiality. The Act enjoins that the data collected under this statute will be used only for statistical purposes. Publication of information without suppressing the identification of informants is prohibited under this COS Act. In special situations where the respondents have specifically consented in writing to disclosure or where the information is otherwise publicly available can be disclosed. However where the individual data are disseminated to users for statistical or research purposes, the disclosure can be done after deleting the name and address of the informant; also every user provided with the individual data has to comply with the directions given by the agency authorized to collect them. Principle 6 of the fundamental principles of official statistics adopted by the UN Statistical Commission also states that the 'individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes'.

1.1.4 *Ethical and moral Issues for SDC*: Besides the legal requirements to restrict disclosure of information provided by a respondent, there is also an ethical element to the need. Statistical data collection to a great extent depends on the voluntary cooperation of the respondent and the trust the respondent holds with the statistical agency authorizing the data collection. This trust is based on the ultimate benefits the data delivers to the society and the Government. Surveys often gather data through interview method as opposed to collection of information from records where the data collector requires only permission to access the records. In the interview method trust and cooperation are vital for the successful collection of data. The agency in all such cases provides an undertaking that the data provided by the respondent will be kept confidential and used only for statistical purposes after suitable aggregations. It is, therefore, ethically important that the statistical disclosure truthfully adheres to this trust and assurance.

The quality of data collected also depends to an extent on the respondent's belief that details especially of a sensitive or personal nature are kept strictly confidential. A professional approach to data collection assumes that the information will not be shared with others and used for purposes other than that intended by the statistical agencies authorizing the data collection.

When the question of confidentiality is considered a key factor in data dissemination, the question arise as to why one may not adopt an approach to completely detach the data from its respondent. Such an extreme procedure has the potential to adversely affect the very usability of the data for which it is acquired, defeating the very purpose of data dissemination. In many cases the data sets from several sources are used for a study and a complete detachment from any identity may make such efforts impossible.

Secondly the data are collected at huge cost and it is essential to find maximum users outside the agency that collect the data to derive its full benefits. It is, therefore, logical that the Statistical Agencies attempt to disseminate maximum possible extent of data.

It is also the professed policy of the Government that the data in its possession should be accessed and shared by others. Initiatives like the Open Government Data (OGD), Right to Information Act (RTI) and the National Data Sharing and Access Policy (NDSAP) enjoins free and liberal dissemination of information held by Government.

1.1.5 *Reliability of estimates and SDC:* Estimates based on a random sample of a population are subject to sampling variability. The most common measure of the extent of sampling variability is the Standard Error (SE). It is important to consider the Standard Error when using the estimates as it affects the accuracy of the estimates and, therefore, the importance that can be placed on the interpretations drawn from the data.

The relative standard error (RSE) provides a measure of reliability for statistical estimates. The RSE is computed by dividing the standard error of the estimate by the estimate and multiplying by 100 to convert it to a percentage. When the RSE is large, the estimate is

imprecise. There is no single national standard for deciding when the RSE is so large that one should not present the data.

The vast number of different study variables or population characteristics and the different domains of interest in a survey make it impractical and almost impossible to calculate and publish standard errors for each statistic (estimated value of a population variable or characteristic) and for each domain individually. However, RSE for the important indicators estimated from the survey may be published.

The random variation may be substantial when the measure, such as rate or percentage, has a small number of events in the numerator or a small denominator. Typically, rates based on large numbers provide stable estimates of the true, underlying rate. Conversely, rates based on small numbers may fluctuate dramatically from year to year, or differ considerably from one small place to another, even when differences are not meaningful. Meaningful analysis of differences in rates between geographic areas or over time requires that the random variation in rates be quantified; this is especially important when rates or percentages are based on small numerators or denominators.

Apart from protecting confidentiality of the respondents, the SDC mechanism should also provide guidance to the users of the data about reliability of the estimates while also maximizing the availability of survey data. No nationally accepted standard for suppression of data due to potential breaches of confidentiality or statistical stability is available at present. The lack of a single national standard, perhaps, speaks to the problems inherent and requires data analysts to use judgment in determining whether aggregated data available to the public protect confidentiality and are precise and stable enough to allow users to draw reasonable conclusions.

1.2 *Constitution of the Expert Committee to evolve standards for Statistical Disclosure Control for Sample Surveys*

1.2.1 The Ministry of Statistics and Programme Implementation (MoSPI) has been disseminating unit level data ever since the National Data Dissemination Policy was announced in 1999. The policy had mandated that the data confidentiality should be maintained by deleting the identification particulars of respondents, etc.

1.2.2 Certain amount of anonymization is therefore applied on the data to remove the possibility of identifying the respondents. There are other Government Ministries/Departments, who also carry out sample surveys and bring out publications like reports/documents/tables/statements, etc. They also release unit level data for use of various users within the Government and outside.

1.2.3 The National Statistical Commission (NSC) in its 77th meeting held during 4th -5th August, 2015 noted that:

a) Statistical Disclosure Control (SDC) primarily addresses two issues, namely, protecting the identification of the individual units, and release of reliable results.

b) At present, no established standards exist for taking an objective decision about release of results or suppression of results based on the criteria of reliability of the results. It is necessary to strike a balance between the release of reliable results and the need of the users to have results which are not reliable but nevertheless useful, since in some situations, the results which are not reliable are still considered useful by the users, especially when no other estimates are available.

1.2.4 The NSC recommended that the MoSPI may constitute an Expert Committee to evolve detailed standards on SDC in respect of microdata and different types of tables publishing survey results. As a follow-up action on the recommendations made by the National Statistical Commission (NSC), an Expert Committee (henceforth referred to as the Committee) to evolve standards for Statistical Disclosure Control for sample surveys conducted by the NSSO was constituted under the Chairmanship of Dr. U.C. Sud, Director, IASRI, New Delhi with the following members:

a) Dr. U. C. Sud, Director, IASRI, Delhi, Chairman,
b) Dr. Manisha Pal, Professor, Department of Statistics, University of Calcutta, Member,
c) Dr. Sarat Kumar Chettri, Asstt. Professor, Assam Don Bosco University, Member,
d) Dr. Diganta Mukherjee, Associate Professor, ISI, Kolkata, Member
e) Shri P. C. Mohanan, Deputy Director General (Retd.), MoSPI, Member
f) Additional Director General NSSO (DPD), MoSPI, Kolkata, Member
g) Additional Director General NSSO (SDRD), MoSPI,, Kolkata, Member Secretary

1.2.5 The broad Terms of Reference (TOR) of the Committee is given below:

(i) To review the existing system of dissemination of microdata (unit level data) and release of estimates, results, reports and documents based on the sample surveys conducted by the National Sample Survey Office (NSSO) on different socio-economic subjects.

(ii) To evolve and recommend standards on Statistical Disclosure Control (SDC) in respect of sample surveys for dissemination of unit level data and release of estimates & results at various disaggregated levels by NSSO and other Government Ministries/Departments through electronic/digital media, various reports/ documents/ tables/statements, etc., by taking into consideration reliability of estimates vis-à-vis requirement of the users for detailed data & results, confidentiality of data and such related aspects.

(iii) To recommend a suitable mechanism/procedure for dissemination of data & results of sample surveys in line with the National Data Sharing and Accessibility Policy-2012.

1.2.6 The Office Memorandum constituting the Committee is in Annexure I.

1.3 Details of the activities of the Committee after its formation

1.3.1 ***First Meeting of the Committee***: The first meeting the Committee was held at Kolkata on 22.02.2016. In the first meeting of the Committee, the TOR of the committee, the concepts relating to SDC and its role in sample surveys were discussed in detail. The Committee members discussed at length findings of the study of RSE of the estimates of selected characteristics from the Employment and Unemployment and Consumer Expenditure surveys of NSSO presented by SDRD Officers. This study of RSE along with its findings is presented in Chapter Two of this document.

1.3.2 The Minutes of the first meeting of the Committee is given in Annexure II. In the first meeting of the Committee, broadly, the following decisions were taken.

(i) It was decided that letters will be addressed to some selected Ministries with a request to share information relating to dissemination of survey data during the last 10 years, to evolve standards of SDC for such Ministries also.

(ii) It was decided Prof. Sarat Kumar Chettri, Member, Expert Committee will look into different aspects of SDC for microdata using the R-package sdcMicro.

(iii) In the first meeting, SDRD presented the findings of a study on the Relative Standard Errors (RSE) and sample observation. It was decided that further analysis will be done considering the denominator rule which is based on number of sample observations appearing in the marginal totals in rows or columns of the Tables of a Tabulation Plan.

1.3.3 ***Second Meeting of the Committee***: The second meeting of the Committee was held at Kolkata on 19.08.2016.

1.3.4 As a follow-up action of the discussion held in the first meeting, various Ministries were requested to provide detail information on Survey data collected during the last 10 years, System of processing and Dissemination policy along with policy practised to protect the confidentiality of the respondents of the survey data. Inputs received only from Ministry of Human Resource Development, Ministry of Labour and Employment, Ministry of Health and Family Welfare and Ministry of Housing & Urban Poverty Alleviation were discussed in the meeting and are presented in detail in Chapter Five of this Report.

1.3.5 The extended study of RSE of the estimates of selected characteristics from the survey data of NSSO to include analysis of rare events was presented before the Committee members and findings were discussed. The extended study of RSE is available in detail in Chapter Three of this Report.

1.3.6 Prof. Sarat Kumar Chettri, Member of the Committee presented a paper titled *Statistical Disclosure Control for Microdata Using the R Package 'sdcMicro'*. The paper which explored feasibility of using R Package 'sdcMicro' to implement SDC methods to

evaluate and anonymize confidential microdata sets is included in Chapter Four of this Report.

1.3.7 The Minutes of the second meeting of the Committee is given in Annexure III. In the second meeting of the Committee, broadly, the following decisions were taken:

(i) It was opined by the members that a data dissemination policy conforming to the SDC may be developed for MoSPI first so that various aspects and consequences of SDC are fully understood and analysed before such a policy is extended to other Ministries/Departments outside MoSPI.

(ii) The Committee felt that generation of estimates based on very small observations may not be appropriate. As such there is a need to avoid such situations, by reformulating Tabulation Plans after taking care of specific requirements from different subjects of NSSO. The Committee suggested that for a set of key indicators Tables at the all India level may be prepared at a more detailed format whereas Tabulation at the State/ UT level may be prepared at a broad level of disaggregation.

(iii) The study of Relative Standard Error (RSE) shows that the estimates having RSE≤30% and based on sample number of observations > 40 can be accepted for release. However, such a rule cannot be applied uniformly across all the Tables in a Tabulation Plan on a subject of survey as well as across all the subjects of survey conducted by NSSO. Therefore the Committee suggested that subject-wise separate rules may be developed after thorough analysis of data and considering the requirements of data users.

(iv) The procedure adopted in 'sdcMicro' package may change some records of the unit level data; the modified data will not generate same estimates as the unit level data due to loss of information. The Committee felt that such loss of information is not desirable and the loss of information may also not be uniform across all units. Also, reduction of the chance of identifiability increases the information loss, which can affect the estimates considerably.

1.3.8 ***Third Meeting of the Committee:*** The report of the Committee was finalised in its third meeting held on 6th December 2016 at NSSO, Kolkata.

Chapter Two : Data analysis for Relative Standard Error and Sample Size

2.1 Survey topics studied

2.1.1 The Committee discussed the findings of the analysis of the relative standard error of the estimates published by NSS through its two important surveys, viz., Employment and Unemployment survey and Consumer Expenditure survey. Form these two surveys, some important indicators were selected for the study. The details of the findings of the study are given below.

2.2 Employment and Unemployment Survey

2.2.1 Distribution **of persons by activity status in the usual status (ps+ss[1]) for age groups**: Table 19 of the Tabulation Plan of NSS 68th round survey on Employment and Unemployment gives the '*the per 1000 distribution of persons by usual status (ps+ss) for each age-group'.* The Table 19 at the all-India level has detailed level of disaggregation at '*sector x gender x 5-year age groups x status x industry*' from which key employment and unemployment parameters can be generated for any domain (state *x* sector *x* gender *x* age). The Structure of Table 19 is given in Annexure IV.

For studying the reliability of the estimates given in different cells of Table 19, the sample number of observations as well as the RSE for each of the cells of this Table were worked out at detailed level of disaggregation. The scatter diagrams presented in Graphs 1 to 4 (Annexure V) of the RSEs of the estimates and the corresponding sample observations of each cells show negative correlation. The pattern displayed by the Graphs show that RSEs remain generally within 30% when sample number of observations netted in a cell is more than 40. The detailed results are presented in Table 2.1 combining all the States/UTs/all-India. Table 2.1 shows that when the sample number of observations are more than 40 for any cell, such estimates are generally reliable.

Table 2.1: Number of sample observations and corresponding RSE for the estimates of per 1000 distribution of person by usual activity status (ps+ss)
[*all the cells of the table for state/UT/all-India have been considered*]

numerator	total no. of cells	percentage distribution of cells	no. of cells with RSE≤30%	proportion of cells with RSE ≤30%
1-40	17395	51.4	5068	29.1
> 40	16426	48.6	16426	99.4
all	**33821**	**100.0**	**21389**	**63.2**

[1] Usual activity status (US) relates to the activity status of a person during the reference period of last 365 days preceding the date of survey. The activity status on which a person spent relatively longer time (major time criterion) is considered the usual principal activity status. Besides the usual principal activity status, a person may be engaged in economic activity, during the last 365 days, for a period of 30 days or more. The status in which such economic activities are pursued is the subsidiary economic activity (SS) of the person. The usual status (ps+ss) is obtained by considering the usual principal status and the subsidiary status together.

The RSE and the sample count were also studied in respect of each of the 35 States/UTs, by calculating the percentage of cells with RSE less than 30% for the following two categories of sample observations in the cells:

1. Cells where the sample number of observations are 6-40
2. Cells where the sample number of observations are more than 40.

The findings are presented in the Graphs 5 to 8 (Annexure VI). It is observed from the graphs that when the sample count of the cells are more than 40, almost all such cells have RSE less than 30% irrespective of the domain of study (state *x* sector *x* gender). For sample number of observations less than 41 (say 6-40), the percentage of cells with RSE less than 30% are very small.

2.2.2 **Wage rate**: For the study of reliability of wage rate per day, Table 44 of the approved Tabulation Plan of NSS 68th round survey on Employment and Unemployment was considered which *gives average wage/salary earnings per day for regular wage salaried work in current daily status for persons of age 15-59*. The structure of Table 44 is given in Annexure VII.

In Graphs 9 to 12 (Annexure VIII), the RSE and the sample number of observations corresponding to the estimates presented in the cells of the Table 44 have been presented. It is seen that for the cells where more than 40 sample observations are netted, RSEs of the estimates of such cells are generally reliable in the sense that the RSE is less than 30%.

The detailed results combining all the States/UTs/all-India are presented in Table 2.2.The table shows that 41.8% of the cells where the estimate of average wage rate per day is based on 1- 40 number of sample observations have less than 30% RSE while 94% of the cells with more than 40 sample observations have less than 30% RSE. Thus, wage rates for regular wage/salaried works calculated on the basis of more than 40 observations can be accepted as reliable.

Table 2.2: Number of sample observations and corresponding RSE of the estimates of wage rage per day for regular wage/salaried work

[*all the cells of the table for state/UT/all-India have been considered*]

no. of sample observations in the cells	total no. of cells	percentage distribution of cells	no. of cells with less than 30% RSE	percentage of cells with RSE less than 30%
1-40	10658	76.7	4451	41.8
> 40	3242	23.3	3050	94.1
All	13900	100.0	7501	54.0

At the State/UT level it is seen that generally for each State/UT the percentages of cells with number of observations 6-40 have more than 30 per cent RSE while the cells with more than 40 sample observations have less than 30 per cent RSE (Graphs 13 to 16 in Annexure IX). Thus, it is seen that the estimates of the cells with more than 40 observations are generally reliable (RSE less than 30 per cent) for all the States/UTs.

2.3 Consumer Expenditure Survey

2.3.1 **Per capita value of consumption**: For the Survey on Consumer Expenditure, reliability of the estimates of *per capita value of consumption different items* (Table 8.2 of Report 2 of the Tabulation Plan, Schedule 1.0) were studied. The structure of Table 8.2 is given in Annexure X. For the purpose of the study, the items of consumption were grouped into six groups as follows:

(i) food group[2]
(ii) energy[3]
(iii) clothing and footwear
(iv) education and medicine (institutional),
(v) miscellaneous goods and services including medical (non-institutional), rents and taxes, and
(vi) durable goods

The detailed results of the cell level number of observations and RSE of the cell level estimates combining all the domains of measurements (state *x* sector) are presented in Table 2.3.

Table 2.3: Number of sample observations and corresponding RSE of the estimates of the cells of per capita value of consumption of (i) food group, (ii) energy, (iii) clothing and footwear, (iv) education and medicine (institutional), (v) miscellaneous goods and services including medical (non-institutional), rents and taxes, and (vi) durable goods

[*all the cells of the table for state/UT/all-India have been considered*]

no. of sample observations in the cells	total no. of cells	percentage distribution of cells	no. of cells with less than 30% RSE	percentage of cells with RSE less than 30%
		food group		
1 - 40	5215	31.3	607	11.6
> 40	11467	68.7	10729	93.6
all	16682	100.0	11336	68.0
		energy		
1 - 40	519	34.8	24	4.6
> 40	973	65.2	926	95.2
all	1492	100.0	950	63.7
		clothing and footwear		
1 - 40	703	16.6	147	20.9
> 40	3537	83.4	3391	95.9
all	4240	100.0	3538	83.4
		education and medicine		
1 - 40	404	24.1	49	12.1
> 40	1274	75.9	1063	83.4

[2] Food group: cereals, pulses, milk and milk products, sugar and salt, edible oil, egg, fish and meat, vegetables, fruits, spices, beverages and processed food and pan, tobacco and intoxicants
[3] Energy: fuel, light and household appliances

Table 2.3: Number of sample observations and corresponding RSE of the estimates of the cells of per capita value of consumption of (i) food group, (ii) energy, (iii) clothing and footwear, (iv) education and medicine (institutional), (v) miscellaneous goods and services including medical (non-institutional), rents and taxes, and (vi) durable goods

[*all the cells of the table for state/UT/all-India have been considered*]

no. of sample observations in the cells	**total no. of cells**	**percentage distribution of cells**	**no. of cells with less than 30% RSE**	**percentage of cells with RSE less than 30%**
all	1678	100.0	1112	66.3
		miscellaneous goods and services		
1 - 40	2766	32.0	291	10.5
> 40	5884	68.0	5271	89.6
all	8650	100.0	5562	64.3
		durable goods		
1 - 40	2920	48.8	223	7.6
> 40	3063	51.2	2359	77.0
all	5983	100.0	2582	43.2

It is seen that for the cells with more than 40 observations, for 'food group', 'energy', 'clothing and footwear', more 90 per cent of the estimate are reliable while only 77 per cent of the estimates relating to 'durable goods', 90 per cent of the estimates relating to 'miscellaneous goods and services' and 83 per cent of the estimates relating to 'education and medicine' are reliable.

Thus, for 'durable goods', 'miscellaneous goods and services' and 'education and medicine' groups, even more than 40 observations are not sufficient for getting reliable estimates.

A study of the MPCE for each of the 12 fractile classes of MPCE (*5% each for the bottom two and top two classes and 10 per cent each for the remaining 8 classes*) shows that for each of the States/UTs, the estimates are reliable.

The RSE at the all-India level of each of the Tables (for *Employment and Unemployment as well as for Consumer Expenditure*) studied in this document is given in Annexure XI.

2.4 Observations from the studies

2.4.1 From the study of the RSE and the sample sizes of the estimates of employment and unemployment and consumer expenditure, the following points emerge:

(i) estimates relating to employment and unemployment:

a) For the employment and unemployment estimates, when number of sample observation in a cell is >40, the estimates are in general reliable in the sense that the estimates of the corresponding cells have RSE ≤30%.

b) There are some instances when the RSE ≤30% but sample observation of the corresponding cells are <40.

c) The above two findings holds true irrespective of the domain for which the estimates are generated.

(ii) estimates relating to consumer expenditure:

a) RSE of estimates of each of the 12 fractile classes of MPCE for each of the States/UTs are reliable.

b) RSEs of estimates of 'food group', 'energy', 'clothing and foot wear', are generally reliable when number of sample observation in a cell is >40.

c) RSEs of estimates of 'education and medicine', 'miscellaneous goods and services', 'durable goods', are not reliable even when the of sample observation in a cell is >40

(iii) Requirement of further studies: The findings on the study of the estimates of selected parameters of employment and unemployment and consumer expenditure may not be true for other type of estimates brought out by NSSO for which further detailed study will be necessary on the following topics:

- Quarterly estimates of employment and unemployment indicators
- Estimates of some smaller States/UTs for some specific characteristics
- Estimates of absolute number of different characteristics (i.e., aggregate estimates in place of ratio estimates)
- Other characteristics of Consumer Expenditure
- Estimates of different parameters for social group and other specific groups
- Enterprise survey
- Health survey
- Education
- Disability
- Debt and Investment
- Land and Livestock Holding
- Situation Assessment Survey of Agricultural Households
- Tourism
- Housing
- Slum

2.5 **Observations of the Committee:** The Committee was of the view that the relation of RSE and sample sizes emerging from the study of the surveys of Employment and Unemployment and Household Consumer Expenditure may not hold good for the estimates of all the parameters generated from these studies and for different subjects on which NSSO undertakes surveys. Therefore, detailed study may be required for each and every subject of NSS surveys.

Chapter Three : Data analysis for Relative Standard Error of Rare Events

3.1 Need for study of the RSE of rare events

3.1.1 The findings of the study of the Relative Standard Error (RSE) and the sample size as discussed in Chapter Two (*ref: Table 19 of the Tabulation Plan of the Employment and Unemployment survey of NSSO*) led to the formation of a decision matrix of the following structure:

Table 3.1: Decision matrix for RSE of the estimate and value of numerator (for N≥1)

RSE / N	RSE≤30%	RSE>30%	all
1≤N≤40→	since RSE is smaller but N is also smaller further analysis based on marginal total (i.e., denominator) is needed (no. of cell:5068, percentage: 15.0%)	estimates are not acceptable (no. of cell: 12327, percentage: 36.4%)	no. of cell: 17395, percentage: 51.4%
N>40→	estimates are acceptable (no. of cell 16321:, percentage: 48.3%)	estimates are acceptable but with caution (no. of cell: 105, percentage: 0.3%)	no. of cell: 16426, percentage: 48.4%
all	no. of cell: 21389, percentage: 63.2%	no. of cell: 12432, percentage: 36.8%	no. of cell: 33821, percentage: 100%

Note: N: numerator, RSE: Relative Standard Error

Table 3.1 shows that for some of the estimates, though the sample sizes are smaller (1≤N≤40), the estimates are reliable (RSE≤30%) while in some other cases of small sizes, the estimates are not reliable (RSE>30%). Such observations, led to the necessity to look into the denominator for those rare events.

3.2 Magnitude of RSE of rare events

3.2.1 Let p be the proportion obtained from a sample corresponding to the population parameter P. The formula $e = \sqrt{\frac{p(1-p)}{n}}$ gives the absolute error. The relative standard error (RSE) in such cases is represented as $\sqrt{\frac{(1-p)}{np}}$. This shows that for rare events (*small p,*) the relative standard error can be large, even with very big samples. In Employment and Unemployment surveys, such rare events may be proportion unemployed, proportion of regular wage/salaried persons in the agricultural sector, proportion of workers in age groups say, 5-9 years, 10-14 years, etc. The study of the magnitude of the estimates and the corresponding RSE is presented in Table 3.2.

Analysis of estimate and the corresponding RSE obtained from the Employment and unemployment survey of NSSO (*ref: Table 19 of the Tabulation Plan of the Employment*

and Unemployment survey of NSSO) show that a higher percentage of estimates with bigger values have smaller RSE than estimates with smaller values. The results are given in Table 3.2. Table 3.2 shows that when the values of the estimates are smaller (say ≤9 per cent), the RSEs are generally higher while the RSEs are lower when the values of the estimates are higher (say>9 per cent).

The formula of RSE for rare events discussed above show that magnitude of an estimate at the all-India level may be small (*p small*) but since it is based on large number of observations (*n big), np* will be moderately large and as such the RSE of that estimate at the all-India level will be small. However, such rare events at the all-India level when are estimated for small domain (say State/UT level when *n* is small but magnitude of p is similar to all-India level), the RSE will be larger, since *np* in such cases will be smaller. Such arguments may lead to conclude that the rare events may be estimated with higher degree reliability for the bigger domains than the smaller domains.

The analysis of RSE of the estimates which are very small at the all India level (≤0.1 per cent) and their corresponding RSE at the State/UT level have been presented in Table 3.3. It is seen that at the all-India level, nearly 47.3 per cent of such estimates are reliable. However, at the State/UT level small percentages of such estimates are reliable.

Table 3.2: Percentage of estimates

RSE \ estimate	value of estimate	
	small (≤0.09)	big (>0.09)
RSE undefined	36.64	-
≤30%	27.49	94.91
>30%	35.86	5.39
all	100.00	100.00

Table 3.3: Percentage of smaller estimates (value of estimate≤0.001 at all India level)

state level RSE \ all-India RSE	≤30% (47.37%)	>30% (52.63%)
RSE undefined	71.43	1.11
≤30%	2.43	11.11
>30%	23.60	87.78
all	100.00	100.00

3.3 **Magnitude of denominator for the estimates based on small samples**

3.3.1 For rare events (*defined for the purpose of this study as the estimates based on small ≤40 sample observations*) a findings of the study on the relationship between RSE and the value of the denominator has been presented in Table 3.4 (*ref: Table 19 of the Tabulation Plan of the Employment and Unemployment survey of NSSO*). It is seen that for large value of denominator (D) also, only 30 per cent of the estimates of the rare events have RSE ≤30%.

Table 3.4: Value of denominator and corresponding RSE for the estimates of per 1000 distribution of person by usual activity status (ps+ss)

[*all the cells of the table for state/UT/all-India have been considered*]

Denominator (for numerator≤40)	total no. of cells	percentage distribution of cells	no. of cells with RSE≤30%	proportion of cells with RSE RSE≤30%
D≤ 40	800	5.4	323	40.4
D > 40	13932	94.6	4402	31.6
all	14732	100.0	4725	32.1

Based on the study of numerator, denominator and RSE, the revised decision matrix is presented in Table 3.5:

Table 3.5: Decision matrix for RSE of the estimate, value of numerator and denominator

RSE / N	RSE≤30%		RSE>30%		all
	D ≤ 40	**D>40**	**D ≤ 40**	**D>40**	
2≤N ≤ 40→	estimates are not acceptable (no. of cell: 323, percentage:1.0)	estimates are acceptable with caution (no. of cell: 4402, percentage:14.1)	estimates are not acceptable (no. of cell: 477, percentage:1.5)	estimates are acceptable with caution (no. of cell: 9530, percentage:30.6)	no. of cell: 14732, percentage:47.3
N>40→	estimates are acceptable (no. of cell: 16321, percentage:52.4)		estimates are acceptable with caution (no. of cell: 105, percentage:0.3)		no. of cell: 16426, percentage:52.7
	no. of cell: 21046, percentage:67.5		no. of cell: 10112, percentage:32.5		no. of cell:, 31158, percentage:100

Note: N: numerator, D: denominator, RSE: Relative Standard Error

3.4 **Observations of the Committee:** The Committee is of the view that for the rare events, even a large value of denominator does not provide reliable estimates for a significant proportion of the estimates. Thus, RSE of the estimate may not be the only factor for taking a decision on release of estimates and for Statistical Disclosure Control purpose a rule based on RSE may not be appropriate. Moreover, it was felt by the Committee that no uniform rule based on RSE and sample size can be formulated to comment on the reliability of the estimates for data dissemination purpose. In this respect detailed level of tabulation may be avoided and necessary abridgement of the Tabulation Plan may be adopted to avoid deeper level of classification specially for the rare events.

Chapter Four: Statistical Disclosure Control for Microdata Using R Package sdcMicro

4.1 Methods and Practices in SDC

4.1.1 In recent years several techniques have been evolved by researchers and statistical agencies in anonymizing microdata. While everyone recognizes the need for anonymizing the microdata, data users would like to get data without much loss of information that was originally collected from the respondent. Mere deletion of identification does not give adequate protection from identity disclosure.

In this context two approaches are generally suggested to ensure the 'safety' of data. One is the 'Safe Data' where restricted data is provided by modifying removing portions of data to protect against disclosures. This can be done by combining categories to enlarge the frequencies under any specific categories or by adding noise or perturbations to the data.

The other approach is to provide a 'safe setting' for data access. In this case, users may access data in controlled settings like safe data centers. This has the disadvantage that the data is no more portable besides denying access to people who are physically removed from the safe settings.

Another possibility is for providing remote access to users where the user can send their programs that are then run on the data in the data center and only output screened for appropriate disclosure are forwarded to the users. Such remote access comes with a higher cost of setting up the data centers besides limiting the number of users. In the Indian context the second approach is no more feasible as a more liberal data dissemination policy is currently being practiced. We therefore need to focus on the safe data approach.

It is necessary to measure the disclosure risk in microdata by carefully analyzing the possibilities of identity disclosure from the data. Identification and tabulating the combination of key variables such as sex, age, place of resident, occupation, industry classification and number of workers etc usually provide enough clues to identify the disclosure risks involved. The agency releasing the data also should have a microdata review panel responsible for clearing the data for release.

There are methods to estimate the disclosure risks based on probabilistic models based on the sample data. One common procedure is to mask the data by recoding some key variables. Other methods are to alter the data by sub-sampling from the actual sample, adding noise to some data especially where the data is continuous and it is possible to make some approximation of its distribution, record swapping between regions, randomizing some values etc. However, in such cases it is necessary to first examine the impact of such techniques in the comparability of the resulting estimates from those published by the agency

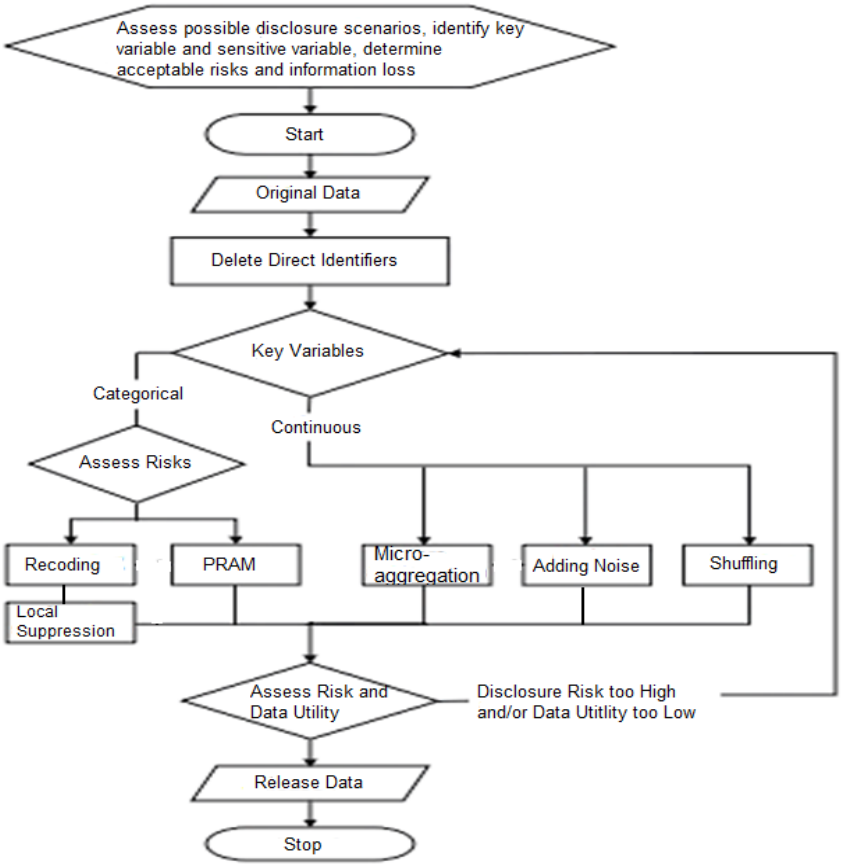
The International Household Survey Network (IHSN) has recently released a software called 'sdcMicro'[4] which is free, R-based open-source software for the generation of protected microdata for researchers and public use. The package provides multiple options for reducing the statistical disclosure risk in categorical or continuous variables.

4.2 Use of sdcMicro for SDC

4.2.1 The feasibility of use of SDC for microdata using the R-package sdcMicro was explored. Prof. Sarat Kumar Chettri prepared a paper on different aspects of sdcMicro for use in SDC. The major topics discussed in the paper are given below:

4.2.2 Introduction: The R package sdcMicro serves as an easy implementation of SDC methods to evaluate and anonymize confidential microdata sets. It includes all popular disclosure risk and perturbation methods. It serves as an easy-to-handle, highly interactive tool for users who want to use the sdcMicro package but are not familiar with the native R command line interface. The package performs automated recalculation of frequency counts, individual and global risk measures, information loss and data utility after data anonymization. All methods are highly optimized in terms of computational costs to be able to work with large data sets. Reporting facilities that summarize the anonymization process can also be easily used by practitioners.

4.2.3 Workflow for Applying common SDC Methods to microdata



[4] Guidelines for the Anonymization of Micro-data Using R-package sdcMicro (Version 1.0) Bernhard Meindl, Matthias Templ and Alexander Kowarik Vienna, April 22, 2013

4.2.4 Anonymization Methods: Generally, two kinds of anonymization methods can be distinguished: deterministic and probabilistic. For categorical variables, recoding and local suppression are deterministic procedures while swapping and PRAM (Gouweleeuw et al. 1998) are based on randomness and considered probabilistic methods. For continuous variables, micro-aggregation is a deterministic method while adding correlated noise (Brand 2002) and shuffling (Muralidhar and Sarathy 2006) are probabilistic procedures. In this report anonymization methods for continuous variables namely micro-aggregation, noise addition and shuffling are discussed.

4.2.5 Micro-aggregation: Micro-aggregation is a perturbative method that is typically applied to continuous variables. The idea is that records are partitioned into groups; within each group, the values of each variable are aggregated. Typically, the arithmetic mean is used to aggregate the values, but other methods to calculate the mean are also possible (e.g., the median). Individual values of the records for each variable are replaced by the group aggregation value.

Following five micro-aggregation methods are available in sdcMicroGUI:

- **mdav** (Domingo-Ferrer and Mateo-Sanz, 2002) is the Maximum Distance to Average Vector method that groups records based on classical (Euclidean) distances in a multivariate space.
- **rmd** (Templ and Meindl, 2008) groups records based on Robust Mahalanobis Distance
- **pca** (Templ, 2008) is a projection method that sorts data on the first principal component.
- **clustpppca** (Templ, 2008) applies the robust counterpart of the pca method to clustered data; it is feasible for small or medium-size datasets.
- **influence** (Domingo-Ferrer et al., 2002) clusters the data and sorts the records by the most influential variable in each cluster.

For computational reasons, mdav method is recommended (Templ et al. 2015). The algorithm is shown in Table 4.1 with a sample code in R console using sdcMicro package (See Table 4.2).

Table 4.1: MDAV (Maximum Distance to Average Vector) Method

1. Compute a global mean X of the data vectors.
2. Find the most distant record Xr from X and Xs from Xr.
3. Form two groups around Xr and Xs with their respective closest k-1 vectors.
4. Repeat steps 1 to 3 as long as there exist at least 2k vectors.
5. If there exist data vectors between k and 2k-1 which does not belong to any of the formed groups then form a new group and exit.
6. If there are less than k data vectors left, then assign the vectors to their closest group.
7. Assign the arithmetic mean of the respective group to each vector.

Table 4.2: A Sample Code to implement MDAV method in R console using sdcMicro package

```
R> sdc <- microaggregation(sdc, aggr = 4, strata_variables = "age", method = "mdav")

R> print(sdc, "numrisk")

Disclosure Risk is between: [0% ; 2.77%] (current)

(orig: ~100%)

Information Loss: IL1: 0.42 (orig: Information Loss: 0)

Difference Eigenvalues: 1.5 %
```

4.2.6 Adding Noise: Adding noise is a perturbative protection method for microdata which is typically applied to continuous variables. It is possible to add uncorrelated random noise which preserves means but variances and correlation coefficients between variables cannot be preserved. Correlated noise is preferred, as it preserves the covariance of the original data. The method of adding noise should be used with caution, as the results depend greatly on the parameters chosen. A sample code to implement add noise method is shown in table 3.

Table 4.3: A Sample Code to implement correlated2 method in R console using sdcMicro package

```
R> sdc <- undolast(sdc)

R> sdc <- addNoise(sdc, method = "correlated2")

R> print(sdc, "numrisk")

Disclosure Risk is between: [0% ; 32.84%] (current)

(orig: ~100%)

Information Loss: IL1: 0.11 (orig: Information Loss: 0)

Difference Eigenvalues: 0.88 %
```

4.2.7 Shuffling: Shuffling (Muralidhar and Sarathy 2006) generates simulated values for selected sensitive variables based on the conditional density of sensitive variables given non-sensitive variables. The method then replaces the ranked new values with the ranked original values. As a rough illustration, assume we have two sensitive variables, income and savings, which contain confidential information. We first use age, occupation, race and education variables as predictors in a regression model to simulate a new set of values for income and savings. We then apply reverse mapping (i.e., shuffling) to replace ranked new values with the ranked original values for income and savings. This way, the shuffled data consists of the original values of the sensitive variables. To implement this method, we need to select the shuffling method, regression method and covariance method in sdcMicroGUI. The methods are not discussed in this report.

4.2.8 Measuring Data Utility: Measuring data utility of the microdata set after disclosure limitation methods is encouraged to assess the impact of these methods. However, in practice it is not possible to create an anonymized file with the same structure as the original file. An important goal, however, should always be that the difference in results of the most important statistics based on anonymized and original data should be very small or even zero.

4.2.9 Information Loss: To assess the impact of the perturbation methods after applying the disclosure limitation on the microdata sets the parameter *Information Loss* (*IL1*) (Mateo-Sanz et al. 2004) is used which is computed as shown below:

$$IL1 = \frac{1}{p}\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{|x_{ij} - x'_{ij}|}{\sqrt{2S_j}}$$

Where, x_{ij} and x'_{ij} are the values of the original data set X and perturbed data set X' respectively. Suppose both datasets consist of n records and p variables each and S_j is the standard deviation of j-th variable.

4.2.10 Eigen: There is also another parameter used to assess the impact of the perturbation methods after applying the disclosure limitation on the microdata sets. It is the relative absolute differences between eigenvalues of the co-variances from standardized original and perturbed values of continuous key variables (Templ and Meindl, 2008). Eigenvalues can be estimated from a robust or classical version of the co-variance matrix.

4.2.11 Measuring Disclosure Risk: The assessment of the quality of an SDC method cannot be limited to information loss; disclosure risk is another magnitude that should be measured. The method that optimizes the tradeoff between both magnitudes subject to some user requirements turns out to be the best option. Mateo-Sanz, Sebe and Domingo-Ferrer (2004) introduced (a). Distance-based record linkage and (b). Interval disclosure.

4.2.12 Distance-based Record Linkage: It is based on the probability of inferring the original record from the anonymized table. It can be defined as for any anonymized record X' in an anonymized table D' if we compute a distance to other records in the original table D, we can get a nearest record X_1 and a second nearest record X_2. If X_1 or X_2 refers to the original record X, then the record X is called a linked_record. We proceed the same way for every record in the protected data set D'. The Distance Linkage Disclosure Risk (DLD) model is adopted to compute.
Let num_linked_record be the number of linked records in an anonymity table, total_num_record be the total number of records in an anonymity table, DLD (Pagliuca 1999) model is computed as shown below:

$$DLD = \frac{num\_of\_linked\_record}{total\_no\_of\_record}.100$$

4.2.13 Interval Disclosure: Interval disclosure (Pagliuca 1999) simplifies the distance-based record linkage and thus is more applicable for large datasets. In this approach, after

applying SDC methods to the original values, an interval around each masked value is constructed. The width of the interval is based on the rank of the value the variable takes on or its standard deviation. We then examine whether the original value of the variable falls within the interval. The measure of disclosure risk is the proportion of original values that fall into the interval. It is the default disclosure risk in sdcMicro.

4.2.14 Conclusion: This brief report prepared according to the manual (Templ et. al. 2015) gives an idea about the different perturbation techniques which can be implemented in R using package sdcMicro. The techniques described here are mainly applied to continuous data. All computation shown here are performed using R version 3.2.3 (2015-12-10) x86_64-w64-mingw32/x64 (64-bit) using sdcMicro version 4.6.0. It includes all popular disclosure risk and perturbation methods and serves as an easy-to-handle, highly interactive tool for users who want to use the sdcMicro package but are not familiar with the native R command line interface.

References

Brand R (2002). "Microdata Protection Through Noise Addition." In Inference Control in Statistical Databases, Lecture Notes in Computer Science, pp. 97–116. Springer-Verlag.

Domingo-Ferrer J. and J.M. Mateo-Sanz J. M. (2002). "Practical data-oriented microaggregation for statistical disclosure control". IEEE Trans. on Knowledge and Data Engineering, 14(1):189-201.

Gouweleeuw J, Kooiman P, Willenborg L, De Wolf PP (1998). "Post Randomisation for Statistical Disclosure Control: Theory and Implementation." Journal of Official Statistics, 14(4), 463–478.

Mateo-Sanz J., Sebe F., Domingo-Ferrer J. (2004). "Outlier Protection in Continuous Microdata Masking." In Privacy in Statistical Databases, volume 3050 of Lecture Notes in Computer Science, pp. 201–215. Springer-Verlag.

Muralidhar and R. Sarathy R (2006). Data shuffling–a new masking approach for numerical data. Management Science, 52(2):658-670.

Pagliuca D (1999). "Some results of individual ranking method on the system of enterprise accounts annual survey". Esprit SDC Project, Deliverable MI-3/D.

Templ M. and Meindl B. (2008). "Robust statistics meets SDC: New disclosure risk measures for continuous microdata masking". Privacy in Statistical Databases. Lecture Notes in Computer Science. Springer, 5262:113-126, ISBN 978-3-540-87470-6, DOI 10.1007/978-3-540-87471-3 10.

Templ M. (2008). "Statistical disclosure control for microdata using the R-package sdcMicro". Transactions on Data Privacy, 1(2):67-85, URL http://www.tdp.cat/issues/abs.a004a08.php

Templ M., Kowarik A. and Meindl B. (2015). "sdcMicro: Statistical Disclosure Control methods for the generation of public- and scientific-use files. Manual and Package." Vol 67(4). URL http://CRAN.Rproject.org/package=sdcMicro. R package version 4.2.0.

4.3 Comments on the feasibility of using sdcMicro for use in SDC

4.3.1 Applying statistical disclosure control techniques to a micro-data file results in information loss to the user, thereby reducing the utility of the data for research. As potential uses of microdata files are vast, it is simply impossible to undertake an exhaustive assessment of information loss. In many cases the research output has direct relevance to policy making, as such research are undertaken at the behest of Government Departments by third party.

4.3.2 It is acknowledged that loss of information is not desirable. The impact of the perturbation methods on the microdata file may considerably change the estimates from the perturbed data and therefore the magnitude of such estimates may vary from those generated from the unperturbed data. This will create confusion among the users.

4.3.3 A more prudent approach would be to have a realistic framework for data disclosure control rather than a purely restrictive and legalistic approach that will hinder access and optimum use of microdata files.

4.3.4 The use of R-package sdcMicro for SDC does not seem feasible at this point for use in the release of NSS survey results. While there is a need for anonymizing the microdata, data users would like to get data without much loss of information that was originally collected from the respondent. Moreover, due to application of sdcMicro, the loss of information may not be uniform on all units. Moreover, reduction of the chance of identifiability increases the information loss, which can affect the estimates considerably.

Chapter Five: SDC in Ministries/Departments other than Minsitry of Statistics and Programme Implementation

5.1 Status of release and maintenance of survey data in different Ministries

5.1.1 Since the TOR of the Committee extends to evolve and recommend the standards of SDC in respect of surveys of NSSO as well as those of other Government Ministries and Departments, it was decided to collect information on availability of unit level data and other related aspects in some of the Ministries. Specific information considered essential by the Committee for taking stock of the status of release and maintenance of survey data in different Ministries were as follows:

- Survey data collected during the last 10 years
- System of processing and maintaining the data
- Dissemination policy
- Policy practised to protect the confidentiality of the respondents of the survey data.

The following Ministries of the Government of India have informed about the status Reply from four of the following Ministries have been received:

- Ministry of Human Resource Development
- Ministry of Labour and Employment
- Ministry of Health and Family Welfare
- Ministry of Housing & Urban Poverty Alleviation

5.1.2 Information received from these Ministries has been summarized and are placed in Table 5.1

Table 5.1: Status of unit level data and other related aspects in some Ministries

Ministry	Survey Data collected	System of data processing and maintaining of unit level data	Dissemination policy	Protection of the confidentiality of the respondents
(i) Ministry of Human Resource Development	(i) Higher education data are collected directly from Institutions of higher education through a dedicated portal (www.aishe.gov.in). (ii) School education data are collected by National University for Educational Planning (NUEPA) under U-DISE since 2012-13.	-	(i) An annual report is published : All India Survey on Higher Education (ii) Last publication on school education statistics brought out by M/o HRD for 2011-12	-

Table 5.1: Status of unit level data and other related aspects in some Ministries				
Ministry	**Survey Data collected**	**System of data processing and maintaining of unit level data**	**Dissemination policy**	**Protection of the confidentiality of the respondents**
(ii) Ministry of Labour and Employment	**For Socio-Economic Surveys of Different Segments of Labour (SESDSL) component:** (i) SC/ST communities (Valsad ST Belt, Gujarat (2006) (ii) SC/ST communities (KBK ST Belt, Orissa (2008-09) (iii) Unorganised Sector: Leather Industry (2007) (iv) Women Workers in Plantation Industry (2009) and Match Industry (2015) (v) Evaluation of the Minimum Wages Act, 1948: Stone Breaking and Stone Crushing (Rajasthan, 2006/Karnataka/2007-08/Gujarat (2007-08) (vi) Evaluation of the Minimum Wages Act, 1948: Beedi Making Establishments (Madhya Pradesh, 2015) **Others:** (vii) Quarterly Quick Employment Survey in selected labour intensive and export-oriented sectors like textiles, metals, gems & jewellery, automobiles, transport, IT/BPO,	(i) For SESDSL components, data collection and processing are done by the regular staff. (ii) For the Annual Employment-Unemployment Survey and Quarterly Employment Survey data processing, Table Generation is outsourced and Report Writing is done in-house. (iii) From fourth Annual Employment-Unemployment survey the unit level data are maintained by Labour Bureau.	Reports are released. There is no specific policy/practice for dissemination of unit level data. However, data for Fourth Annual Employment-Unemployment Survey to the users is provided to users on demand.	The identity of the respondents are not disclosed.

Table 5.1: Status of unit level data and other related aspects in some Ministries				
Ministry	**Survey Data collected**	**System of data processing and maintaining of unit level data**	**Dissemination policy**	**Protection of the confidentiality of the respondents**
	leather and Handloom/powerloom to assess the economic slowdown on employment in India since 2009. So far 28 such surveys have been conducted. (viii) Annual Employment and Unemployment Surveys. So far 5 such surveys have been conducted.			
(iii) Ministry of Health and Family Welfare	(i) NFHS- 3 (2005-06) (ii) DLHS-3 (2007-08) (iii) Concurrent Evaluation of NRHM (2009-10) (iv) Three rounds of Annual Health Survey (2010-11, 2011-12 and 2012-13) including the CAB Component (2014) (v) DLHS-4 (2012-13) (vi) NFHS-4 (2015-16)	Unit level data is processed and handled by the nodal agency (IIPS) engaged by the Ministry for conducting the survey.	In the recent past, the Ministry took decision to place the unit level data of AHS and DLHS-4 in public domain. As a result the unit level data from 3 round of AHS including CAB[5] and DLHS-4 has been uploaded on HMIS[6] portal/website of Ministry. The unit level data of surveys conducted prior to the above surveys are handled by the nodal agency, i.e., IIPS who provide unit level data to users on demand.	To protect the confidentiality of respondents ID particulars are removed from the unit level data while disseminating survey data in the public domain.
	(i) Slum & Livelihood Survey 2008 (ii) Ministry of HUPA is associated with the conduct of Socio-Economic Caste Census (SECC)-2011 (Urban) carried out by respective State Government/UT	Unit level data is used and processed by the concerned State/UTs.	The data so collected is tabulated and made available to the users in the website.	-

5. CAB: Clinical, Anthropometric & Bio-chemical (CAB) Survey
3. HMIS: Health Management Information System (HMIS)

Table 5.1: Status of unit level data and other related aspects in some Ministries				
Ministry	**Survey Data collected**	**System of data processing and maintaining of unit level data**	**Dissemination policy**	**Protection of the confidentiality of the respondents**
	Administrations. (iii) National Building Organisation collects data through DES of the respective states on • Building Material Prices and Wages of Labour from selected centres spread all over the country on quarterly basis. • Total number of building permits and completion issued by the Municipalities • Total number of Permit Issued for Residential houses.			

5.1.3 The study of the dissemination policies of some of the Ministries/Departments shows that method of data dissemination policies in these Ministries/Departments is not uniform. Some of the Ministries do not release unit level data while others (*Ministry of Labour and Employment, Ministry of Health and Family Welfare*) release unit level data by suppressing the identity of the individuals. But detailed method of suppression of identity of individuals is not clear.

5.2 **Observations of the Committee:** The Committee felt that without Ministry/Department specific details like results/reports published, process of anonymization, data structure of disseminated unit level data, etc., it is not wise to recommend policies for Statistical Disclosure Control (SDC) in a uniform manner. In fact, the Committee strongly felt that Ministry/Department specific SDC policies should be formulated after taking into consideration their requirement. As such it was decided that the Committee will not recommend policies for Statistical Disclosure Control (SDC) of unit level data for other Ministries/Departments.

The committee suggested that as a starting point, SDC policy for dissemination will be developed for different Divisions/Offices within MoSPI. Once the SDC policy of dissemination for MoSPI is standardised and implemented it may further be extended to other Ministries/Departments, if they agree to such policies.

Chapter Six: Recommendations of the Committee

The Committee held comprehensive discussion over the issues of Statistical Disclosure Control (SDC) to protect the identification of the individual units, and release of reliable results. The Indian microdata are very large data sets collected more often from huge populations. The large size of the domain has the natural potential to mask the identity of most respondents. Some of the prescriptions applicable for other developed countries are not fully relevant in our context, where data confidentiality is an evolving concept.

The Committee reviewed the existing system of dissemination policy adopted by MoSPI and other Ministries. Considering all the aspects of TOR and taking note of the concern expressed by NSC, the Committee has made specific recommendations covering the following three aspects:

A. Recommendations relating to protecting the identification of the individual units

6.1 *Need to continue the present practice of data release*: MoSPI has been disseminating data since 1998 and any reversal of the practices followed gets constrained by user expectations. This implies that techniques that lead to information loss and mismatch of survey findings with that already published cannot be adopted. If this is indeed done then it would be necessary to inform the data users and provide them with error margins for their estimates from the original findings reported in the survey reports. Such process is complex and difficult to implement in large scale surveys conducted by NSSO.

6.2 *Need of a trade off between disclosure risk and information loss*: The data disseminating agency has to strike a proper balance between information loss and reduced disclosure risk. Since potential uses of microdata files are vast, it is simply impossible to undertake an exhaustive assessment of information loss. Any data dissemination policy has to strike a balance between data confidentiality and the information necessary for meaningful data analysis. Promoting data sharing is one of the key aspects of the modern day information revolution. Data are usually collected at huge cost and proper sharing of data can reduce overall cost besides reducing respondent burden.

6.3 *Establishment of a data review panel*: The Committee recommends for constituting a *data review panel* consisting of producers and users of data which will function on a continuous basis. This panel will examine the microdata that MoSPI plans to release and if necessary, suggest techniques of SDC before their release to users. The concerned data disseminating agencies should thereafter apply the techniques to ensure compliance with the recommended SDC norms.

6.4 *More stringent clauses in the undertaking*: The Ministry may review the existing undertaking obtained from data users to ensure that the responsibility of maintaining confidentiality of the respondents also lies with the data users.

6.5 *The present system of anonymization is adequate:* The present system of anonymization practised in MoSPI is considered adequate to suppress the identity of the particular respondent. However, possibility of indirectly revealing the identity of the respondent from data specific attributes still remains. The Committee is of the view that complete elimination of such possibilities cannot be built into the disclosure control policies of MoSPI for the time being, in view of the effort required and the possibility of reducing the utility of data. No further anonymization is recommended for the time being.

6.6 *Use of the provision regarding disclosure of data collected from respondent in the Collection of Statistics (COS) Act:* The COS Act has a provision that the disclosure of data is permitted in case the respondent specifically provides that he/she has no objection to that. Statistical Agencies may consider the inclusion of a provision in the survey instruments that the respondent has no objection for the use of the data for statistical purposes.

6.7 *Grouping of data for smaller UTs:* For smaller UTs covered in NSS like Lakshadweep, Daman, Diu it is suggested that the unit level data when disseminated may be grouped into a single unit to avoid possible data disclosure. However, the present practice of publishing separate estimates for each of the States and UTs should continue.

6.8 *Issue on the release of sample list:* There have been several requests for access to information contained in the sample list used for socio-economic surveys. Strictly speaking the sample list does not contain any individual information or information that can reveal the identity of the respondents except where the entire households/establishments in the sample unit are surveyed. As per the current practice, the sample identifiers in the data are scrambled to remove any possibility of linking the unit level data to a specific sample village/block. Location of the sample village/block can be sometimes very important for any research using geographical modelling. NSSO may therefore consider making the information available in the sample list to researchers subject to restrictions/conditions as may be necessary for the specific research project.

6.9 *Adherence to the National Data Sharing and Accessibility Policy-2012 (NDSAP-2012):* The objective of this policy is to facilitate the access to Government of India owned shareable data and information in both human readable and machine readable forms through a network all over the country in a proactive and periodically updatable manner, within the framework of various related policies, Acts and Rules of Government of India, thereby permitting a wider accessibility and use of public data and information.

MoSPI has been disseminating unit level data for all its surveys since the National Data Dissemination Policy was announced in 1999. All the Reports/Key Indicators released on the basis of the surveys undertaken by NSSO are placed in the website of the Ministry (*www.mospi.gov.in*). Besides, MoSPI is also contributing data to the National Data Portal (*www.data.gov.in*).

The Committee is of the view that the present dissemination practice followed in MoSPI is fully compliant with the *NDSAP-2012*.

B. Recommendations relating to release of reliable results

6.10 *Release of reliable estimates*: The study of Relative Standard Error (RSE) undertaken for the specific subjects of survey (*Employment-Unemployment and Household Consumer Expenditure*) shows that the estimates having RSE≤30% and based on sample number of observations > 40 can be accepted for release only in certain situations. Such a rule cannot be applied uniformly across all the Tables in a Tabulation Plan even for a single subject, and far less applicable to other subjects of survey conducted by NSSO. Therefore, there is a need to develop separate rules for different surveys after intense data analysis.

6.11 *Need for caution about the reliability of the estimates*: The Committee is of the view that it is not feasible to formulate a uniform rule based on RSE and sample size to comment on the reliability of the estimates for data dissemination purpose. The Committee recognises the immense use of the released results of NSSO in the form of reports/documents etc. Therefore, the Committee recommends to release all the estimates generated on the basis of the Tabulation Plan, even if such estimates are based on small sample sizes, with an appropriate caution that such estimates are likely to be less reliable.

6.12 *Necessity of abridgement of the Tabulation Plan:* To avoid publishing of less reliable estimates in Reports, based on small sample sizes, necessary abridgement of the Tabulation Plan may be done to reduce the level of disaggregation. Level of disaggregation may be decided using the experience gained from the study of results of similar surveys.

C. Recommendations relating to SDC in other Ministries/Departments

6.13 The Committee observed that inputs received from other Ministries / Departments in respect of results/reports published, process of anonymization, etc., lack detailed information. As such the Committee feels that it is difficult to recommend a uniform policy for Statistical Disclosure Control (SDC) across all Ministries.

6.14 The Committee recommends that SDC policy for dissemination should be developed, standardised and implemented first for different Divisions/Offices within MoSPI. Only, thereafter, the SDC policy can be extended to other Ministries/Departments after due consultation.

Annexure I

No.M-12011/25/2015-NSSO (CPD)
Government of India
Ministry of Statistics & Programme Implementation
National Sample Survey Office (NSSO)
Coordination & Publication Division

Room No. 425 J, Sardar Patel Bhavan,
Sansad Marg, New Delhi-110001
Dated: 22.12.2015

OFFICE MEMORANDUM

Subject: Constitution of Expert Committee to evolve standards for Statistical Disclosure Control for sample surveys conducted by the NSSO

As a follow-up action on the recommendation made by the National Statistical Commission (NSC) in its 77th meeting held during 4th-5th August, 2015, an Expert Committee to evolve standards for Statistical Disclosure Control for sample surveys conducted by the NSSO is constituted as under:

I. Composition:

Non-official:

(1)	Dr. U. C. Sood, Director, IASRI, Delhi	Chairman
(2)	Dr. Manisha Pal, Professor, Department of Statistics, University of Calcutta, Kolkata	Member
(3)	Dr. Sarat Kumar Chettri, Asstt. Professor Department of Computer Science & Engineering and Information Technology, Assam Don Bosco University, Guwahati	Member
(4)	Dr. Diganta Mukherjee, Associate Professor, Indian Statistical Institute, Kolkata	Member
(5)	Shri P. C. Mohanan, Deputy Director General (Retd.), MOS&PI, New Delhi	Member

Official:

(6)	Additional Director General, NSSO (DPD), MoS&PI, Kolkata	Member
(7)	Additional Director General, NSSO (SDRD), MoS&PI, Kolkata	Member Secretary

II. The Terms of Reference (TOR) of the Committee are:

a) To review the existing system of dissemination of micro-data (unit level data) and release of estimates, results, reports and documents based on the sample surveys conducted by the National Sample Survey Office (NSSO) on different socio-economic subjects.

b) To evolve and recommend standards on Statistical Disclosure Control (SDC) in respect of sample surveys for dissemination of unit level data and release of estimates & results at various disaggregated levels by NSSO and other Govt. Ministries/Departments through electronic/digital media, various reports/documents/tables/statements, etc. by taking into consideration reliability of estimates vis-à-vis requirement of the users for detailed data & results, confidentiality of data

and such related aspects.

c) To recommend a suitable mechanism/procedure for dissemination of data & results of sample surveys in line with the National Data Sharing and Accessibility Policy-2012.

2. The tenure of the Committee will be for six months from the date of issue of this Office Memorandum.

3. The Chairman and non-official members would be entitled to a sitting fee of Rs. 1000/- per day for attending meetings. They will also be entitled to transport/transport charges for local travel for attending the meetings of the Committee. Besides, they will be eligible to travel by Air India flight in economy class or by rail in air-conditioned first class while undertaking domestic tours in connection with the meetings of the Committee. The room rent and TA/DA for the outstation meetings will be regulated in accordance with the extant rules of Ministry of Finance from time to time.

4. TA/DA of the official members will be borne by their respective Ministries/ Departments/ Organizations.

5. The expenditure on conducting the meetings of the Expert Committee and on payment/ reimbursement made to the non-official members would be borne by the NSSO (SDRD), Kolkata, Ministry of Statistics and Programme Implementation under the Head 3454-Census, Surveys and Statistics (Major Head), 02-Surveys and Statistics (Sub-Major Head), 02.204-Central Statistical Office (Minor Head), 19-Capacity Development (Capacity Development of CSO and Institutional Development & Capacity Building), 01-Establishment, 19.01.13-Office Expenses under Plan Scheme.

6. NSSO (SDRD) will provide Secretariat support to the Committee and NSSO (DPD) will provide the necessary data and computational assistance to the Committee.

7. This issues with the concurrence of AS&FA vide Dy. No. 1479/AS&FA dated 11.12.2015.

(S. P. Srivastava)
Joint Director
Tel: 23360893

To
Chairman and members of the Expert Committee.

Copy for information to:

1. ADG's of SDRD, DPD and FOD
2. PSO to Secretary (S&PI), Sardar Patel Bhavan, New Delhi
3. Sr. PPS to DG&CEO (NSSO), Sardar Patel Bhavan, New Delhi
4. PPS to DDG, NSC, Sardar Patel Bhavan, New Delhi

23/12/15
(S. P. Srivastava)
Joint Director

Annexure II

Minutes of the first meeting of Expert Committee to evolve standards for Statistical Disclosure Control (SDC) held at Kolkata on 22.02.2016

The first meeting the Expert Committee to evolve standards for Statistical Disclosure Control (SDC) was held at Kolkata on 22.02.2016. At the outset, Sri Satya Narain Singh, ADG, SDRD welcomed all the members to the first meeting of the Expert Committee and requested Sri P. C. Mohanan, Member of the Committee to Chair the meeting, since Prof. U. C. Sud, Chairman of the Expert Committee could not attend this meeting. The meeting was held under the Chairmanship of Sri. P. C. Mohanan.

Sri P. C. Mohanan, outlined the Terms of Reference (TOR) of the Committee. Then he made a presentation on the concepts relating to SDC and its role in survey data.

The following decisions were taken in the meeting:

1. Since the TOR of the Committee extends to evolve and recommend the standards of SDC in respect of surveys of NSSO as well as those of other Government Ministries and Departments, it was decided to collect information on availability of unit level data and other related aspects in some of the Ministries. To start with letters will be addressed to Ministry of Rural Development, Ministry of Labour and Employment, Ministry of Health and Family Welfare, Ministry of Housing and Urban Poverty Alleviation, Ministry of Agriculture & Farmers Welfare, Ministry of Human Resource Development with a request to share information on the following aspects:

    a. Survey data collected during the last 10 years
    b. System of processing and maintaining the data
    c. Dissemination policy
    d. Policy practised to protect the confidentiality of the respondents of the survey data.

2. It was decided Prof. Sarat Kumar Chettri, Member, Expert Committee will look into different aspects of SDC for microdata using the R-package sdcMicro to be discussed in the next meeting.

3. In the context of the presentation made by SDRD for Statistical Disclosure Control (SDC) in sample surveys, both confidentiality and statistical issues in working with survey data were discussed. In general, problems with confidentiality arise when there are small denominators (population size represented in a specific cell in a table); and, problems with data reliability arise when there are small numerators (cases in a specific cell in a table). Thus, it is desirable to have rules for privacy protection which consider both denominator size and numerator size. Rules to address statistical reliability can be limited to consideration of numerator size.

4. The presentation made by SDRD for SDC with respect to Employment and Unemployment Survey and Consumer Expenditure Survey of NSSO, mainly addressed the issues of data reliability by studying the frequency of sample observation occurring in the numerators for a specific cell in a table along with the Relative Standard Errors (RSE). It was decided that further analysis will be done considering the denominator rule which focuses on the fact that the risk of violating confidentiality increases substantially when data are tabulated for small subgroups of the population within small geographic areas.

5. It was also decided that the study undertaken by SDRD and presented in the meeting will be extended to have an idea about the reliability of the results for other surveys of NSS, especially for the estimates of those characteristics which are rare in the population.

The meeting ended with vote of thanks to the Chair.

Annexure III

Minutes of the Second meeting of the Expert Committee to Evolve Standards for Statistical Disclosure Control (SDC) for Sample Surveys Conducted by the NSSO held at Mahalanobis Bhawan, Kolkata on 19.08.2016

The second meeting of SDC was held at Mahalanobis Bhavan, Kolkata on 19.08.2016 under the Chairmanship of Prof. U. C. Sud. At the outset Shri K.P. Unnikrishnan, ADG, SDRD welcomed all the members. Shri D. Mukhopadhyay, DDG, SDRD briefly introduced the agenda points for discussion in the meeting. He also outlined the decisions taken in the first meeting of SDC.

The list of participants in the meeting is given in Annexure.

In this meeting detailed discussion was held on the following aspects:

- Data collection and dissemination policies of different Ministries/Departments.
- The findings of the study of SDRD on *Analysis of Relative Standard Error (RSE) of the estimates of Employment and Unemployment parameters for the purpose of SDC based on sample size in the numerator and denominator.*
- The presentation made by Prof. Sarat Kumar Chettri, Member on *Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro.*
- Methods of data dissemination by CSO, IS Wing and Directorate General of Commercial intelligence and Statistics (DGCIS)

The following decisions were taken in the Meeting:

1. The study of the dissemination policies as practised by by some of the Ministries/Departments shows that method of data dissemination policies in these Ministries/Departments are not uniform. Some of the Ministries do not release unit level data while others (*Ministry of Labour and Employment, Ministry of Health and Family Welfare*) release unit level data by suppressing the identity of the individuals. But detailed method of suppression of identity of individuals is not clear. The communications received from the Ministries/Departments were not adequate to prescribe policies for Statistical Disclosure Control (SDC) for different Ministries/Departments. It was felt that separate policies for such Minsitries/Departments will be needed by close interaction with them.

2. 2. It was opined by the members that a data dissemination policy conforming to the SDC may be developed for MoSPI so that various aspects and consequences of SDC are fully understood and analysed before such a policy is extended to other Ministries/Departments outside MoSPI.

3. The Committee examined and analysed some important tables of Employment and Unemployment survey and Consumer Expenditure survey of NSS 68th (2011-12) round. The Committee observed that a large number of cells in these studied Tables had very small number of observations (number of observations 1 or 0). The Committee felt generation of estimates based on very small observations may not be appropriate. As such there is a need to avoid such situations by reformulating

Tabulation Plans after taking care of specific requirements from different subjects of NSSO. The Committee suggested that for a set of key indicators Tables at the all India level may be prepared at a more detailed format whereas Tabulation at the State/ UT level may be prepared at a broad level of disaggregation.

4. The study of Relative Standard Error (RSE) shows that the estimates having RSE≤30% and based on sample number of observations > 40 can be accepted for release. However, such a rule cannot be applied uniformly across all the Tables in a Tabulation Plan on a subject of survey as well as across all the subjects of survey conducted by NSSO. Therefore the Committee suggested that subject-wise separate rules may be developed after thorough analysis of data and considering the requirements of data users.

5. Further the Committee observed that the tabulation plan provides for estimates of rare events such as unemployment rate in different age groups. The theoretical discussion of rare events prove that RSE of such estimates will always be high. Therefore, RSE of the estimate may not be the only factor for taking a decision on release of estimates and for Statistical Disclosure Control purpose a rule based on RSE may not be appropriate.

6. The Committee appreciated the highly informative presentation made by Prof. Sarat Kumar Chettri, on *Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro*. The R package sdcMicro serves as an easy implementation of SDC methods to evaluate and anonymize confidential micro-data sets. It includes all popular disclosure risk and perturbation methods. All methods are highly optimized in terms of computational costs to be able to work with large data sets. Reporting facilities that summarize the anonymization process can also be easily used by practitioners. Since the procedure adopted in the package may change some records of the unit level data, the modified data will not generate same estimates as the unit level data due to loss of infomation. The Committee felt that such loss of information is not desirable and the loss of information may also not be uniform across all units. Also, reduction of the chance of identifiability increases the information loss, which can effect the estimates considerably.

7. The members of the committee of SDC were requested to send their inputs for preparing a draft report on the recommendation of the committee regarding Statistical Disclosure Control in NSS surveys for discussion in the next meeting of SDC.

8. The next meeting will be SDC will be held in first week of November, 2016. In the next meeting, the draft SDC report based on the inputs and comments from the members will be placed for discussion.

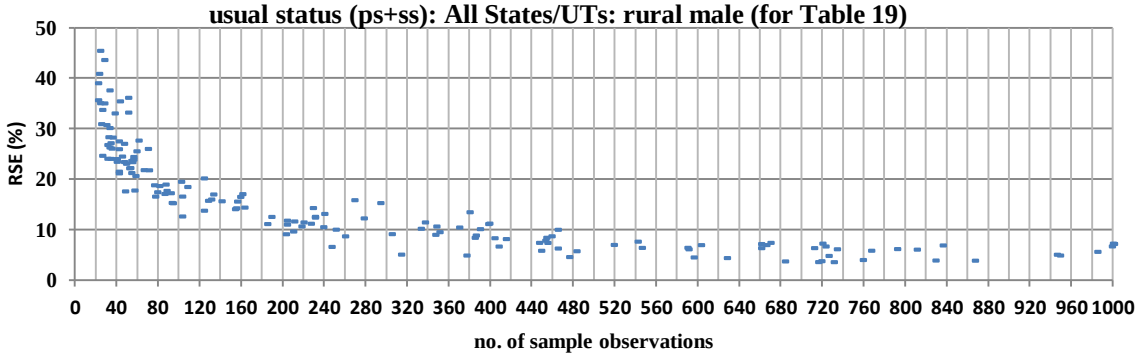
The meeting ended with a vote of thanks to the Chair.

Table (19): Per 1000 distribution of persons by usual status (ps+ss) for each age-group

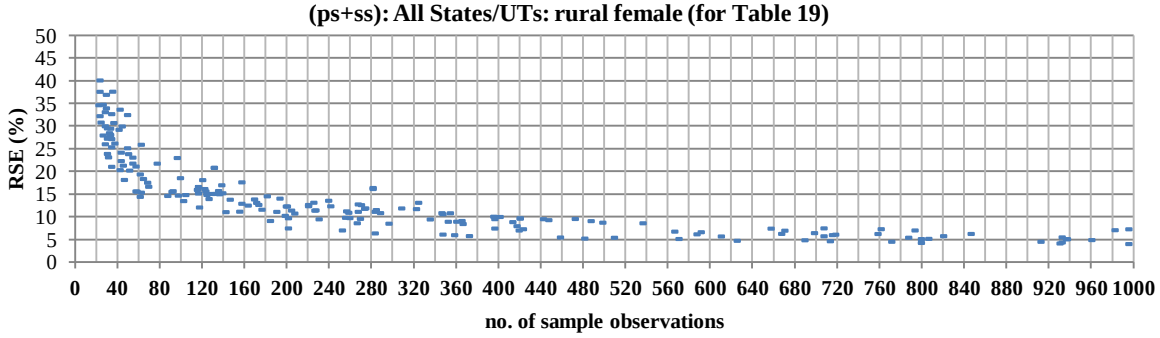
state/u.t./all-India estimate/ RSE rural/urban/ rural+urban male/female/persons

usual activity (ps+ss)		age group (years)																	persons	
		0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60 -64	65 & above	15 & above	n.r.	all	estd. (00)	sam-ple
status	industry																			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)
11,12,21	01-03																			
	05-43																			
	45-96																			
	01-96																			
31	01-03																			
	05-43																			
	45-99																			
	01-99																			
41	x																			
51	01-03																			
	05-43																			
	45-99																			
	01-99																			
41 & 51	01-99																			
11-51	01-03																			
	05-43																			
	45-99																			
	01-99																			
81	x																			
11-81	x																			
91	x																			
92	x																			
93	x																			
94	x																			
95	x																			
97	x																			
99	x																			
91-99	x																			
11-99	x	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000		
estd. persons (00)																			x	x
sample persons																			x	x

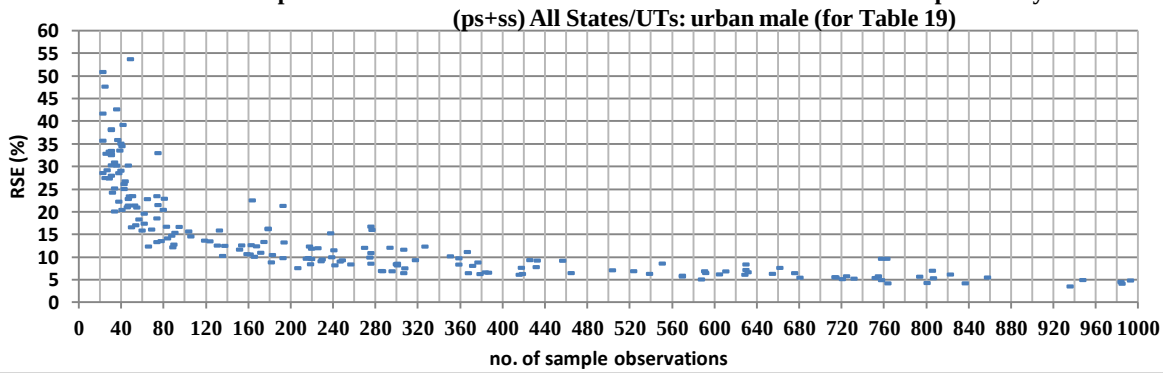
Graph 1: RSE and no. of observations of the cells for distribution of persons by usual status (ps+ss): All States/UTs: rural male (for Table 19)



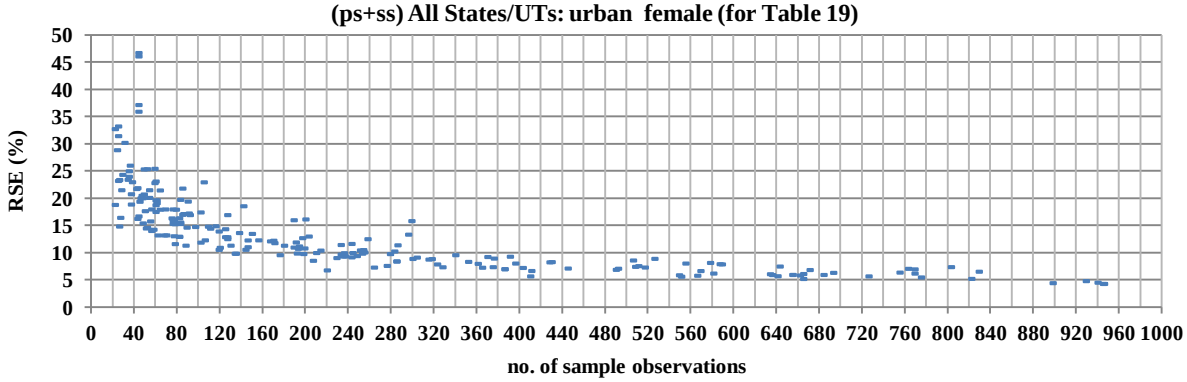
Graph 2: RSE and no. of observations of the cells for distribution of persons by usual status (ps+ss): All States/UTs: rural female (for Table 19)



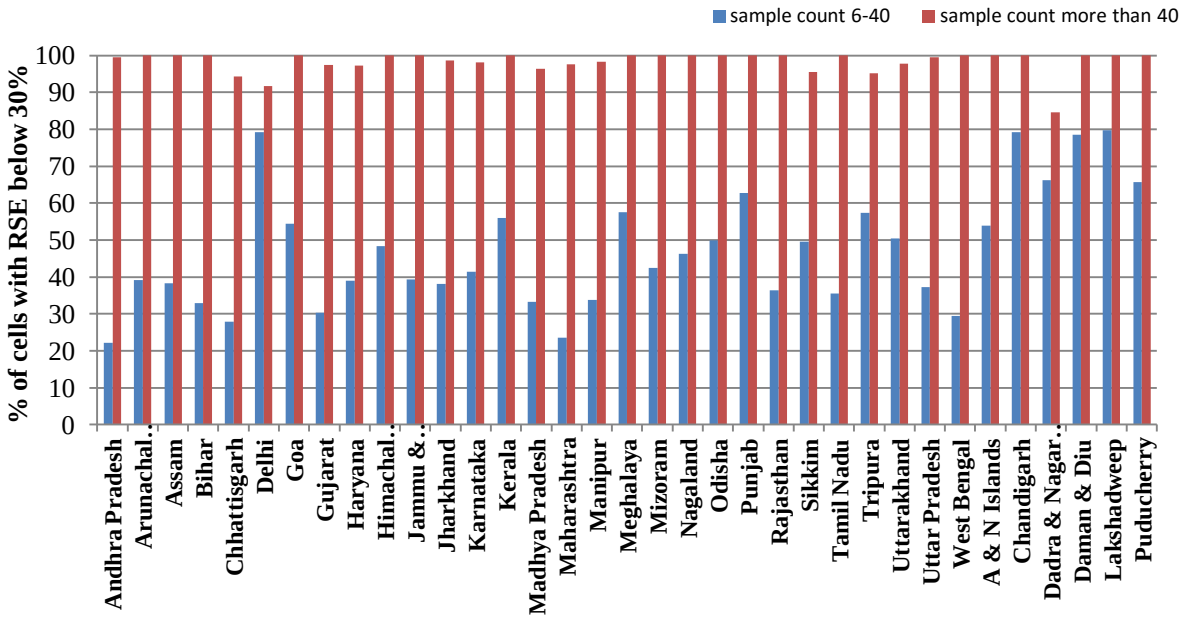
Graph 3: RSE and no. of observations of the cells for distribution of persons by usual status (ps+ss) All States/UTs: urban male (for Table 19)



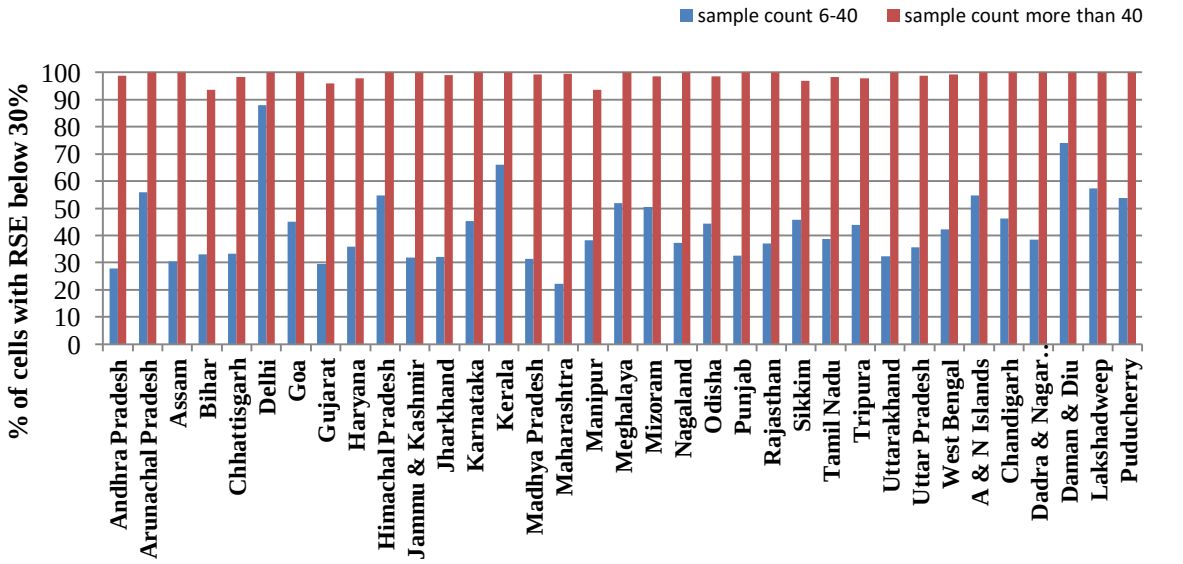
Graph 4: RSE and no. of observations of the cells for distribution of persons by usual status (ps+ss) All States/UTs: urban female (for Table 19)

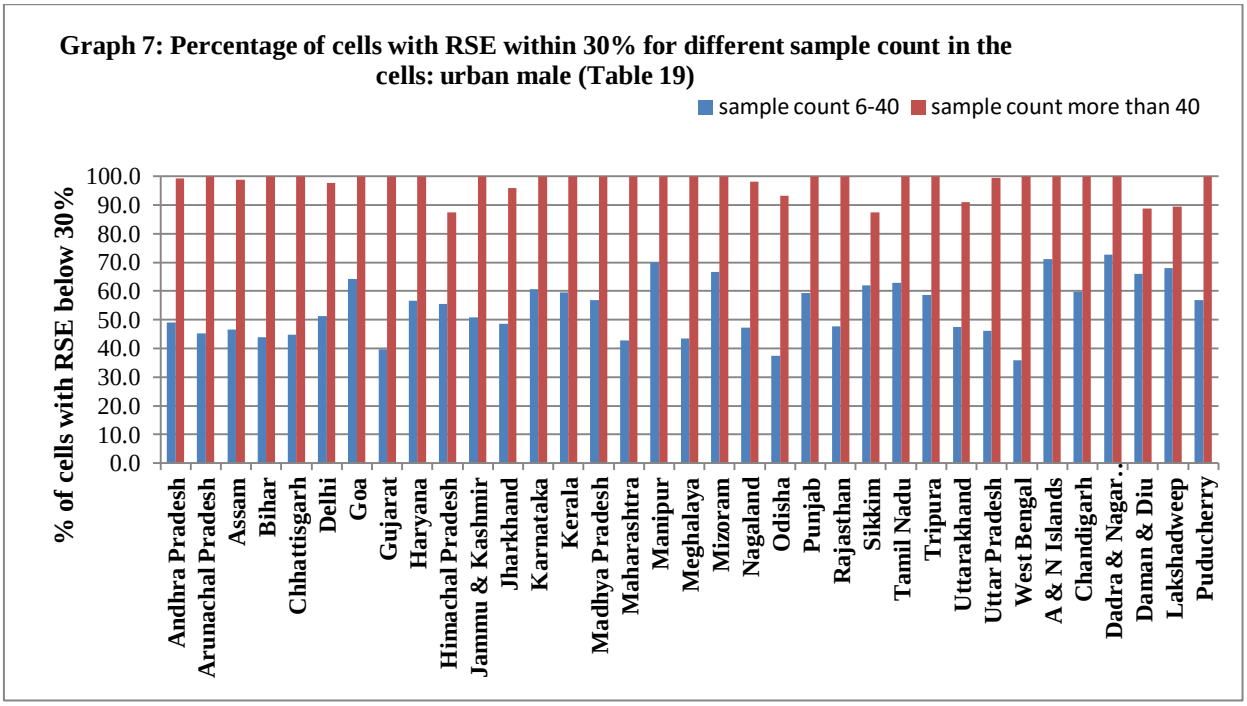


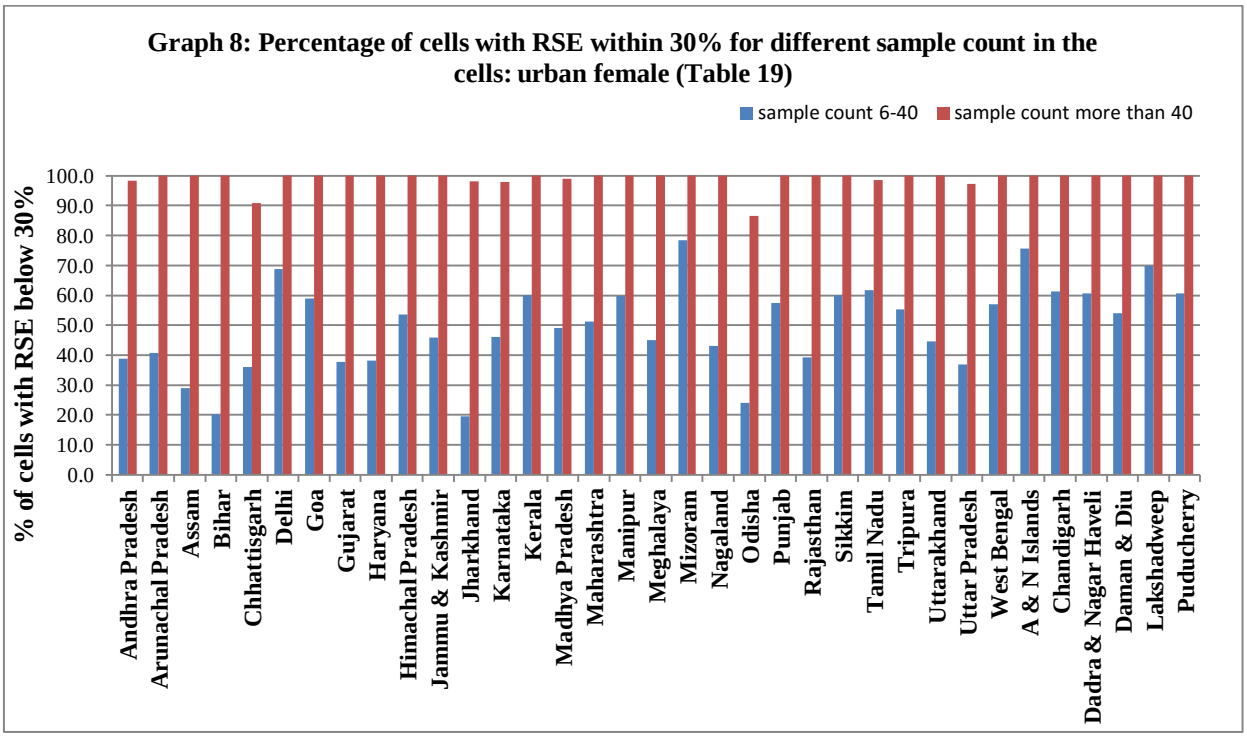
Graph 5: Percentage of cells with RSE within 30% for different sample count in the cells: rural male (Table 19)



Graph 6: Percentage of cells with RSE within 30% for different sample count in the cells: rural female (Table 19)




Graph 7: Percentage of cells with RSE within 30% for different sample count in the cells: urban male (Table 19)
sample count 6-40
sample count more than 40
% of cells with RSE below 30%
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chhattisgarh
Delhi
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu & Kashmir
Jharkhand
Karnataka
Kerala
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Punjab
Rajasthan
Sikkim
Tamil Nadu
Tripura
Uttarakhand
Uttar Pradesh
West Bengal
A & N Islands
Chandigarh
Dadra & Nagar..
Daman & Diu
Lakshadweep
Puducherry


Graph 8: Percentage of cells with RSE within 30% for different sample count in the cells: urban female (Table 19)
sample count 6-40
sample count more than 40
% of cells with RSE below 30%
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chhattisgarh
Delhi
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu & Kashmir
Jharkhand
Karnataka
Kerala
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Punjab
Rajasthan
Sikkim
Tamil Nadu
Tripura
Uttarakhand
Uttar Pradesh
West Bengal
A & N Islands
Chandigarh
Dadra & Nagar Haveli
Daman & Diu
Lakshadweep
Puducherry

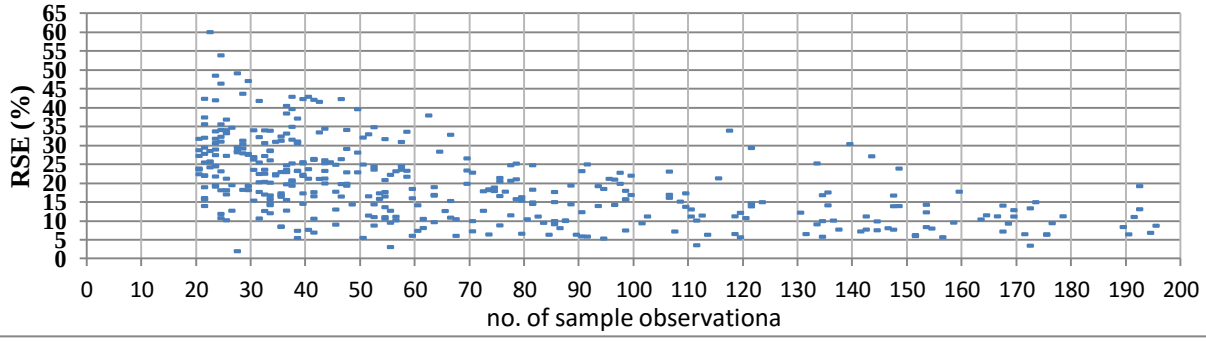
Annexure VII

Table (44) : Average wage/salary earnings (Rs. 0.00) per day received by regular wage/salaried employees (31,71 & 72) of age 15-59 years by industry of work and broad education category

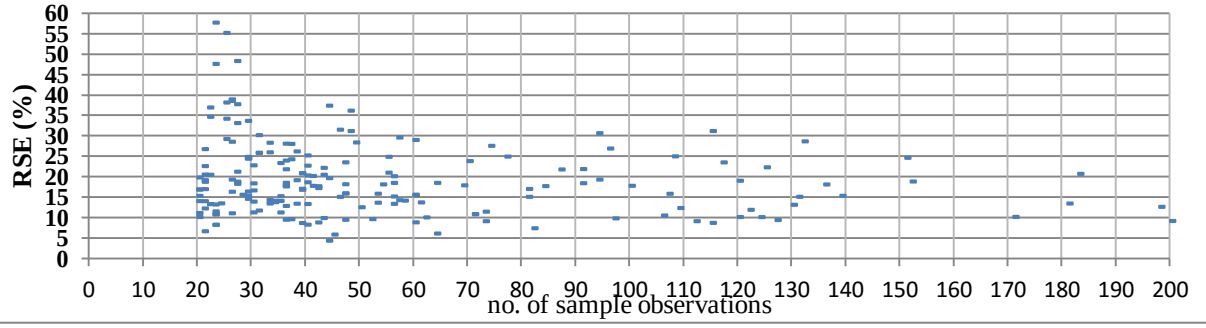
state/u.t./ all-India rural/urban/ rural+urban male/female/persons

industry of work (industry division)	education category						person-days	
	not literate (01)	literate &upto middle (02-07)	secondary & higher secondary (08 & 10)	diploma / certificate (11)	graduate & above (12, 13)	all (incl. n.r.)	estd. (00)	sample
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
agriculture (01-03)								
mining & quarrying (05-09)								
manufacturing1 (10-18)								
manufacturing2 (19-33)								
electricity, gas and water (35-39)								
construction (41-43)								
trade (45,46,47,55,56)								
transport& storage etc. (49-53)								
services (58-96)								
privatehhs. with emp. persons (97)								
others (99)								
all								
estimated person-days (00)								
sample person-days								

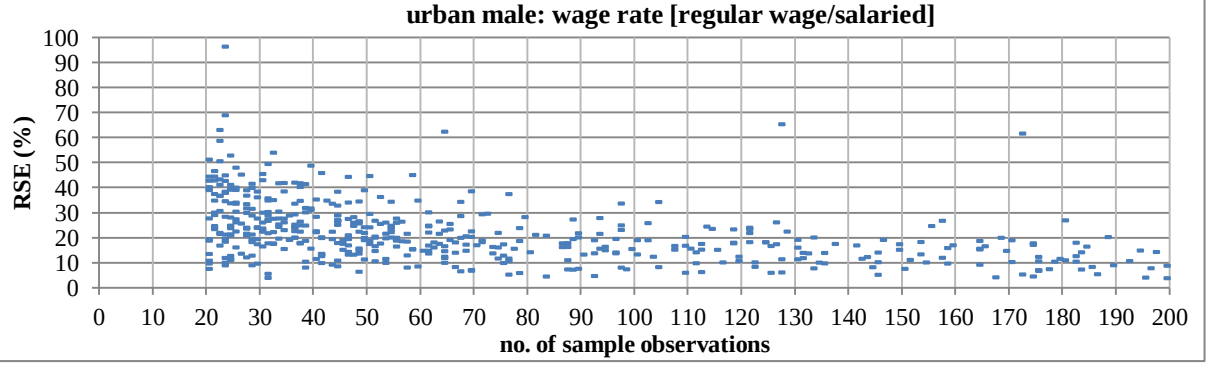
Graph 9: RSE and no. of observations in the cells for wage rate per day: All-India: rural male: wage rate [regular wage/salaried]



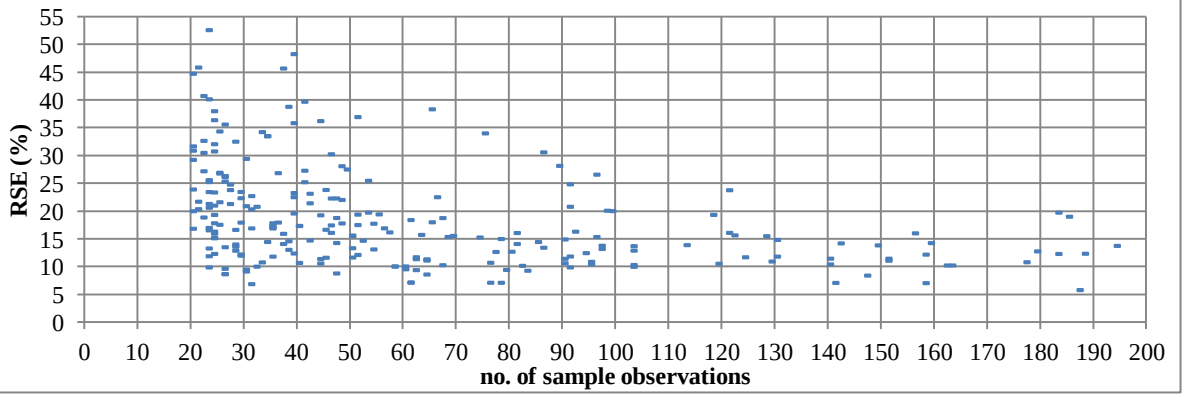
Graph 10: RSE and no. of observations in the cells for wage rate per day: All-India: rural female: wage rate [regular wage/salaried]

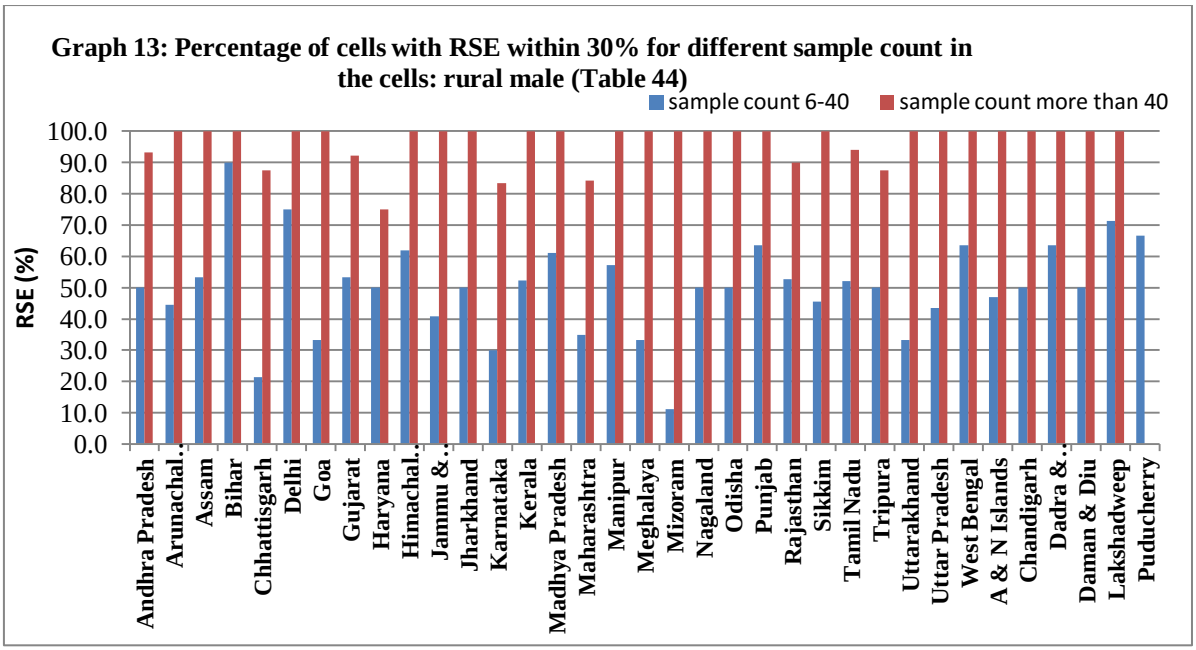


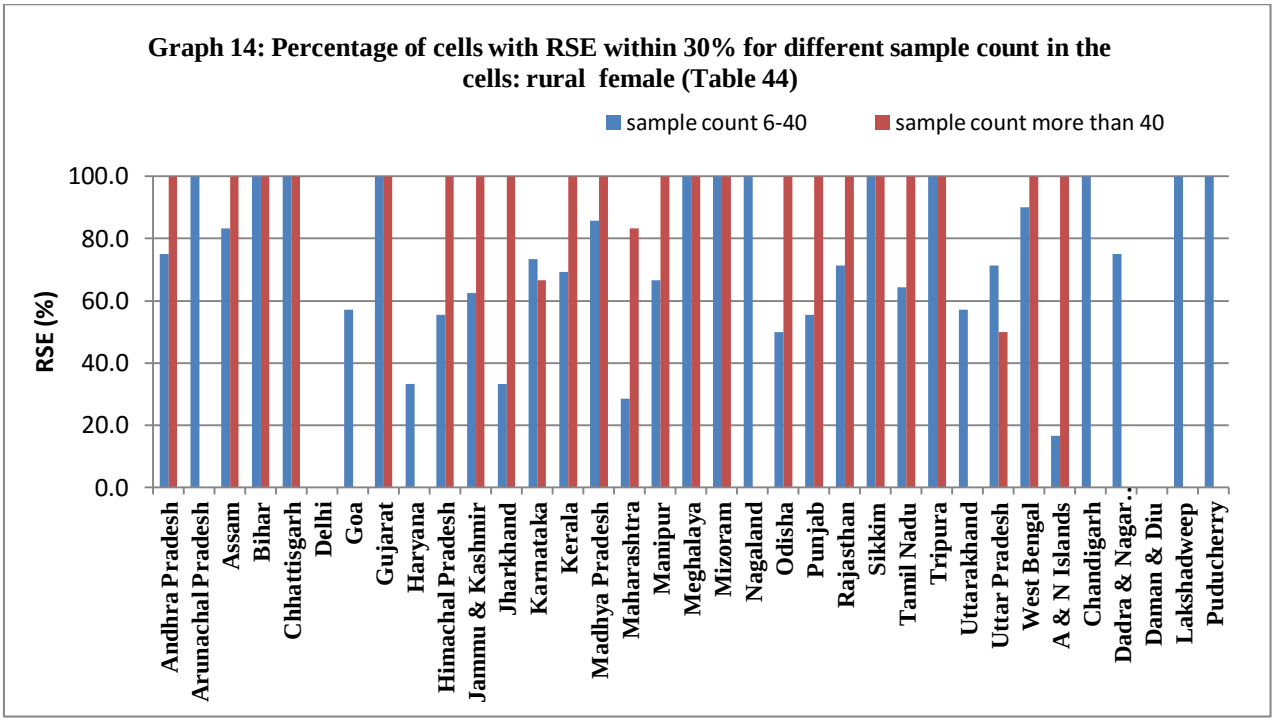
Graph 11: RSE and no. of observations in the cells for wage rate per day: All-India: urban male: wage rate [regular wage/salaried]

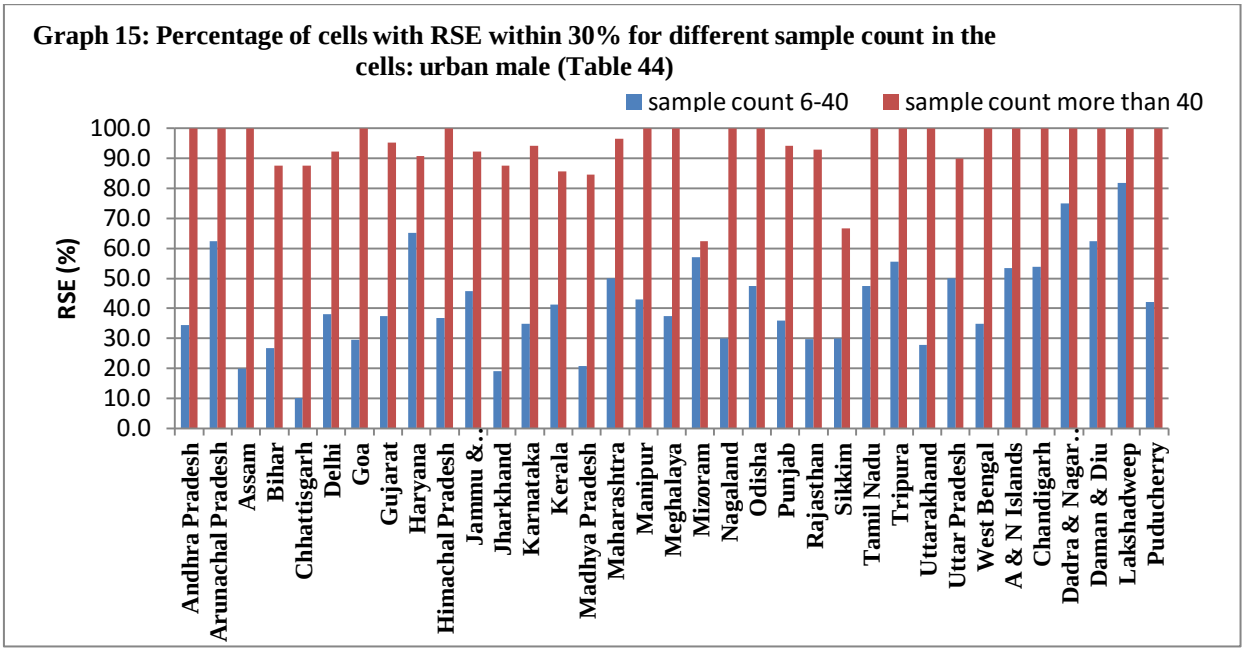


Graph 12: RSE and no. of observations in the cells for wage rate per day: All-India: urban female: wage rate [regular wage/salaried]




Graph 13: Percentage of cells with RSE within 30% for different sample count in the cells: rural male (Table 44)
sample count 6-40
sample count more than 40
RSE (%)
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
Andhra Pradesh
Arunachal..
Assam
Bihar
Chhattisgarh
Delhi
Goa
Gujarat
Haryana
Himachal..
Jammu &..
Jharkhand
Karnataka
Kerala
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Punjab
Rajasthan
Sikkim
Tamil Nadu
Tripura
Uttarakhand
Uttar Pradesh
West Bengal
A & N Islands
Chandigarh
Dadra &..
Daman & Diu
Lakshadweep
Puducherry


Graph 14: Percentage of cells with RSE within 30% for different sample count in the cells: rural female (Table 44)
sample count 6-40
sample count more than 40
RSE (%)
100.0
80.0
60.0
40.0
20.0
0.0
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chhattisgarh
Delhi
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu & Kashmir
Jharkhand
Karnataka
Kerala
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Punjab
Rajasthan
Sikkim
Tamil Nadu
Tripura
Uttarakhand
Uttar Pradesh
West Bengal
A & N Islands
Chandigarh
Dadra & Nagar..
Daman & Diu
Lakshadweep
Puducherry


Graph 15: Percentage of cells with RSE within 30% for different sample count in the cells: urban male (Table 44)
sample count 6-40
sample count more than 40
RSE (%)
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chhattisgarh
Delhi
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu &..
Jharkhand
Karnataka
Kerala
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Punjab
Rajasthan
Sikkim
Tamil Nadu
Tripura
Uttarakhand
Uttar Pradesh
West Bengal
A & N Islands
Chandigarh
Dadra & Nagar..
Daman & Diu
Lakshadweep
Puducherry

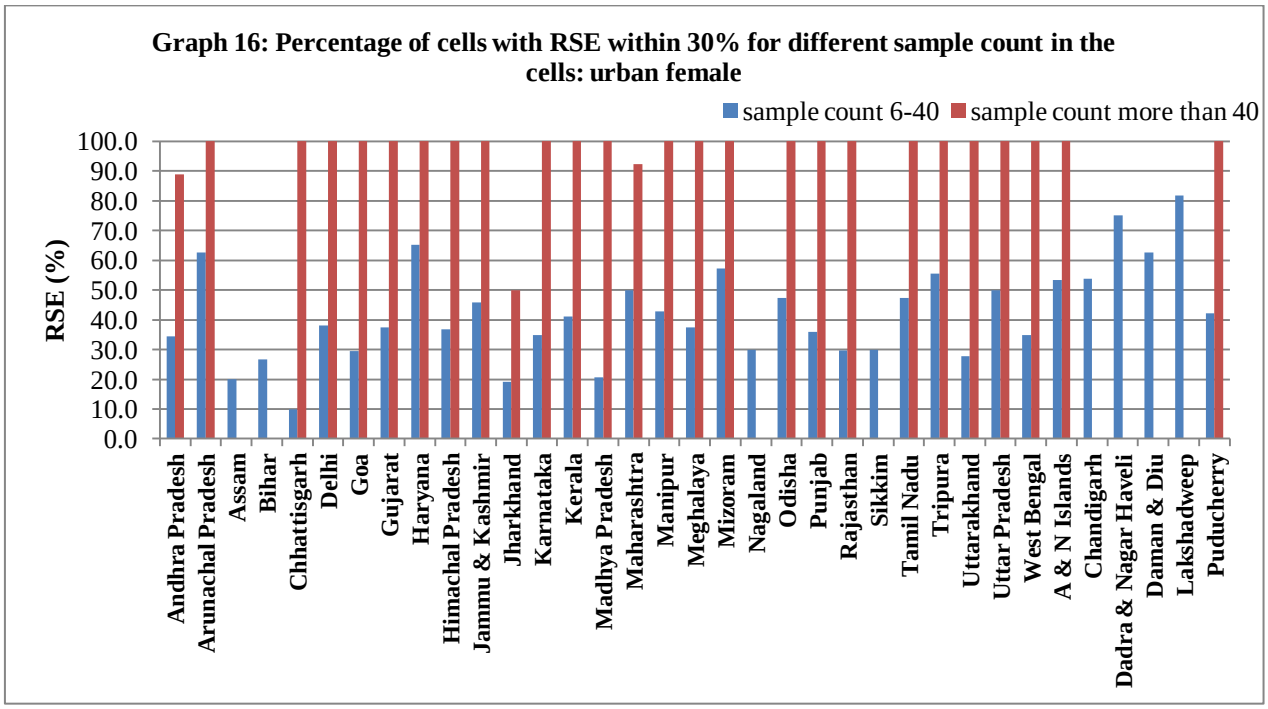

Graph 16: Percentage of cells with RSE within 30% for different sample count in the cells: urban female
sample count 6-40
sample count more than 40
RSE (%)
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chhattisgarh
Delhi
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu & Kashmir
Jharkhand
Karnataka
Kerala
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Punjab
Rajasthan
Sikkim
Tamil Nadu
Tripura
Uttarakhand
Uttar Pradesh
West Bengal
A & N Islands
Chandigarh
Dadra & Nagar Haveli
Daman & Diu
Lakshadweep
Puducherry

Table 8.2: Monthly per capita quantity and value of consumption at detailed item level for all items of consumption

State/UT/all-India Sub-sample: 1/2/all Fractile class of MPCE**: Rural/Urban

code	item	quantity*	value (Rs.[0.00])	no. per 1000 hhs reporting consumption	no. of sample hhs reporting consumption
(1)	(2)	(3)	(4)	(5)	(6)
101	rice- pds				
102	rice-other sources				
103	chira				
……					
129	**cereal - subtotal**				
……					
……					
299	**packaged processed food - subtotal**				
777	**food: total**				
300					
……					
309	**pan – subtotal**				
…..					
……					
643	other ornaments				
649	jewellery and ornaments – subtotal				
659	durable goods – total				
888	**non-food: total**				
999	**total (777+888)**				

Annexure XI

Table (19RSE): RSE of the per 1000 distribution of persons by usual activity category taking also into consideration the subsidiary economic status of persons categorized 'not working' in the principal status for each age-group

(NSS 68th round (July 2011- June 2012), Schedule 10)

State: all-India Sector: ALL Sex: Person

usual activity		RSE (%) of per 1000 distribution of persons by age group																
		Age groups																
(ps+ss)		0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60 -64	65 & above	15 & above	all	estd
status	Industry																	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
11,12,21	01-03		76.12	11.08	4.03	2.74	2.55	2.29	2.12	2.15	2.16	2.43	2.62	2.68	2.88	1.06	1.08	1339338
	05-43		59.06	14.86	7.14	5.27	4.55	4.08	3.86	4.32	4.15	5.00	5.95	6.61	7.96	2.25	2.27	310674
	45-96		57.54	18.22	6.49	3.68	2.84	3.01	2.61	2.56	2.77	3.34	3.89	4.51	4.64	1.31	1.31	545736
	01-96		40.56	8.27	3.14	1.99	1.70	1.52	1.38	1.44	1.47	1.77	1.92	2.08	2.31	0.70	0.71	2195747
31	01-03			137.98	23.73	24.99	19.69	18.58	18.21	22.57	17.71	23.76	29.10	39.18	36.73	10.40	10.42	17208
	05-43		124.46	53.09	8.55	7.54	5.74	4.61	4.41	5.23	6.49	6.37	6.78	14.64	19.81	2.93	3.01	228697
	45-99		93.38	19.99	6.56	3.50	2.77	2.81	2.71	2.87	3.15	3.33	3.87	9.30	10.69	1.29	1.31	504681
	01-99		75.09	42.94	5.43	3.66	2.57	2.27	2.26	2.49	2.80	2.97	3.25	7.65	9.35	1.25	1.29	750587
41	x			73.11	19.52	11.29	10.23	10.17	8.13	9.50	10.77	11.97	13.37	14.15	15.69	4.18	4.15	63453
51	01-03		96.38	15.01	5.72	4.32	3.68	3.39	3.19	3.36	3.59	4.16	4.51	4.97	6.41	1.69	1.69	694525
	05-43			22.88	5.04	3.45	3.27	3.13	3.36	3.83	3.99	4.91	5.63	7.16	9.20	1.80	1.79	422021
	45-99		94.29	37.21	10.94	7.94	7.40	7.68	6.67	8.74	9.36	11.28	11.94	12.74	17.02	3.66	3.61	78177
	01-99		71.98	12.23	3.74	2.56	2.33	2.29	2.19	2.50	2.63	3.11	3.53	4.12	5.21	1.16	1.15	1194723
41 & 51	01-99		71.98	12.16	3.67	2.47	2.26	2.23	2.08	2.37	2.53	2.98	3.40	3.89	4.96	1.11	1.10	1258176
11-51	01-03		68.71	8.83	3.17	2.09	1.81	1.57	1.51	1.49	1.50	1.71	1.89	2.09	2.52	0.72	0.74	2056013
	05-43		57.62	15.96	3.69	2.75	2.35	2.12	2.05	2.30	2.54	2.93	3.35	4.33	5.49	1.17	1.21	1019903
	45-99		44.60	14.32	4.30	2.35	1.92	2.00	1.77	1.82	1.95	2.31	2.64	3.87	4.08	0.85	0.86	1128594
	01-99		36.17	7.24	2.06	1.07	0.75	0.69	0.66	0.64	0.71	0.89	1.10	1.47	1.92	0.30	0.33	4204511
81	x		36.41	21.52	6.40	4.08	5.48	7.70	13.23	21.54	24.58	29.16	53.86	42.77	38.43	3.22	3.25	94080
11-81	x		32.18	6.87	1.92	0.97	0.71	0.68	0.66	0.64	0.70	0.89	1.10	1.47	1.92	0.29	0.32	4298591
91	x		0.41	0.36	0.84	2.21	6.09	22.90	41.66	62.02	145.01	36.87	50.13	35.77	31.86	1.06	0.51	2992687

Table (19RSE): RSE of the per 1000 distribution of persons by usual activity category taking also into consideration the subsidiary economic status of persons categorized 'not working' in the principal status for each age-group

(NSS 68th round (July 2011- June 2012), Schedule 10)

State: all-India Sector: ALL Sex: Person

usual activity		RSE (%) of per 1000 distribution of persons by age group																
		Age groups																
(ps+ss)		0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60 -64	65 & above	15 & above	all	estd
status	Industry																	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
92	x		26.21	8.65	3.56	2.27	2.03	2.03	2.22	2.33	2.42	2.52	2.66	2.77	2.76	0.82	0.84	1176137
93	x		31.72	10.50	3.99	2.56	2.28	2.42	2.51	2.69	3.22	3.63	3.87	4.56	4.68	0.98	0.99	871795
94	x		42.76	60.54	48.50	38.35	27.80	20.96	21.07	16.37	13.54	9.58	6.67	3.47	2.60	2.39	2.42	157465
95	x		16.92	21.90	12.83	12.53	14.39	14.59	14.76	18.51	13.45	9.76	8.81	6.57	3.40	2.75	2.71	128115
97	x		3.08	7.59	8.48	12.28	22.69	27.24	33.69	20.75	17.96	11.92	8.80	5.28	2.36	2.30	2.06	346129
99	x	0.00															1.04	911638
91-99	x	0.00	0.02	0.21	0.60	1.01	1.26	1.37	1.50	1.57	1.74	1.84	1.78	1.47	0.81	0.37	0.21	6583966
11-99	x	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10882557

Table (44RSE) : RSE of average wage/salary earnings (Rs. 0.00) per day received by regular wage/salaried employees (31,71 & 72) of age 15-59 years by industry of work and broad education category

State: all-India Sex: Person Sector: ALL

		RSE (%) of average wage/salary earnings (Rs. 0.00) per day					
		education category					
sector of work	Industry	not literate	literate & upto Middle	secondary & higher secondary	diploma / certificate	graduate & above	all (incl. n. r.)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(17)
Agriculture	(01-03)	9.26	11.00	18.97	14.47	33.23	11.10
mining & quarrying	(05-09)	21.02	18.05	16.67	34.40	37.22	23.46
manufacturing1	(10-18)	10.28	6.07	5.67	15.92	10.60	4.46
manufacturing2	(19-33)	8.59	7.33	7.03	9.07	10.61	6.03
electricity, gas & water	(35-39)	15.04	13.38	13.48	15.89	16.42	9.83
Construction	(41-43)	21.90	11.71	12.09	19.46	18.99	8.22
Trade	(45,46,47,55,56)	10.64	5.52	5.17	20.40	9.28	4.11
transport & storage etc.	(49-53)	7.54	7.38	5.80	16.25	11.26	4.99
Services	(58-96)	8.07	4.20	2.92	5.60	2.14	2.16
private hhs. with emp. persons	(97)	10.46	10.50	18.78	20.98	26.33	8.14
Others	(99)		103.42			25.50	54.00
All		2.68	1.69	1.48	2.91	2.16	1.57

Table 1RSE: RSE of average MPCE for each fractile class of MPCE(MMRP)

Sector: Rural

State/UT		RSE (%) of average MPCE (Rs. 0.00) for each fractile class											
	1	2	3	4	5	6	7	8	9	10	11	12	All
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Andhra Pradesh	1.94	0.57	0.36	0.31	0.34	0.30	0.23	0.36	0.34	0.58	0.72	7.28	1.95
Arunachal Pradesh	2.38	0.50	1.30	0.74	0.59	0.71	0.66	0.81	0.92	1.12	0.86	3.95	4.72
Assam	2.12	0.47	0.52	0.33	0.24	0.33	0.28	0.37	0.41	0.59	0.45	4.88	1.87
Bihar	1.80	0.48	0.53	0.39	0.27	0.22	0.26	0.28	0.40	0.53	0.62	1.71	1.54
Chhattisgarh	5.06	0.70	0.53	0.60	0.47	0.41	0.24	0.45	0.79	0.52	0.85	4.56	1.63
Delhi	0.02		1.70	0.00	1.33	0.11	0.12	1.63	2.37	0.00	0.00	18.64	17.01
Goa	4.14	1.45	0.54	1.68	1.80	0.61	0.52	0.31	0.64	2.04	2.64	15.80	6.63
Gujarat	4.56	1.45	0.62	0.43	0.44	0.45	0.35	0.52	0.36	0.67	1.17	3.70	1.99
Haryana	3.00	0.75	0.64	0.51	0.42	0.35	0.76	0.52	0.60	0.61	1.91	5.32	2.37
Himachal Pradesh	1.44	0.60	0.60	0.41	0.36	0.44	0.36	0.35	0.35	0.57	0.85	7.95	2.63
Jammu & Kashmir	3.00	0.94	0.52	0.30	0.32	0.19	0.26	0.31	0.33	0.55	0.63	3.48	1.58
Jharkhand	4.76	0.72	0.67	0.39	0.25	0.33	0.27	0.31	0.47	0.58	0.49	4.92	1.72
Karnataka	1.33	0.89	0.57	0.42	0.34	0.29	0.37	0.28	0.55	0.73	0.83	10.02	2.72
Kerala	1.52	0.42	0.44	0.42	0.34	0.35	0.43	0.27	0.42	0.54	0.88	13.83	4.12
Madhya Pradesh	2.23	0.57	0.59	0.33	0.31	0.34	0.30	0.30	0.44	0.58	0.65	4.24	1.94
Maharashtra	3.11	0.60	0.37	0.33	0.22	0.27	0.30	0.31	0.36	0.50	0.60	9.40	2.14
Manipur	1.07	0.87	0.44	0.73	0.32	0.32	0.33	0.48	0.29	0.69	0.64	1.42	2.11
Meghalaya	3.81	1.17	0.61	0.43	0.40	0.26	0.30	0.51	0.41	0.44	0.42	2.13	2.80
Mizoram	6.46	1.03	1.39	1.01	0.57	0.79	0.95	0.83	0.73	0.40	0.69	3.05	3.28
Nagaland	1.86	1.51	0.95	0.49	0.29	0.56	0.41	0.53	0.32	0.39	0.93	3.64	2.18
Odisha	1.48	0.50	0.42	0.43	0.25	0.32	0.32	0.24	0.34	0.48	0.65	2.77	1.33
Punjab	1.85	0.58	0.67	0.49	0.46	0.34	0.36	0.36	0.36	0.73	0.72	3.10	2.24
Rajasthan	2.32	0.48	0.48	0.44	0.31	0.31	0.31	0.34	0.38	0.47	0.68	**19.64**	2.95
Sikkim	1.70	1.44	0.78	0.68	0.55	0.45	0.64	0.77	0.53	1.12	0.73	2.06	2.99
Tamil Nadu	2.13	0.93	0.72	0.47	0.42	0.28	0.33	0.30	0.47	0.68	0.73	2.50	1.61
Tripura	1.51	0.57	0.62	0.26	0.36	0.36	0.22	0.60	0.47	0.55	0.66	2.41	3.30
Uttarakhand	1.83	0.75	0.55	0.47	0.31	0.31	0.59	0.50	0.76	0.98	0.89	7.26	3.71
Uttar Pradesh	1.46	0.33	0.32	0.28	0.21	0.21	0.19	0.21	0.28	0.42	0.47	3.20	1.26

Table 1RSE: RSE of average MPCE for each fractile class of MPCE(MMRP)

Sector: Rural

State/UT		RSE (%) of average MPCE (Rs. 0.00) for each fractile class											
	1	2	3	4	5	6	7	8	9	10	11	12	All
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
West Bengal	3.41	0.53	0.36	0.29	0.25	0.24	0.28	0.36	0.38	0.61	0.61	3.82	1.54
A & N Islands	3.84	1.76	0.83	1.11	1.13	0.98	1.38	0.34	1.09	1.51	1.61	14.08	4.14
Chandigarh	5.95		1.38	2.53	0.93	0.14	0.71	2.04	1.41	7.12	2.23	5.23	7.13
Dadra & Nagar Haveli	0.84	1.33	1.30	2.97	0.24	0.57	1.79	2.04	0.46	5.52	0.31	3.57	9.63
Daman & Diu	0.08	0.00	0.00	0.00	0.06	2.74	3.30	1.20	0.61	1.53		0.00	9.90
Lakshadweep	7.04	0.00		0.87	3.75		0.00	0.05	1.04	2.32	5.93	3.67	15.36
Puducherry	2.52	0.90	1.67	0.40	1.19	1.72	1.53	0.30	0.75	0.66	1.34	19.52	3.93
all-India	0.64	0.17	0.15	0.10	0.09	0.08	0.09	0.09	0.11	0.15	0.20	2.17	0.54

Table 1RSE: RSE of average MPCE for each fractile class of MPCE(MMRP)

Sector: Urban

State/UT		RSE (%) of Average MPCE (Rs. 0.00) for each fractile class											
	1	2	3	4	5	6	7	8	9	10	11	12	All
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Andhra Pradesh	1.51	0.46	0.75	0.37	0.46	0.56	0.36	0.43	0.60	1.18	0.86	4.00	2.18
Arunachal Pradesh	2.03	2.07	1.48	1.11	0.69	0.63	1.13	0.92	0.94	1.69	2.38	4.23	6.59
Assam	2.59	0.80	1.23	1.32	0.71	0.61	0.87	0.61	1.35	0.82	2.52	5.15	6.93
Bihar	2.21	0.53	0.87	0.76	0.48	0.79	0.43	0.66	0.61	1.44	1.32	5.54	3.38
Chhattisgarh	2.34	1.38	1.20	0.67	0.37	1.05	1.15	0.95	0.93	1.24	1.58	6.39	8.27
Delhi	3.07	0.39	0.89	1.01	0.86	0.56	0.90	0.71	0.86	1.42	2.61	5.42	4.25
Goa	5.67	0.99	0.70	0.85	0.76	1.45	0.94	0.96	1.17	1.24	1.47	9.16	6.38
Gujarat	1.43	0.52	0.72	1.12	0.39	0.91	0.26	0.39	0.45	0.68	0.70	6.28	3.44
Haryana	3.27	1.71	1.25	0.50	0.51	0.63	0.57	0.74	0.68	2.01	1.73	8.33	5.11
Himachal Pradesh	1.75	2.58	1.65	0.75	0.90	0.93	0.91	1.10	1.19	1.38	1.96	5.88	6.44
Jammu & Kashmir	1.70	0.61	0.64	0.64	0.44	0.35	0.44	0.48	0.54	0.95	0.94	6.50	3.17
Jharkhand	2.02	0.81	0.62	0.54	1.13	1.23	0.49	0.64	0.82	1.08	1.04	5.18	3.30
Karnataka	0.96	0.36	0.73	0.40	0.62	0.75	0.96	0.45	0.49	0.98	1.15	10.19	5.56
Kerala	2.39	0.64	0.61	0.61	0.37	0.45	0.56	0.50	0.55	0.73	1.25	10.85	3.89
Madhya Pradesh	1.43	0.70	0.77	0.52	0.65	0.49	0.40	0.55	0.64	0.92	1.45	4.69	4.58
Maharashtra	1.48	0.48	0.41	0.36	0.31	0.30	0.30	0.38	0.34	0.71	0.97	7.47	2.72
Manipur	0.87	0.14	1.20	0.41	0.21	0.22	0.17	0.18	0.64	0.73	0.75	2.63	1.56
Meghalaya	0.90	3.32	1.84	0.54	0.39	0.72	0.41	0.54	1.02	1.25	1.12	2.96	2.76
Mizoram	1.48	0.66	1.26	0.64	0.57	0.38	0.23	0.67	0.61	0.52	0.60	2.05	2.02
Nagaland	1.28	1.88	0.90	2.60	1.16	0.76	1.23	0.41	0.38	1.47	0.77	4.17	6.73
Odisha	1.40	1.81	0.78	0.97	0.79	0.70	0.70	1.17	0.85	1.10	1.37	3.33	8.63
Punjab	2.11	1.12	0.71	0.57	0.35	0.49	0.39	0.60	0.46	0.64	0.97	4.50	3.40
Rajasthan	2.08	0.80	0.89	0.46	0.60	0.68	0.48	1.02	0.67	1.11	1.06	11.34	4.00
Sikkim	5.43	1.46	3.11	0.36	1.89	1.79	0.29	1.45	3.03	2.25	3.52	2.78	1.98
Tamil Nadu	1.74	0.51	0.56	0.38	0.34	0.26	0.34	0.42	0.43	0.75	0.76	4.76	2.17
Tripura	6.73	0.80	0.54	0.77	0.49	0.73	0.95	0.44	0.84	1.68	1.32	7.41	3.56
Uttarakhand	4.48	0.98	0.70	1.57	1.07	1.42	1.26	1.21	0.85	0.86	1.54	4.10	5.96
Uttar Pradesh	1.58	0.48	0.54	0.66	0.31	0.37	0.37	0.67	0.76	1.08	1.77	4.23	4.68
West Bengal	1.45	0.84	0.66	0.52	0.64	0.60	0.46	0.38	0.51	0.76	0.97	5.32	3.02

Table 1RSE: RSE of average MPCE for each fractile class of MPCE(MMRP)

Sector: Urban

State/UT		RSE (%) of Average MPCE (Rs. 0.00) for each fractile class											
	1	2	3	4	5	6	7	8	9	10	11	12	All
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
A & N Islands	2.65	1.21	0.78	0.54	1.09	0.55	0.70	0.83	1.12	2.88	5.97	3.67	6.95
Chandigarh	7.58	2.89	2.92	2.34	1.01	1.65	2.52	1.09	1.40	4.04	3.41	4.06	15.29
Dadra & Nagar Haveli	2.90	0.68	3.23	2.01	0.50	0.74	1.96	1.18	4.32	4.18	0.00	6.95	13.05
Daman & Diu	1.78	3.01	4.05	6.41	0.03	0.81	0.42	0.16	1.77	3.91	0.07	0.19	10.63
Lakshadweep	1.48	2.25	0.59	2.05	0.69	0.41	1.42	1.64	1.92	2.81	1.34	27.37	11.61
Puducherry	4.48	0.96	1.02	0.99	0.98	0.69	0.47	0.87	0.62	1.02	1.06	14.49	6.04
all-India	0.61	0.20	0.21	0.16	0.15	0.15	0.14	0.13	0.16	0.25	0.30	2.26	1.02

Table 8.2RSE: Monthly per capita quantity and value of consumption at detailed item level for all items of consumption

State: all-India Sector: Urban

Code	item	RSE (%) of value (Rs.[0.00])	no. of sample hhs reporting consumption
(1)	(2)	(3)	(4)
550	bedstead	12.08	1098
551	almirah, dressing table	10.38	519
552	chair, stool, bench, table	9.12	1712
553	suitcase, trunk, box, handbag and other travel goods	8.51	1747
554	foam, rubber cushion	25.53	124
555	carpet, daree & other floor mattings	12.34	822
556	paintings, drawings, engravings, etc.	29.44	190
557	other furniture & fixtures (couch, sofa, etc.)	16.45	402
559	furniture & fixtures: sub-total (550-557)	6.73	5140
560	radio, tape recorder, 2-in-1	23.84	386
561	television	11.07	2204
562	VCR/VCD/DVD player	16.67	468
563	camera & photographic equipment	25.17	186
564	CD, DVD, audio/video cassette, etc	4.96	3445
565	musical instruments	21.52	78
566	other goods for recreation	28.71	125
569	goods for recreation: sub-total (560-566)	8.88	6116
570	stainless steel utensils	3.30	11481
571	other metal utensils	6.51	2656
572	casseroles, thermos, thermoware	10.08	1076
573	other crockery & utensils	4.60	7625
579	crockery & utensils: sub-total (570-573)	3.02	16420
580	electric fan	4.81	3464
581	air conditioner, air cooler	9.82	2825
582	inverter	11.74	1100
583	lantern, lamp, electric lampshade	13.19	356
584	sewing machine	16.96	1834
585	washing machine	9.64	727
586	stove, gas burner	5.80	5517
587	pressure cooker/ pressure pan	5.08	5804
588	refrigerator	9.62	1267
590	water purifier	9.98	1255
591	electric iron, heater, toaster, oven & other electricappliances	14.19	1140
592	other cooking/ household appliances	11.34	824
599	cooking & other household appliances: sub-total (580-592)	4.51	14961
600	bicycle	2.85	13802
601	motor cycle, scooter	5.17	13638
602	motor car, jeep	13.78	2934
603	tyres & tubes	6.14	6855
604	other transport equipment	24.04	209
609	personal transport equipment: sub-total (600-604)	8.28	24056

Table 8.2RSE: Monthly per capita quantity and value of consumption at detailed item level for all items of consumption

State: all-India Sector: Urban

Code	item	RSE (%) of value (Rs.[0.00])	no. of sample hhs reporting consumption
(1)	(2)	(3)	(4)
610	contact lenses, hearing aids & orthopaedic equipment	14.18	328
611	other medical equipment	43.87	39
619	therapeutic appliances: sub-total (610-611)	14.21	365
620	clock, watch	5.62	7709
621	other machines for household work	43.88	44
622	PC/ Laptop/ other peripherals incl. software	10.41	1100
623	mobile handset	4.54	5161
624	telephone instrument (landline)	22.90	56
625	any other personal goods	31.81	274
629	other personal goods: sub-total (620-625)	5.44	12299
630	bathroom and sanitary equipment	27.15	1162
631	plugs, switches & other electrical fittings	12.52	2112
632	residential building & land (cost of repairs only)	5.34	8265
633	other durables (specify)….....	36.97	241
639	residential building, land and other durables: st(630-633)	5.33	10163
640	gold ornaments	8.57	2336
641	silver ornaments	7.37	1596
642	jewels, pearls	36.46	195
643	other ornaments	13.95	4976
649	jewellery & ornaments: sub-total (640-643)	7.68	7842
659	durable goods: total 559+569+579+599+609+619+629+639+649)	4.47	35024