STATISTICS AND APPLICATIONS

GUIDELINES TO AUTHORS FOR PREPARING ARTICLES
(FOR MS WORD FILE)

(1) **Paper size** (Portrait A4): Width: 21 cms; Height: 29.7 cms

(2) **Margins**: Top: 2.54 cms; Bottom: 2.54 cms; Left: 2.54 cms; Right: 2.54 cms; Gutter position: left 0 cm

(3) Headers:
On page 1 there would be no Header
But on page 1, one line space above the title of the paper, left aligned please insert [FONT: TIMES NEW ROMAN; FONT SIZE: 12 - REGULAR]
Statistics and Applications {ISSN 2454-7395 (online)}
Volume ##, No. #, yyyy (New Series), pp 1-last page number of the paper
From page 2 onwards:
[FONT SIZE: 10; FONT: TIMES NEW ROMAN – REGULAR]
Different odd and even page header; Different first page; Don't show page number on first page
Even Page Header: Left most position: Page Number; Centre: Authors Names in Capitals; Right most position: [Vol. ##, No. #
Odd Page Header: Left most position: yyyy]; Centre: Short Running Title in Capitals; Right most position: Page Number
(The page numbers may be 1 onwards. We shall change at our end according to placement in the volume.)
Footers: (Only on page 1)
[There would be no footer from page 2 onwards. But if the authors wish to include a footer based upon their requirement, they may do so. But do follow the font size and type as of Header.]
Footer on page 1
[FONT SIZE: 10; FONT: TIMES NEW ROMAN – REGULAR]
Corresponding Author: Name of the corresponding author
E-mail address: of the corresponding author

An Example

First page - Above the title of the paper
Statistics and Applications {ISSN 2454-7395 (online)}
Volume 18, No. 1, 2020 (New Series), pp 21–34

First page; Footer
Corresponding Author: Priyanka Anjoy
E-mail: anjoypriyanka90@gmail.com

Even Number Page; Header

22 V.K. GUPTA, H. CHANDRA, J. SARKAR, B.N. MANDAL AND R. PARSAD [Vol. 18, No. 1

{If it is not possible to write all names with initials, then write as

22 V.K. GUPTA, H. CHANDRA, J. SARKAR, ET AL. [Vol. 18, No. 1 }

Odd Number Page; Header

2020] ESTIMATION AND SPATIAL MAPPING OF INCIDENCE OF INDEBTEDNESS 23

(4) **Font Type** OF THE ENTIRE MANUSCRIT, INCLUDING THE TABLES, FIGURES, CHARTS, GRAPHS, ETC. AND THE ANNEXURES AND APPENDICES, IF ANY: **TIMES NEW ROMAN ONLY**

{with exception only for graphs, pics or charts obtained using software in which the captions and the text are generated through the software itself automatically. The font style and the font size thus generated through use of software would be acceptable. It need not be Ties New Roman. This is the only exception permitted}

(5) **Font size**:

Title of the paper: {Font Size 16 – Boldface; first alphabet of each word is Capital except those of pronouns like and, on, in, with, for, to, …} (Centered)

Authors Names: {Font Size 12 - Boldface; first alphabet of each word is Capital} (Centered)

Affiliation: {Font Size 12 - Italics; first alphabet of each word is Capital} (Centered)

Received: dd mth yyyy; Revised: dd mth yyyy; Accepted: dd mth yyyy (Font size 12 – Regular, Centered)

Key words: Block designs; Unequal block sizes; Resolvability; Intra block analysis; Universal optimality. [Font size 12 – Regular; except the *Key Words*, which is Font size 12 – italics] (Justified)

AMS Subject Classification: Optional [Font size 12 – bold; but the classification number be regular] (Justified)

Section Number and Heading {Font Size 12 – Boldface; first alphabet of each word is Capital}

Sub-section Number and Heading (Font Size 12 – Boldface; in sentence case meaning thereby first alphabet of first word is capital and all other words are small); As far as possible give sub-section numbers to the sub-sections instead of simply having subsections without numbers.

The sections and subsections should be in hanging mode with tab space of 1 cm between section (sub-section) number and its heading

Sections like Acknowledgements, References etc. will not have any section number and so there would be no indent of 1.0 cm in this case. But structurally they will be same as other sections.

Acknowledgements

References

The Text in the entire manuscript from this point onwards will have FONT: TIMES NEW ROMAN and FONT SIZE 12. This will hold also for Table headings, Graphs, Charts, figures, etc. captions and headings, with exception only for graphs, pics or charts obtained using software in which the captions and the text are also generated through the software itself automatically

(6) Line spacing in the entire body of the text is 1.0 cm (single). This also includes references and the line spacing between references. This also includes line spacing after the section heading and sub-section heading, Tables headings; Figures headings; Charts headings, or anything else

(7) Paragraph marks: Tab (1 cm): Alignment Justified
(This means that the first line of the paragraph would have a tab of 1 cm. and the remaining text would be justified)

(8) Equation numbers to be given at right alignment. All equations numbers be in continuous mode. Do not include section number in the equation number. For example (1), (2), …. , (10), (11), …(15), (16), ……
{**Note**: In case the Appendix / Annexures have equations, then these should have different style of equation numbers. Suppose the Annexures or Appendices are A, B, C, …, then the equations therein would have numbers as (A1), (A2), …, (B1), (B2), …., (C1), (C2),….}

(9) As far as possible, tables and figures should be inserted along with the text. But if the Tables are big, then these should be given in the Annexure. If the Tables give some data, then the source of the Data should be mentioned. The source of data / figure, if any, should be mentioned below the Table in Italics in Font size 11. For example ~ *Source: NSSO Round #*

.

(10) The heading of the table should be in bold face in sentence case starting with Table #, followed by colon, then the title of the table. For example ~ **Table 1: Resolvable block designs for different number of treatments**
The Table numbers would also be continuous and not have section numbers in it

(11) The headings of the Figures (and charts, graphs, etc.) and their numbers should also be given exactly similar to that of Tables

(12) As far as possible, Proofs of Lemmas and Theorems should be given as Appendix. Short proofs may be given in the main body of the paper just below the Lemma or the Theorem

(13) All the equations should be centred as far as possible with equation numbers at extreme right (right aligned). The equations and the equation numbers should be in same line, as far as possible

(14) **IMPORTANT -** In the body of the text all notations and symbols should be in italics. The numerals should be in regular font. All matrices and vectors should be in bold face and regular face. As far as possible, all matrices should be in caps and all vectors should be in small face. This would also hold in the equations as well. All symbols, notations etc. would be italics and all the numerals should be in regular font. And consistency of notations has to be maintained throughout the text. All vectors would be column vectors only and would be represented by a transpose in case of a row vector.

Abbreviations like *e.g., i.e., viz., etc*., should be in italics throughout.

For example, write

Let there be v treatments arranged in b of size k_j each, $j = 1,2, \ldots, b$, with the replication of ith treatment being r_i, $i = 1,2, \ldots, v$. Let $\mathbf{N} = ((n_{ij}))$, where n_{ij} is the replication of ith treatment in jth block. Let $\mathbf{r}'' = (r_{\&}, r_{(}, \cdots, r_v)$. Define N_{ij}^{uv} and $U_{sv}^{(\alpha,\beta)}$ or ST_1. $\mathbf{0}_u$ will denote a column vector of order $u \times 1$ and $\mathbf{1}_s$ would denote a vector of order $s \times 1$. Please note that the multiplication sign throughout the text and in the equations would be as just displayed in the preceding sentence and not as x. Similarly a minus sign, be it in text or in tables or in equations, will be represented as "–" and not as "-"

(15) Citation of References in Text: References in the text may be cited as "Dey (2016); Dey and Nigam (1995); or Dey et al. (2015) [in case there are more than two authors] {The references contain only the last name of the authors and the first and middle name will not be included in the references given in the text}. But the authors should refrain themselves from writing et al. in the references listed in the end of the paper under the section "References". We may also give references in the text as Chander (2016 a, b, c) or Parsad and Dash (2016 a, b) [in case there is more than one paper for the same set of authors in a year]." In extreme case when there are more than four authors in a paper, then in references give the names of first four authors and then comma and then et al.

(16) **Format for References**:
Hanging {left: 0"; Hanging: 1 cm.; Line spacing: Single, Spacing between references also Single}
IMPORTANT - Complete name of Journal should be given in the references, rather than using its acronym. Name of the Journal should in Italics; Volume and issue number in boldface type. Names of the authors would be written as (Last name, initials with a space of one and separated by a stop. Do not write the full form of initials)

Parsad, R., Chandra, H., Mandal, B. N. and Gupta, V. K. (2020).

Format of References

A. Research papers

Dibley, M. J., Godsby, J. B., Staehling, N. W. and Trowbridge, F. L. (1987). Development of normalized curves for the international growth reference: historical and technical considerations. *American Journal of Clinical Nutrition*, **46**(**5**), 736-748.

Datta, G. S., Torabi, M., Rao, J. N. K. and Liu, B. (2018). Small area estimation with multiple covariates measured with errors: A nested error linear regression approach of combining multiple surveys. *Journal of Multivariate Analysis*, **167**, 49-59.

B. Book/ Project/Technical reports

Tiwari, N., Joshi, J. and Upreti, P. K. (2012). *Investigation of Geo-Spatial Hotspots for the Occurrence of Tuberculosis*. Lambert Academic Publishing, Germany (ISBN: 978- 3-8484-0806-1).

Rao, J. N. K. and Molina, I. (2015). *Small Area Estimation*. 2nd Edition, Wiley, Hoboken, NJ.

Csorgo, M., Dawson, D. A., Rao, J. N. K. and Saleh, A. K. Md. E. S. (Eds.) (1981). *Statistics and Related Topics*. North-Holland, Amsterdam.

Binder, D. A., Gratton, M., Hidiroglou, M. A. and Rao, J. N. K. (1984). Analysis of categorical data from surveys with complex designs: some Canadian experiences. In *Proceedings of a Seminar held in Luxemburg*, Eurostat News, Special Number, 75-94.

C. Government Reports

National Institute of Nutrition and Institute of Applied Statistics and Development Studies (2002). *Nutrition Profile of Community in Uttar Pradesh. District Level Reports*. Department of Women and Child Development/Food and Nutrition Board, Government of India.

D. Websites

For websites, give citation as suggested on the Website or give only URLs.

(17) Layout of Manuscript as given below

Statistics and Application {ISSN 2454-7395(online)}
Volume **, No. **, yyyy (New Series), pp ***-***
{One blank line: line size 12, spacing 1.0 cm. (Single)}
Title {Font Size 16; Boldface, first alphabet of each word is Capital except pronouns like in, of, or, for, with,…)
{One blank line: line size 12, spacing 1.0 cm. (single)}
Authors {Font Size 12; Boldface, first alphabet of each word is Capital)
Affiliation {Font Size 12*; Italics)*
{One blank line: line size 12, spacing Single}
Received dd mth yyyy; Revised dd mth yyyy; Accepted dd mth yyyy
Follow it by **Abstract**, *Key Words*: **AMS Subject Classification** (optional)
These three would be placed within thin borders, one at top of Abstract and the other at the bottom of AMS Subject Classification (or Key Words)

Abstract (Font: Times New Roman; Font size 12, Bold)
{One blank line, spacing Single, text justified with first line at a tab of 1 cm.}
Font: Times New Roman; Font size 12, Regular
First line – indent 1 cm.
All other lines – no indent, start from left.
{One blank line, spacing Single}
Key words: (Font: Times New Roman; Font size 12, Italics; The actual key words will be regular with first letter of first word capital and all others small and two key words separated by a colon)
Key words: Block designs; Unequal block sizes; Resolvable designs; Intra block analysis; Universal optimality.
{One blank line: line size 12, spacing Single}
AMS Subject Classification: 62K99; 62J05 (Optional)

---

1. Introduction
{One blank line: line size 12, spacing Single}

A regression model describes the influence of various factors on the response under study. In agricultural experiments, regression models are used to study the effect of manure on the yield of a crop.

2. The Problem and Its Perspectives

{One blank line: line size 12, spacing Single}

Suppose the manure applied during sowing of the crop is a q-component mixture with mixing proportions given by $\mathbf{x}'' = (x_{\&}, x_{(}, \cdots, x_q)$, where $\mathbf{x} \in \Xi \{(x_{\&}, x_{(}, \cdots, x_q) | x_i \geq 0, 1 \leq x_{\mathrm{i}} \geq 1, \sum_{\mathrm{i}?\&}^{q} x_{\mathrm{i}} = 1\}$.

Acknowledgements
References
ANNEXURE, if any
APPENDIX, if any

IN THE SEQUEL IS GIVEN A TEMPLATE THAT CAN ALSO BE OF HELP IN PREPARATION OF MANUSCRIPT

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 2, 2021 (New Series), pp 121–135

Estimation and Geo-Spatial Mapping of Incidence of Indebtedness in the State of Karnataka in India by Juxtaposing Survey and Census Data

V.K. Gupta[1], Hukum Chandra[2], Jyotirmoy Sarkar[3], B.N. Mandal[2] and Rajender Parsad[2]

[1]*Former ICAR National Professor at ICAR-Indian Agricultural Statistics Research Institute, New Delhi*
[2]*ICAR-Indian Agricultural Statistics Research Institute, New Delhi*
[3]*Department of Mathematical Sciences, Indiana University–Purdue University Indianapolis, Indiana, USA*

Received: 20 August 2020; Revised: 19 November 2020; Accepted: 27 December 2020

Abstract

Information about the household debt behaviour in different occupational categories is of key importance to the Governmental organization for taking effective policy measures targeting the vulnerable groups. The purpose of this paper is to illustrates small area estimation (SAE) methodology to estimate extent of indebtedness in rural areas for the two major occupation categories- rural cultivator and rural non-cultivator as well as for both categories combined together across all the 30 districts of Karnataka state in India using the data of All India Debt and Investment Survey 2012-13 and population census 2011. The findings show that the district-level estimates of incidence of indebtedness obtained from SAE are more precise than the direct survey estimates. A spatial map has also been produced to observe the inequality in distribution of indebtedness within districts and in each occupational category across districts. Such maps are definitely useful for framing consistent policy actions and fund disbursement for the indebted household mass.

Key words: Small area estimation; Generalized linear mixed model; Covariates; Principal component analysis; Indebtedness; Spatial Map.

1. Introduction

Agriculture plays an important role in the economy of Karnataka and it is the main occupation for more than 60% of population. Karnataka is a drought-prone region with a large proportion of wasteland and having the second largest arid zone in the country after Rajasthan. And due to these factors, the state has been facing agrarian distress with increasing incidence of farmers' suicides since 1997. In fact, the rate of farmer suicides in Karnataka has hit the highest level in a decade, topping the list after Maharashtra, highlighting agrarian distress in the state, according to the report Accidental Deaths and Suicides in India 2015 published by National Crime Records Bureau (NCRB). According to NCRB 2015 data, about 1,197 farmers committed suicides in Karnataka during 2014-15; the state was just behind Maharashtra and Telangana. The NCRB also found that about 79% suicides (946 out of 1,197) in Karnataka were due to bankruptcy or indebtedness. The pre-requisite for any effective policy approach taken in this regard is a proper statistical and

Corresponding Author: B.N. Mandal
Email: mandal.stat@gmail.com

economic framework that allows for an effective analysis and monitoring of farmers' distress. Measure of disaggregated level indebtedness can be an important tool to the policy makers to mark certain region or group for upliftment and reduce the situation of agrarian distress or farmers' suicides. In this study we attempt to estimate such micro or disaggregated level incidence of indebtedness at micro or local level using the area level small area model.

Most of the large-scale surveys are planned to produce reliable estimates at macro or higher geographical (*e.g.* national and state) level, and cannot be used directly to generate reliable micro or local (also referred to as small area) level estimates because of the small sample sizes (Rao and Molina 2015). This is because, large scale survey designed for a large population (*e.g.* national and state level) may select a small number of units or even no unit from the small area of interest (*e.g.* district or further disaggregation of district). Hence, sample sizes from small areas (or small domains) are too small to justify the use of traditional direct survey estimates. The underlying theory in the literature of survey sampling that helps in resolving the problem of smaller sample sizes is referred as small area estimation (SAE) technique. The technique is model-based methods that links the variable of interest from survey with the auxiliary information available from other data sources for small areas and hence increase the overall (effective) sample size and precision. In this paper we employ area level SAE technique to produce reliable estimates of the incidence of indebtedness among cultivators and non-cultivators categories as well as for both the categories combined in different districts of rural areas of Karnataka in India by linking data from the All-India Debt and Investment Survey (AIDIS) 2012-13 of National Sample Survey Office (NSSO), and the Population Census 2011. This work will enable us to obtain spatial distribution of incidence of indebtedness as well as regional inequality in such measure of indebtedness among the farm families and other families of rural areas in Karnataka. The rest of the paper has been organized into five sections. In Section 2, we discuss the data used in the paper. Section 3 provides an overview of SAE technique that has been used to generate incidence of indebtedness among occupational category by districts in Karnataka. In Section 4, we present diagnostic procedures to examine model assumptions and validate small area estimates including discussion about the results. Finally, Section 5 provides concluding remarks and some recommendations.

2. Features and Summary of Data

To begin with, it would be in order to describe the data that have been used in this analysis. In particular, the SAE analysis is based on the AIDIS 2012-13 data for rural areas of the State of Karnataka in India and the Population Census of 2011. The sampling design used in the AIDIS 2012-13 data is stratified multi-stage random sampling with districts as strata, the census villages in the rural sector as first stage units and households as the ultimate stage units. For the state of Karnataka, there are a total of 2340 surveyed rural households (including both indebted and non-indebted) spread over 30 districts. The rural households are broadly classified into two types; namely; cultivator and non-cultivator households. As per the concepts and definitions of AIDIS, all rural households operating at least 0.002 hectare of land during the 365 days preceding the date of survey are treated as 'cultivator households'. On the other hand, all rural households operating no land or land less than 0.002 hectare are considered to be non-cultivator households. What follows, based on land holding size (LHS), we denote three categories of households: (i) LHS-A: All households (ii) LHS-C: Cultivator-households with LHS greater than 0.002 ha, and (iii) LHS-NC: Non cultivator-households with LHS less or equal to 0.002 ha. Here, the districts and district by household categories are small areas of interest. Table 1 presents the distribution of district-wise sample sizes for three categories of households. Across all the districts (*i.e.* LHS-A), the sample size ranges between a minimum of 55 households to a maximum of 112 with an average of 78 households. The sample sizes become too small if sub-grouped further by land holding size categories (i.e. district by cultivator and non-cultivator categories). That is, the sample size of rural cultivators (LHS-C) varies from a minimum of 23 to a maximum of 90 households across the 30 districts with an average of 49 households. And for non-cultivators (LHS-NC), the sample size varies from a minimum of 11 to a maximum of 51 households across the districts with an average of 29 households. Such small samples from the districts pose a challenge in deriving reliable direct estimates of indebtedness. Thus, SAE is an obvious choice to address this problem.

Table 1: Distribution of sample size by occupational categories across districts in rural Karnataka

District	All	Cultivator	Non-Cultivator	District	All	Cultivator	Non-Cultivator
Belgaum	112	67	45	Tumkur	112	90	22
Bagalkot	84	57	27	Kolar	56	45	11
Bijapur	112	85	27	Bangalore	56	23	33
Gulbarga	98	60	38	Bangalore Rural	56	34	22
Bidar	84	49	35	Mandya	112	85	27
Raichur	84	55	29	Hassan	84	63	21
Koppal	84	63	21	Dakshina Kannada	84	41	43
Gadag	56	31	25	Kodagu	56	35	21
Dharwad	56	28	28	Mysore	112	71	41
Uttara Kannada	56	32	24	Chamarajanagar	56	39	17
Haveri	84	52	32	Ramanagara	55	24	31
Bellary	112	72	40	Chikkaballapura	56	42	14
Chitradurga	84	33	51	Yadgir	56	44	12
Davanagere	84	58	26	Minimum	55	23	11
Shimoga	87	50	37	Maximum	112	90	51
Udupi	56	28	28	Average	78	49	29
Chikmagalur	56	27	29	Total	2340	1483	857

Two types of variables are utilized in SAE technique, the variable of interest and the auxiliary variable. As noticed in Section 1, the auxiliary (covariates) variables play an important role in SAE. The auxiliary variables for this analysis are available at district level from the Census 2011. The Population Census 2011 provides a number of covariates at district level that can be utilized for small area modeling. We therefore carried out a preliminary data analysis in order to define appropriate covariates for SAE modeling, using Principal Component Analysis (PCA) to derive composite scores for selected groups of variables. In particular, we carried out PCA separately on three groups of variables, all measured at district level and identified as P1, P2 and P3 below. The first group (P1) consisted of literacy rates by gender and proportions of worker population by gender. The first principal component (P11) for this group explained 61% of the variability, while adding the second principal component (P12) increased explained variability to 85%. The second group (P2) consisted of the proportions of main worker by gender, proportions of main cultivator by gender and proportions of main agricultural labourer by gender. The first principal component (P21) for this second group explained 48% of the variability in the P2 group, while adding the second component (P22) increased explained variability to 62%. Finally, the third group (P3) consisted of proportions of marginal cultivator by gender and proportions of marginal agriculture labourers by gender. The first principal component (P31) for this third group explained 37% of the variability in the P3 group, while adding the second component (P32) increased explained variability to 60%. Finally, three variables, P11, P21 and P31 that significantly explained the model with AIC value 51.59, are identified for the use in SAE analysis. In this paper, the *Y*-variable of interest is the indebted households, i.e. whether a household is in debt or not. A household is defined to be indebted if it has outstanding loan (from respective source) as on 30.06.2012. The target is to estimate the proportion of indebted household (*i.e.* the incidence of indebtedness) at the district (LHS-A) and district by household category (LHS-C and LHS-NC) level. Incidence of indebtedness (IOI) is defined as number of households with any one loan (from respective source) divided by all households in that population segment.

3. Methodological Framework

This Section describes the methodology used in the small area analysis considered in this paper. To begin with, we assume a finite population *U* of size *N* which is consisting of D non-overlapping and mutually exclusive small areas (or district in this paper). We assume that a sample *s* of size *n* is drawn from this population using a probability sampling method. Here, a subscript *d* has been used to denote quantities related to small area d. Let U_d and s_d be the population and sample of sizes N_d and n_d in area d, respectively such that $U = \bigcup_{d=1}^{D} U_d$, $N = \sum_{d=1}^{D} N_d$, $s = \bigcup_{d=1}^{D} s_d$ and $n = \sum_{d=1}^{D} n_d$. We use subscript s and r respectively to denote quantities related to sample and non-sample parts of the population. Let y_{di} denotes the value of the variable of interest for unit $i\,(i=1,...,N_d)$in area *d*. The variable of interest, with values y_{di}, is binary (*e.g.* $y_{di} = 1$ if *i*th household is in debt and 0 otherwise) in area d , the aim is to estimate the small area population count, $y_d = \sum_{i\in U_d} y_{di}$, or equivalently the small area proportion, $P_d = N_d^{-1} y_d$, in area *d*. The standard direct survey estimator (hereafter denoted by DIR) for P_d is, $p_{d\mathrm{w}} = \sum_{(\in s_d} \dot{w}_{d(}\, y_{d(}$ where $\dot{w}_{d(} = \frac{\mathrm{w}_{d\mathrm{i}}}{\sum_{\mathrm{i}\in s_d} \mathrm{w}_{d\mathrm{i}}}$ is the normalized survey weight with $\sum_{(\in s_d} \dot{w}_{d(} = 1$ and w_{di} is the survey weight for unit *i* in area *d*. The estimated design-based variance of DIR is approximated by $v(p_{d\mathrm{w}}) = \sum_{(\in s_d} \dot{w}_{d(}\,(\dot{w}_{d(} - 1)(y_{d(} - p_{d\mathrm{w}})^5$, with the simplifications $w_{di} = a_{di}^{-1}$, $a_{di,di} = a_{di}$ and $a_{di,dj} = a_{di}a_{dj}, i \neq j,$ where a_{di} is the first order inclusion probability of unit *i* in area *d* and $a_{di,dj}$ is the second order inclusion probability of units *i* and *j* in area *d*. Under simple random sampling (SRS), $w_{di} = N_d n_d^{-1}$ and DIR is then $p_d = n_d^{-1} y_{sd}$, with estimated variance $v(p_d) \approx n_d^{-1} p_d (1-p_d)$, where $y_{sd} = \sum_{i\in s_d} y_{di}$ denotes the sample count in area *d*. Similarly, $y_{rd} = \sum_{i\in s_r} y_{di}$ denotes the non-sample count in area *d*. If the sampling design is informative, this SRS-based version of DIR may be biased. Furthermore, DIR is based on area-specific sample data and can therefore be very imprecise when the area specific sample size is small or may even be impossible to compute if this sample size is zero. However, model-based SAE procedures that ‘borrow strength’ via a common statistical model for all the small areas can be used to address this problem. If we ignore the sampling design, the sample count y_{sd} in area (*i.e.* district) *d* can be assumed to follow a Binomial distribution with parameters n_d

and π_d , *i.e.* $y_{sd} = Bin(n_d, \pi_d)$, where π_d is the probability of occurrence of an event for a population unit in area *d* or the probability of prevalence in area *d*. Similarly, for the non-sample count, $y_{rd} \sim Bin(N_d - n_d, \pi_d)$. Further, y_{sd} and y_{rd} are assumed to be independent binomial variables with π_d being a common success probability.

Let $\mathbf{x}_d$ be the *k*-vector of covariates for area *d* from available data sources. Following Chandra *et al.* (2011) the model linking the probability π_d with the covariates $\mathbf{x}_d$ is the logistic linear mixed model of the form

$$logit(\pi_d) = \ln\left\{\pi_d(1-\pi_d)^{-1}\right\} = \eta_d = \mathbf{x}_d^T\boldsymbol{\beta} + u_d, \tag{1}$$

with $\pi_d = \exp(\mathbf{x}_d^T\boldsymbol{\beta} + u_d)\left\{1+\exp(\mathbf{x}_d^T\boldsymbol{\beta} + u_d)\right\}^{-1}$. Here $\boldsymbol{\beta}$ is the *k*-vector of regression coefficients, often known as fixed effect parameters, and u_d is the area-specific random effect that captures the area dissimilarities. We assume that u_d's are independently and normally distributed with mean zero and variance σ_u^2. Here, we observe that model (1) relates the area level proportions (direct estimates) from the survey data to the area level covariates. The Fay and Herriot (FH) method for SAE is based on area level linear mixed model and their approach is applicable to a continuous variable. Model (1), a special case of a generalized linear mixed model (GLMM) with logit link function, is suitable for modelling discrete data, particularly the binary variables. (Chandra, 2013; Chandra *et al.*, 2017). Under model (1), an empirical predictor (EP) of the population count y_d in area *d* is

$$\hat{y}_d^{EP} = y_{sd} + \hat{y}_{rd} = y_{sd} + \left(N_d - n_d\right)\left[\exp(\mathbf{x}_d^T\hat{\boldsymbol{\beta}} + \hat{u}_d)\left(1+\exp(\mathbf{x}_d^T\hat{\boldsymbol{\beta}} + \hat{u}_d)\right)^{-1}\right]. \tag{2}$$

An estimate of the corresponding proportion in area *d* is obtained as $\hat{p}_d^{EP} = N_d^{-1}\hat{y}_d^{EP}$. It is obvious that in order to compute the small area estimates by equation (2), we require estimates of the unknown parameters β and $\mathbf{u} = (u_1, \ldots, u_D)^T$. We can observe that the parameters $\boldsymbol{\beta}$ and σ_u^2 are the same for every area; *i.e.*, they can be estimated using the data from all small areas. We use an iterative procedure that combines the Penalized Quasi-Likelihood (PQL) estimation of β and $\mathbf{u}$ with REML estimation of σ_u^2 to estimate unknown parameters (Chandra *et al.*, 2011).

The mean squared error (MSE) estimates are computed to assess the reliability of estimates and also to construct the confidence interval (CI). The MSE estimate of (2) is:

$$mse(\hat{p}_d^{EP}) = M_1(\hat{\sigma}_u^2) + M_2(\hat{\sigma}_u^2) + 2M_3(\hat{\sigma}_u^2). \tag{3}$$

Following Chandra *et al.* (2011) we define few notations to express different components of (3). We denote by $\hat{\mathbf{V}}_s = diag\left\{n_d\hat{p}_d^{EP}(1-\hat{p}_d^{EP})\right\}$ and $\hat{\mathbf{V}}_r = diag\left\{(N_d - n_d)\hat{p}_d^{EP}(1-\hat{p}_d^{EP})\right\}$ the diagonal matrices defined by the corresponding variances of the sample and non-sample parts, respectively. We then define $\mathbf{A}=\left\{diag(N_d^{-1})\right\}\hat{\mathbf{V}}_r$, $\mathbf{B}=\left\{diag(N_d^{-1})\right\}\hat{\mathbf{V}}_{rd}\mathbf{X}-\mathbf{A}\hat{\mathbf{T}}\hat{\mathbf{V}}_s\mathbf{X}$ and $\hat{\mathbf{T}}=\left(\hat{\sigma}_u^2\mathbf{I}_D+\hat{\mathbf{V}}_s\right)^{-1}$, where $\mathbf{X}=(\mathbf{x}_1^T,\dots,\mathbf{x}_D^T)^T$ is a $D\times k$ matrix, and $\mathbf{I}_D$ is an identity matrix of order *D*. We further write $\hat{\mathbf{T}}_{11}=\left\{\mathbf{X}^T\hat{\mathbf{V}}_s\mathbf{X}-\mathbf{X}^T\hat{\mathbf{V}}_s\hat{\mathbf{T}}\hat{\mathbf{V}}_s\mathbf{X}\right\}^{-1}$ and $\hat{\mathbf{T}}_{22}=\hat{\mathbf{T}}+\hat{\mathbf{T}}\hat{\mathbf{V}}_s\mathbf{X}\hat{\mathbf{T}}_{11}\mathbf{X}^T\hat{\mathbf{V}}_s^T\hat{\mathbf{T}}$. Under model (1), the components of MSE estimate are: $M_1(\hat{\sigma}_u^2)=\mathbf{A}\hat{\mathbf{T}}\mathbf{A}^T$, $M_2(\hat{\sigma}_u^2)=\mathbf{B}\hat{\mathbf{T}}_{11}\mathbf{B}^T$ and $M_3(\hat{\sigma}_u^2)=trace\left(\hat{\nabla}_i\hat{\Sigma}\hat{\nabla}'_j v(\hat{\sigma}_u^2)\right)$ with $\hat{\Sigma}=\hat{\mathbf{V}}_{sd}+\hat{\phi}\mathbf{I}_D\hat{\mathbf{V}}_{sd}\hat{\mathbf{V}}_{sd}^T$. Let us write $\Delta=\mathbf{A}\hat{\mathbf{T}}$ and $\hat{\nabla}_i=\partial(\Delta_i)/\partial\phi\big|_{\phi=\hat{\phi}}=\partial(\boldsymbol{A}_i\hat{\mathbf{T}})/\partial\sigma_u^2\big|_{\sigma_u^2=\hat{\sigma}_u^2}$, where $\boldsymbol{A}_i$ is the i^{th} row of the matrix $\boldsymbol{A}$. Here $v(\hat{\sigma}_u^2)$ is the asymptotic covariance matrix of the estimate of variance component $\hat{\sigma}_u^2$, which can be evaluated as the inverse of the appropriate Fisher information matrix for $\hat{\sigma}_u^2$. This term also depends upon whether we use ML or REML estimate of $\hat{\sigma}_u^2$. We use REML estimates for $\hat{\sigma}_u^2$ and where $v(\hat{\sigma}_u^2)=2\left((\hat{\sigma}_u^2)^{-2}(D-2t_1)+(\hat{\sigma}_u^2)^{-4}t_{11}\right)^{-1}$ with $t_1=(\hat{\sigma}_u^2)^{-1}trace(\hat{\mathbf{T}}_{22})$ and $t_{11}=trace(\hat{\mathbf{T}}_{22}\hat{\mathbf{T}}_{22})$.

4. Model Settings and Estimators

The semi-parametric or the non-parametric covariance model considered for the studyis of the form

$$\mathrm{Y}=\mathrm{X}\beta+\phi(\mathrm{U})+\varepsilon \tag{4}$$

where, Y is the observation vector, m=Xβ+ϕ(U), is the regression function, **X** is the design matrix, β is the vector of treatment effect, ϕ(U)is the non-parametric function representing the relationship between Y−Xβ and the covariate U which is assumed to be a smooth function and ε is the error term assumed to be *iid* with mean vector 0 and covariance matrixσ^2I. Back-fitting algorithm (Buja *et al.*, 1989) is used to estimate the treatment vector and covariate effect in the regression model and estimates are given by

$$\hat{\beta}=[\mathbf{X}^T(\mathbf{I}-\mathbf{S})\mathbf{X}]^{-S}\mathbf{X}^T(\mathbf{I}-\mathbf{S})\mathbf{Y},\ \ \hat{\boldsymbol{\phi}}=\mathbf{S}\mathrm{V}\mathbf{Y}-\mathbf{X}\hat{\beta}\mathrm{W}\ \text{ and }\ \hat{\mathbf{m}}=\mathbf{X}\hat{\beta}+\hat{\boldsymbol{\phi}}$$

where, *S* is the smoothing matrix derived using local linear regression (Ruppert and Wand, 1994). Let S$_i$ be the i^{th} row of the smoother matrix, then

$$\mathbf{S}=[\mathbf{S}_\mathrm{S}\ \dots\ \mathbf{S}_\backslash]^T_\mathrm{S}$$
$$\mathbf{S}^T_{(}=\mathbf{e}^T\mathrm{V}_\mathrm{S}\mathbf{Z}^T_{u_\mathrm{i}}\mathbf{W}_{u_\mathrm{i}}\mathbf{Z}_{u_\mathrm{i}}\ \mathrm{W}\ \mathbf{Z}^T_{u_\mathrm{i}}\mathbf{W}_{u_\mathrm{i}}$$

where,

$$\mathbf{Z}_{u_\mathrm{i}}=\mathrm{a}\begin{matrix}1 & (u_\mathrm{S}-u_()\\ \vdots & \vdots\\ 1 & (u_n-u_()\end{matrix}\mathrm{e},\ \ \mathrm{e}_1^\mathrm{T}=[1\ 0\ 0]$$

and $\mathbf{W}_{u_\mathrm{i}}=diag\,\mathbf{j}K\,\mathbf{l}\frac{u_\mathrm{m}-u_\mathrm{i}}{\mathrm{n}}\mathrm{o},\dots,K\,\mathbf{l}\frac{u_n-u_\mathrm{i}}{\mathrm{n}}\mathrm{o}$rfor some kernel functions *K* and bandwidth *h*. The properties of the estimates are provided by Jose and Ismail (2001) and Rupert and Wand (1994). Cross-validation (leave-one-out) technique is generally used to estimate the optimum bandwidth *h*. The cross-validation score is given by

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{m}_{(-i)}\right)^2$$

where, y_i, $i = 1,\ldots,n$ are the observations and $\hat{m}_{(-i)}$ is the leave-one-out estimate (estimated value of m_i without using the i^{th} observation) with h as bandwidth. The optimum bandwidth is the value of h which minimizes the cross-validation score $CV(h)$. The estimate, $\hat{\beta}$ is asymptotically unbiased and its asymptotic variance is $\sigma^2(X^TX)^{-1}$ which is same as the fully parametric model (Opsomer and Ruppert, 1999). Cleveland and Devlin (1988) and Hastie and Tibshirani (1990) discussed the estimation of error variance in linear regression smoothers. An approximate estimate of the error variance is given by

$$\hat{\sigma}^2 = \frac{1}{[n-p-2trace(\mathbf{S})+trace(\mathbf{S}^T\mathbf{S})]}\left(\mathbf{Y}-\mathbf{X}\hat{\beta}-\hat{\boldsymbol{\phi}}\right)^T\left(\mathbf{Y}-\mathbf{X}\hat{\beta}-\hat{\boldsymbol{\phi}}\right)$$

The variance of $\hat{\beta}$ is estimated by

$$\hat{V}(\hat{\beta}) = diag(\mathbf{PP}^T)\hat{\sigma}^2$$

where, $P = (X^T(I\text{-}S)\ X)^{-1}X^T(I\text{-}S)$. The significance of the covariate effect ϕ can be tested using the lack of fit statistic or by comparing the mean residual sum of squares (Hart, 1997; Jose, *et al.*, 2009). Under the null hypothesis that the covariate effect $\phi(U)=\mathbf{0}$, the mean residual sum of squares obtained by fitting the model (1) is given by

$$\hat{\sigma}_0^2 = \mathbf{Y}^T[\mathbf{I}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}]^T[\mathbf{I}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}]\mathbf{Y}/(n-p-1)$$

The lack of fit statistic is given by

$$R = \frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}$$

The statistic R asymptotically follows an F distribution with (n–p–1), [n–p–2*trace*($\mathbf{S}$)+ *trace* ($\mathbf{S}^T\mathbf{S}$)] degrees of freedom and it can be used for testing the significance of the covariate effect.

5. Analysis of Data in the Presence of Outliers

The regression estimate and the cross-validation technique can behave very badly in the presence of outliers in the data or when the errors are heavy-tailed (Leung, D., 2005). One remedy is to remove the influential observations from the data. Another approach is to use robust smoother, which is not as vulnerable as the usual smoothing technique. A robust M-type estimate $\hat{m}$ of the regression function can be obtained by minimizing the objective function

$$\sum_{i=1}^{n}\rho\left(\frac{y_i-\hat{m}_i}{s}\right) \qquad (5)$$

where, $\rho(.)$ is an even function with bounded first derivative $\psi(.)$ and a unique minimum at zero. The derivative $\psi(x) = \frac{d\rho(x)}{dx}$ is called the influence function and $w(x) = \frac{\psi(x)}{x}$ is the corresponding weight function. Several M-type estimators have been discussed in literature using different types of influence functions (Huber, 1981; Rey, 1983; Hampel *et al.*, 1986; Tukey, 1977). Tuckey's bi-weight robust function is very popular and it is considered in this paper. The ρ, ψ and *w* functions corresponding to the Tuckey's robust estimator is given by

$$\rho(x) \quad = \begin{cases} \frac{c^5}{6}\,\text{î}1 - \text{ï}1 - \text{l}\,\frac{x}{c}\,\text{o}^{5}\,\text{ñó}^{\text{ò}} & |x| \le c \\ \frac{c^5}{6} & |x| > c \end{cases}$$

$$\psi(x) = \text{ú} \quad \begin{matrix} x\,\text{îï}1 - \text{l}\,\frac{x}{c}\,\text{o}^{5}\,\text{ñó}^{5} & |x| \le c \\ 0 & |x| > c \end{matrix}$$

$$w(x) = \text{ú} \begin{matrix} \text{îï}1 - \text{l}\,\frac{x}{c}\,\text{o}^{5}\,\text{ñó}^{5} & |x| \le c \\ 0 & |x| > c \end{matrix}$$

The turning constant *c* is picked to give reasonably high efficiency. When the errors are normal and *x* is the standardized residual, then *c* = 4.685 produce 95% efficiency.

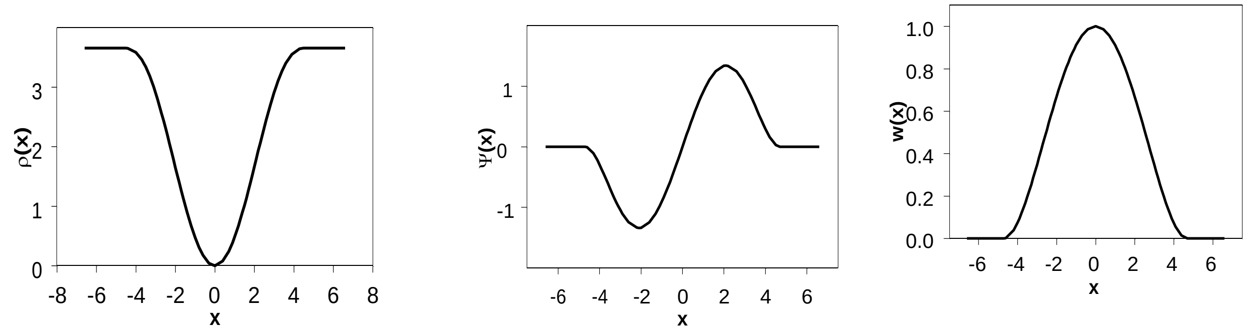


Figure 1: ρ, ψ **and** *w* **functions of Tuckey's bi-weight robust estimate**

Iterated reweighted least squares technique is used to solve the minimization problem in eq. (2) to obtain the robust estimate of the regression function. The estimate of the regression function in the k^{th} iteration is given by

$$\hat{\mathbf{m}}_{(k)} = \mathbf{X}\hat{\beta}_{(k)} + \hat{\boldsymbol{\phi}}_{(k)}$$

$$\hat{\beta}_{(k)} = \text{w}\mathbf{X}^{T}\text{V}\mathbf{I} - \mathbf{S}_{(k)}\text{W}\mathbf{X}\text{y}^{-\text{S}}\mathbf{X}^{T}\text{V}\mathbf{I} - \mathbf{S}_{(k)}\text{W}\boldsymbol{Y}$$

$$\hat{\boldsymbol{\phi}}_{(\mathbf{k})} = \mathbf{S}_{(k)}\text{V}\mathbf{Y} - \mathbf{X}\hat{\beta}_{(k)}\text{W}$$

where, $\mathbf{S}_{(k)}$ is the smoothing matrix in the k^{th} iteration derived using robust local linear regression. Let $\mathbf{S}_{i(k)}$ be the i^{th} row of the smoothing matrix in the k^{th} iteration, then

$$\mathbf{S}_{(k)} = \text{w}\mathbf{S}_{\text{S}(k)} \ \ldots \ \mathbf{S}_{n(k)}\text{y}^{T}$$

$$S_{(k)}^{T} = e_{1}^{T}\left(Z_{u_i}^{T}W_{u_i(k)}^{*}Z_{u_i}\right)^{-1}Z_{u_i}^{T}W_{u_i(k)}^{*}$$

$$W_{u_i(k)}^{*} = diag\{w^{*}(r_{1(k-1)}),\dots,w^{*}(r_{n(k-1)})\}$$

$$w^{*}(r_{j(k-1)}) = \frac{K\left(\frac{u_i-u_j}{h}\right)w(r_{j(k-1)})}{\sum_{l=1}^{n}K\left(\frac{u_i-u_l}{h}\right)w(r_{l(k-1)})}, \qquad j = 1,\dots,n$$

where $w(r_{j(k-1)})$ is the value of the robustness weight function corresponding to y_j in the k^{th} iteration and $r_{(k-1)j} = \frac{(y_j-\hat{m}_{j(k-1)})}{s_{(k-1)}}$ is the standardized residual of the j^{th} datum in the $(k–1)^{th}$ iteration with $\hat{m}_{j(k-1)}$ as the estimated value and $r_{(0)i} = 0$ for $i = 1,\dots,n$. The Median of Absolute Deviation from median (*MAD*) is used for computing a robust estimate for the scale factor *s* and

$$s_{(k-1)} = \frac{\underset{i}{median}\left|e_{(k-1)i} - \underset{j}{median}(e_{(k-1)j})\right|}{0.6745}$$

where $e_{(k-1)i} = y_i - \hat{m}_{i(k-1)}$

The estimate of the regression function in the k^{th} iteration is written as

$$\hat{\mathbf{m}}_{(k)} = \mathbf{X}\hat{\beta}_{(k)} + \mathbf{S}_{(k)}(\mathbf{Y} - \mathbf{X}\hat{\beta}_{(k)})$$

Iteration is continued till there is no significant improvement in the estimated valuesand the final estimate of the regression function is written as

$$\hat{\mathbf{m}}^{*} = \mathbf{X}\hat{\beta}^{*} + \mathbf{S}^{*}(\mathbf{Y} - \mathbf{X}\hat{\beta}^{*})$$
$$\hat{\beta}^{*} = [\mathbf{X}^{T}(\mathbf{I} - \mathbf{S}^{*})\mathbf{X}]^{-1}\mathbf{X}^{T}(\mathbf{I} - \mathbf{S}^{*})\mathbf{Y}$$
$$V(\hat{\beta}^{*}) = diag(\mathbf{P}^{*}\mathbf{P}^{*T})\hat{\sigma}^{*2}$$

where $\mathbf{S}^{*}$ is the smoothing matrix of the final iteration, $\hat{\mathbf{m}}^{*}$, $\hat{\beta}^{*}$ and $\hat{\sigma}^{*}$ are the final estimates of the regression function, treatment vector and scale factor respectively and

$$P^{*} = (X^{T}(I-S^{*})\ X)^{-1}X^{T}(I-S^{*})$$

Optimum bandwidth: Let $w_i^{\#}$ be the final robustness weight assigned to y_i and $\hat{m}_{i(h)}^{\#}$ be the estimated value of m_i with band width *h*. The Mean Squared Error (*MSE*) of the estimated value corresponding to the bandwidth *h* is given by

$$MSE(h) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{m}_{i(h)}^{\#}\right)^{2}$$

The cross-validation score *CV* (*h*) does not work well for the robust smoothers because the *CV* function itself is strongly influenced by the outliers (Wang and Scott, 1994). The cross- validation score is the sum of squares of the prediction errors of the smoother at each of the design points. When there are outliers, the prediction errors corresponding to the outliers will be uncharacteristically extreme and these extreme prediction errors will inflate the *CV*(*h*). Therefore, similar to robust smoothing technique, the influence of extreme prediction errors should be minimized. A robust cross validation score *RCV* (*h*) is defined as

$$RCV(h) = \frac{\sum_{i=1}^{n} w_i^{\#}\left(y_i - \hat{m}^{\#}_{h(-i)}\right)^2}{\sum_{i=1}^{n} w_i^{\#}}$$

where, $w_i^{\#}$ is the final robustness weight defined earlier, $\hat{m}^{\#}_{h(-i)}$ is the robust estimate of y_i with h as bandwidth and without using the i^{th} observation y_i. The value of h which minimizes the robust cross validation score *RCV* (*h*) will be the optimum bandwidth. In the computation of *RCV*(*h*), the effect of outliers is controlled by taking weighted sum of squares of the prediction errors of the smoother at each of the design points with the robustness weight $w_i^{\#}$.

6. Simulation Study

A simulation study was conducted to evaluate the performance of the proposed method. The semi-parametric regression model considered for the simulation study is given by

$$\mathrm{Y} = \mathrm{X}\beta + \phi(\mathrm{U}) + \varepsilon \qquad (6)$$

where **Y** is the $n\times1$ observation vector, **m** = $\mathrm{X}\beta + \phi(\mathrm{U})$, is the regression function, **X** is the $n\times k$ design matrix, β is the $k\times1$ treatment effect vector which is taken as $\beta^T = [-2\ -2\ 0\ 4]$, $\phi(u) = 1+2sin(\pi u)$ and the random error vector ε follows $N(\mathbf{0}, \sigma^2 \mathrm{I})$ and $u \in [0,1]$. Based on the above, 100 sets of data are simulated for different values of *n* (100, 200, 400) and σ (1.0, 2.0) with 0%, 4% and 8% outliers. To generate data with specific percentage of outliers, the required number of random numbers between 0 to n are generated and the value of the regression function *m* corresponding to the data points are replaced with $m+6\sigma$. The Epanechnikov kernel function $K(u) = 0.75(1-u^2)$ is employed in the study. The treatment effect vector $\beta^T = [\beta_1\ \beta_2\ \beta_3\ \beta_4]$, the nonparametric function ϕ and the error variance σ^2 are estimated using the method given in Section 2. Tuckey's bi-weight function with the turning point c=4.685 is used as the robustness function. The Average Mean Squared Errors (*AMSE*) of the estimated values of σ , β, ϕ and **m** with the true values of 100 sets of simulated data for different values of *n* (100, 200, 400) and σ (1.0, 2.0) are given in Table A.2. The *AMSE* of the estimated parameters are calculated as follows:

$$AMSE \text{ of } \hat{\sigma} = \frac{1}{100}\sum_{i=1}^{100}\left(\sigma - \hat{\sigma}_i\right)^2,$$

$$AMSE \text{ of } \hat{\beta} = \sum_{j=1}^{4}\frac{1}{100}\sum_{i=1}^{100}\left(\beta_j - \hat{\beta}_{ji}\right)^2$$

$$AMSE \text{ of } \hat{\phi} = \frac{1}{100}\sum_{i=1}^{100}\frac{1}{n}\sum_{j=1}^{n}\left\{\phi(u_j) - \hat{\phi}_i(u_j)\right\}^2,$$

$$AMSE \text{ of } \hat{\mathbf{m}} = \frac{1}{100}\sum_{i=1}^{100}\frac{1}{n}\sum_{j=1}^{n}\left\{m - \hat{m}_i(u_j)\right\}^2$$

where, $\hat{\sigma}_{(i)}$, $\hat{\beta}_{j(i)}$, $\hat{\phi}_{(i)}$ and $\hat{m}_{(i)}$ are the estimated values of σ , β_j, ϕ and the regression function m corresponding to the i^{th} simulated data set. The bias of the point estimates of $\hat{\sigma}$, $\hat{\beta}_j$, j = 1,…,4 are calculated as follows

$$\text{Bias of } \hat{\sigma} = \frac{1}{100}\sum_{i=1}^{100}\left(\sigma - \hat{\sigma}_{(i)}\right)$$

$$\text{Bias of } \hat{\beta}_j = \frac{1}{100}\sum_{i=1}^{100}\left(\beta_j - \hat{\beta}_{j(i)}\right), j=1,\ldots,4$$

The *AMSE* of the estimates are converging to zero as *n* increases or in other words, the estimated values are converging to the true values as *n* increases. Note that the bias of the point estimates $\hat{\sigma}$, $\hat{\beta}_j$, j = 1,…,4 are also negligible as *n* increases (Table A.2). This indicates the consistency of the estimates. The *MSE* varies with change in the choice of bandwidths. The optimum bandwidth (bandwidth corresponds to the minimum *MSE*) depends on the curvature of the function. The optimum bandwidth for estimating the parameters of the model was obtained based on the robust cross validation technique given in Section 2.

The comparison of Average Mean Squared Errors (*AMSE*) of the estimated values of σ , β, ϕ and **m** with the true values of 100 sets of simulated data for different values of *n* (100, 200, 400) and σ (1.0, 2.0) showed that in the presence of outliers (4% and 8%) the robust method performs much better than the non-robust method. The value of *AMSE* decreases as *n* increases or in other words the estimates converges to the true value.

Acknowledgements

The authors express their gratefulness to the anonymous reviewer for giving some useful suggestions that led to considerable improvement in the presentation of the results. The authors also express their thanks to the Chief Editor for his suggestions on restructuring the contents for better expression.

References

Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). An error-component model for prediction of country crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83**, 28-36.

Chandra, H., Salvati, N. and Sud, U. C. (2011). Disaggregate-level estimates of indebtedness in the state of Uttar Pradesh in India – An application of small area estimation Technique. *Journal of Applied Statistics*, **38**, 2413–2432.

Chandra, H. (2013). Exploring spatial dependence in area-level random effect model for disaggregate-level crop yield estimation. *Journal of Applied Statistics*, **40**, 823-842.

Chandra, H., Salvati, N. and Chambers, R. (2017). Small area prediction of counts under a non-stationary spatial model. *Spatial Statistics*, **20**, 30-56.

Cleveland, W. S. and Devlin, S. J. (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, **83**(**403**), 596–610.

Fay R. E. and Herriot R. A. (1979). Estimation of income from small places: an application of James-Stein procedures to census data.*Journal of the American Statistical Association*,74, 269-277.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics - The Approach Based on Influence Functions*. New York: John Wiley and Sons.

OEIS Foundation Inc. (2020). Number of Hadamard matrices of order 4*n*. *The On-Line Encyclopedia of Integer Sequences*. http://oeis.org/A007299.

Rao J. N. K. and Molina I. (2015). *Small Area Estimation*. 2nd Edition. John Wiley and Sons.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics*, **22**, 1346-1370.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading: Addison-Wesley.

Weisstein, E. W. Hadamard's Maximum Determinant Problem. *MathWorld–A Wolfram Web Resource*. Retrieved 18 July 2020.

https://mathworld.wolfram.com/HadamardsMaximumDeterminantProblem.html

Wikipedia contributors. Hadamard's maximal determinant problem. *Wikipedia, The Free Encyclopedia*. Retrieved 18 July 2020.

Appendix A

Table A.1: Optimum bandwidth ad *AMSE* of the estimates in the simulation study

σ	Outliers (%)	*n*	*h*	AMSE ($\hat{\beta}$)		AMSE ($\hat{\phi}$)		AMSE ($\hat{m}$)		AMSE (σ)	
				SP	Robust SP	SP	Robust SP	SP	Robust SP	SP	Robust SP
1	0	100	0.30	0.1486	0.1471	0.0796	0.0541	0.0731	0.0664	0.0064	0.0182
		200	0.20	0.0720	0.0740	0.0309	0.0256	0.0398	0.0389	0.0030	0.0078
		400	0.15	0.0315	0.0341	0.0167	0.0148	0.0215	0.0218	0.0011	0.0033
	4	100	0.30	0.3712	0.1741	0.0946	0.0608	0.1835	0.0810	0.3194	0.0143
		200	0.25	0.1656	0.0849	0.0522	0.0308	0.1041	0.0408	0.3089	0.0080
		400	0.15	0.0774	0.0384	0.0459	0.0162	0.0944	0.0241	0.2948	0.0038
	8	100	0.25	0.5138	0.1793	0.2440	0.0594	0.4018	0.0846	0.8825	0.0180
		200	0.25	0.2285	0.0985	0.1704	0.0285	0.2723	0.0439	0.8511	0.0133
		400	0.25	0.1250	0.0548	0.1440	0.0179	0.2222	0.0231	0.8331	0.0121
2	0	100	0.30	0.5399	0.5672	0.1913	0.1591	0.2631	0.2588	0.0188	0.0644
		200	0.30	0.3009	0.3142	0.1144	0.0887	0.1451	0.1452	0.0101	0.0415
		400	0.25	0.1463	0.1530	0.0543	0.0401	0.0689	0.0674	0.0043	0.0125
	4	100	0.30	1.4690	0.6348	0.3745	0.1797	0.7428	0.2981	1.2803	0.0432
		200	0.30	0.5670	0.2865	0.2134	0.0931	0.3958	0.1386	1.2292	0.0281
		400	0.20	0.3239	0.1418	0.2455	0.0509	0.3576	0.0810	1.1912	0.0224
	8	100	0.30	1.7699	0.6486	1.0668	0.1936	1.6062	0.3138	3.3692	0.0716
		200	0.30	0.9476	0.3266	0.6910	0.0900	1.0818	0.1394	3.3647	0.0584
		400	0.25	0.4338	0.1461	0.7617	0.0442	0.9799	0.0709	3.3693	0.0576

SP: Semi- parametric

Table A.2: Bias of the robust point estimates in the simulation study

σ	Outliers (%)	*n*	*h*	Bias of				
				$\hat{\beta}_S$	$\hat{\beta}_5$	$\hat{\beta}_ò$	$\hat{\beta}_\mu$	$\hat{\sigma}$
1	0	100	0.30	–0.0184	–0.0187	0.0201	0.0169	–0.0610
		200	0.20	–0.0182	–0.0157	0.0155	0.0185	–0.0282
		400	0.15	0.0026	–0.0053	0.0127	–0.0100	–0.0187
	4	100	0.30	–0.0232	0.0100	–0.0010	0.0143	0.0440
		200	0.25	–0.0161	0.0112	–0.0002	0.0051	0.0346
		400	0.15	0.0033	0.0019	–0.0004	–0.0048	0.0294
	8	100	0.30	–0.0002	0.0166	–0.0077	–0.0088	0.0851
		200	0.25	–0.0008	0.0093	–0.0084	0.0000	0.0789
		400	0.20	0.0010	0.0039	–0.0058	0.0010	0.0745
2	0	100	0.30	–0.0060	–0.0079	0.0055	0.0085	–0.1037
		200	0.25	–0.0090	0.0234	–0.0156	0.0013	–0.0586
		400	0.20	–0.0030	0.0040	–0.0060	0.0051	–0.0398
	4	100	0.30	0.0222	–0.0271	–0.0120	0.0168	–0.0558
		200	0.25	0.0131	–0.0135	–0.0110	0.0114	0.0490
		400	0.20	–0.0060	0.0184	–0.0088	–0.0037	0.0445
	8	100	0.30	–0.0132	0.0294	–0.0131	–0.0031	0.1248
		200	0.30	–0.0143	0.0292	–0.0061	–0.0089	0.1160
		400	0.25	–0.0117	0.0202	–0.0106	0.0020	0.1119

Table A.3: Estimated values with standard errors (weight of nuts) of the field data

Parameter	Linear		Semi-parametric		Robust Semi-	
	Estimate	SE	Estimate	SE	Estimate	SE
$\mu+\beta_1$	9.969	0.683	9.924	0.622	9.925	0.548
$\mu+\beta_2$	9.414	0.683	9.570	0.626	9.573	0.552
$\mu+\beta_3$	10.029	0.638	9.949	0.594	9.950	0.524
$\mu+\beta_4$	9.883	0.675	9.994	0.617	9.991	0.543
$\mu+\beta_5$	9.922	0.691	9.918	0.636	9.916	0.560
$\mu+\beta_6$	10.767	0.630	10.758	0.587	10.758	0.517
σ	4.317		4.312	-	3.803	-

μ: Overall mean

Table A.4: Estimated values with standard errors (number of nuts) of the field data

Parameter	Linear		Semi-parametric		Robust Semi-parametric	
	Estimate	SE	Estimate	SE	Estimate	SE
$\mu+\beta_1$	328.80	22.71	331.96	20.71	330.80	16.83
$\mu+\beta_2$	307.84	22. 70	308.87	20.70	308.85	16.83
$\mu+\beta_3$	331.13	21.12	334.87	19.62	336.45	15.95
$\mu+\beta_4$	332.86	22.40	336.57	20.45	337.32	16.63
$\mu+\beta_5$	324.32	22.91	315.21	21.27	313.69	17.29
$\mu+\beta_6$	370.57	20.87	374.34	19.35	374.71	15.72
σ	143.06		142.96		116.16	

μ: Overall mean

Appendix-B

Figure B.1: Estimated survival functions of fitted models—Chemotherapy group

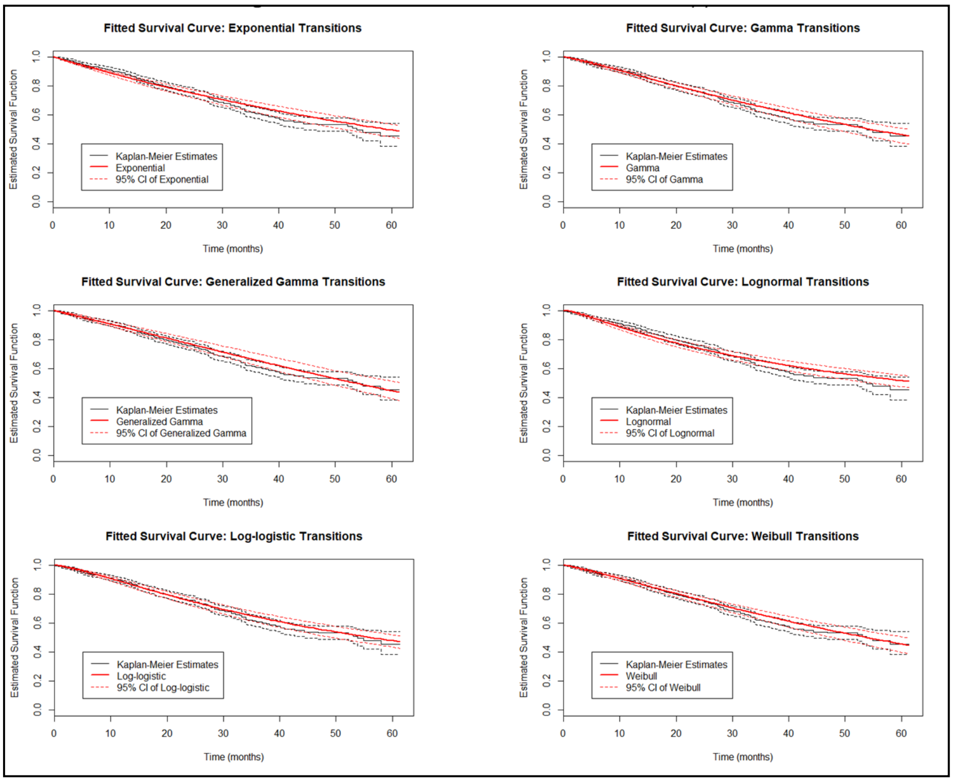
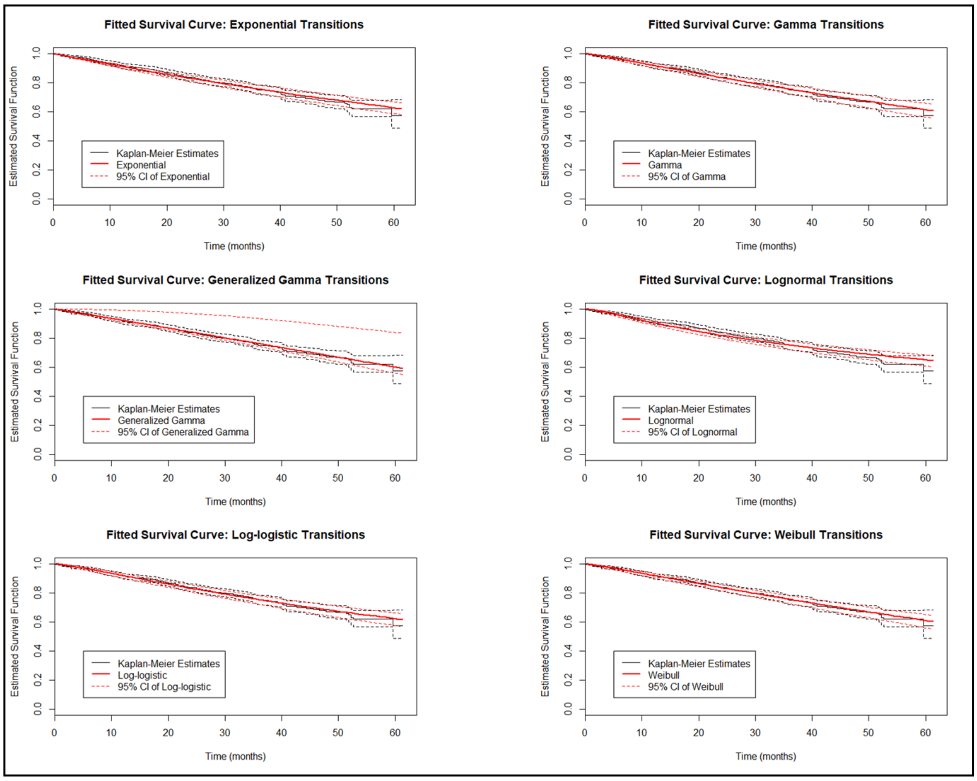
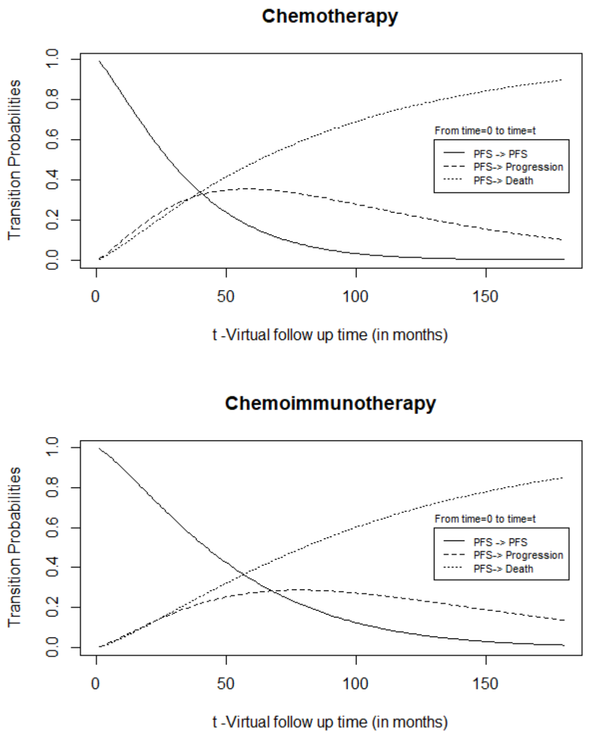


Figure B.2: Estimated survival functions of fitted models—Chemoimmunotherapy group



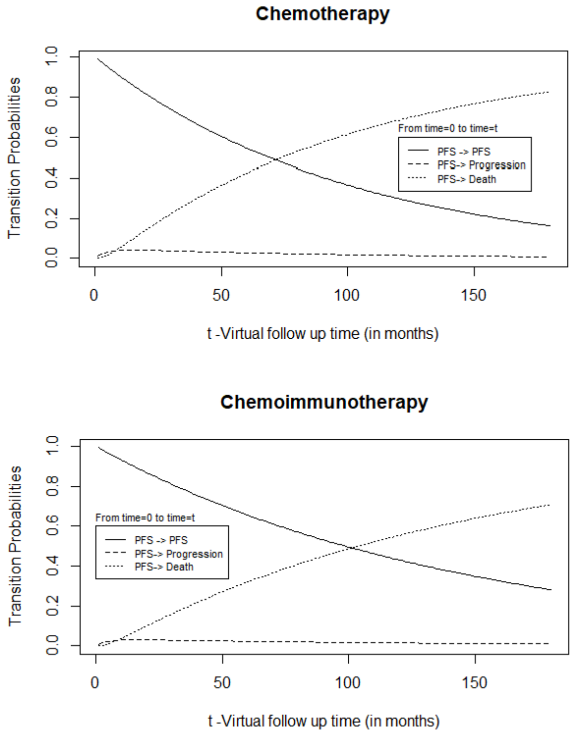
Appendix-C

Figure C.1: Transition probabilities plotted against time—Weibull semi-markov model


Chemotherapy

Chemoimmunotherapy

Figure C.2: Transition probabilities plotted against time—Multinomial-Dirichlet Bayesian model


Chemotherapy

Chemoimmunotherapy