Small Area Estimation Using GIS

Roma Choudhury Sahu, Kasturi Basu and Shibdas Bandyopadhyay
Applied Statistics Unit, Indian Statistical Institute, Kolkata - 700108

Summary

Small area statistics are useful for local level planning. GIS techniques are being increasingly used in small area estimation. NSSO data on fsus (villages within a district, i.e., NSSO stratum) are ideal for this purpose. But, so far, attempt has not been made to use GIS technique on NSSO data. One advantage of GIS method is that, unlike traditional regression method of estimation, no auxiliary information is required.GIS method has been applied in this paper to obtain block level estimates on the basis of 1991 census data as an illustration. It was observed that block level estimates based on GIS turned out to have less error than estimates based on regression method.

1. Introduction

Small areas for this study are small geographic areas. Of several model-based techniques, GIS (Geographical Information System) is being increasingly used for small area estimation [Cai (2004)[1], Heady (1998)[2]]. GIS models essentially assume that nearer areas are more likely to be similar than areas that are further apart. GIS does not require information on any auxiliary variable.

With GIS, one may get estimates for all villages (NSSO fsu's) within a district (NSSO stratum) using NSSO data on selected villages within the district; one may also compute block level estimates for all blocks in the district. NSSO data are ideal for small area estimation using GIS. We first illustrate the use of GIS for small area estimation. Then we shall discuss GIS methodology.

A key component in using GIS is the availability of a digitized map of the boundaries of areas (e.g., NSSO fsu's) of interest within a geographical zone (e.g., NSSO stratum) of interest.

To see if GIS based estimates are any good, we obtained a separate set of estimates based on linear regression equations, a standard method of estimation when an auxiliary variable is available.

Algebraic comparisons of standard errors of small area estimates based on GIS and regression equations will not be attempted here. Instead, we shall use known populations for comparisons and validation.

2. Illustration

For illustration, we shall obtain block-level estimates within a district; we have selected Nadia district in the State of West Bengal and digitized the 17 block boundaries (see Fig. 1).

FLR (Female Literacy Rate) i.e., [{No. of Female Literates / (No. of Female - No. of Female in the age group 0-6)}* 100] at the block level is taken as study variable. FSCSTR (per cent SC/ST among female population in a block) is taken as an auxiliary variable. 1991 census data from 1991 District Primary Census Abstract[3], are used to compute (see Table 1) FLR and FSCSTR and taken as true values.

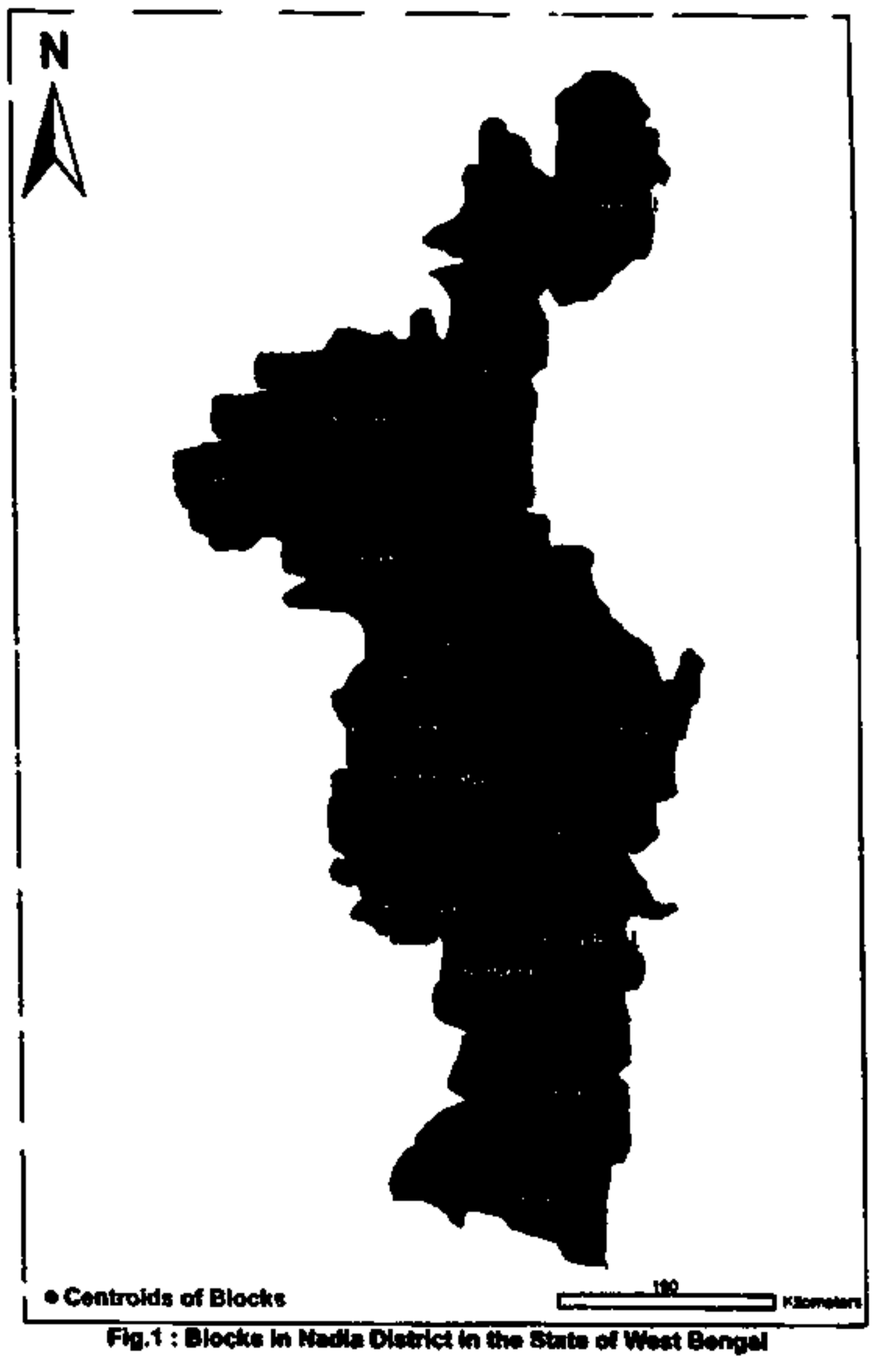

Fig.1 : Blocks in Nadia District in the State of West Bengal

Table 1: 1991 Census Data for 17 Blocks of Nadia District in West Bengal

Block Id No		No of Female	No of Female in the age group (o-6) years	No of Female Scheduled Cast	No of Female Scheduled Tribes	No of Female Literates	Female Literacy Rate (%)*	Female SC/ST Rate (%)**
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	Karimpur-I	69181	12549	12531	1234	21148	37.34	19.90
2	Karimpur-II	78716	15776	13227	1009	17639	28.02	18.08
3	Tehatta-I	89711	16083	31255	1051	28077	38.13	36.01
4	Tehatta -II	55457	9864	9051	294	15669	34.37	16.85
5	kaliganj	114758	22584	20351	786	28410	30.82	18.42
6	Nakashipara	131512	25873	31201	3016	32850	31.10	26.02
7	Chapra	110052	21583	20148	441	27809	31.43	18.71
8	Krishnagar-II	51448	9184	9875	255	14638	34.63	19.69
9	Krishnaganj	54605	9209	24534	2808	18912	41.66	50.07
10	Krishnagar-I	107808	19506	38190	5177	32611	36.93	40.23
11	Nabadwip	50500	9923	7773	455	13944	34.36	16.29
12	Hanskhali	104074	16761	50867	2547	40709	46.62	51.32
13	Santipur	81184	14873	30971	3014	25469	38.41	41.86
14	Ranaghat-I	90532	14433	27952	805	37482	49.25	31.76
15	Ranaghat-II	136664	22824	62725	5389	55783	49.00	49.84
16	Chakdah	145580	23713	63218	7751	58612	48.09	48.75
17	Haringhata	85530	14733	26952	4429	32213	45.50	36.69

(Source : District Primary Census Abstract, Nadia, West Bengal, 1997)

*** Female Literacy Rate = [No. of Female Literates / { No. of Female – No. of Female in the age group (0-6)}]*100**

**** Female SC & ST Rate = {(No. of Female SC + No. of Female ST)/ No. of Female }*100**

2.1 No auxiliary information, GIS technique

Table 2.1 : 10 Randomly Selected Blocks

Block Id No	1991 Census	
	FLR	FSCSTR
(1)	(2)	(3)
3	38.13	36.01
4	34.37	16.85
5	30.82	18.42
8	34.63	19.69
9	41.66	50.07
10	36.93	40.23
11	34.36	16.29

We selected at random 10 of the 17 blocks (at least 10 data points are needed to apply GIS), copied the FLR values for the 10 selected blocks from Table 1; Table 2.1 gives the Id of selected blocks in col. (1) and FLR values in col. (2).

We now wish to estimate FLR values for the remaining 7 block using GIS.

Table 2.2 gives the Id of the remaining 7 blocks in col. (1) and GIS predicted FLR in col. (2); GIS model based standard errors of GIS estimates are given in col. (3), true FLR are given in col. (4) (copied from Table 1, col. (8)) and GIS residuals (difference between true and GIS predicted) are given in col. (5). Average squared error (sum of squares of residuals in col. (5) divided by 7) is 24. Square root of the average squared error, v24 = 4.90, compares well with the GIS model-based standard errors in col. (3)

Table 2.2 : Prediction of Female Literacy Rate for 7 Blocks using GIS

Block Id No	Female Literacy Rate				Observed Female SCST Rate	Regression Predicted	Regression Residuals
	GIS Predicted	Standard Error	Observed *	GIS Residuals			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8).
1	37.93	6.71	37.34	-0.59	19.90	33.75	3.59
2	37.43	5.68	28.02	-9.41	18.08	33.15	-5.13
6	33.44	3.59	31.10	-2.34	26.02	35.78	-4.68
7	37.12	3.76	31.43	-5.69	18.71	33.35	-1.92
14	43.82	3.48	49.25	5.43	31.76	37.68	11.57
15	46.61	4.05	49.00	2.39	49.84	43.66	5.34
17	47.96	4.03	45.50	-2.46	36.69	39.31	6.19
	Average Squared Error			24.00			37.93

* treated as unknown but used for cross validation

Set No. 1: Id. no. of randomly selected blocks : 3,4,5,8,9,10,11,12,13,16

Regression Equation : FLR =27.16 + (0.33) * FSCSTR

2.2 With auxiliary information, Regression technique

For the 10 selected blocks as in section 2.1, FSCSTR values are copied in col. (3) in Table 2.1 from Table 1,col. (9).

The linear regression of FLR on FSCSTR, based on the 10 pairs of observations for the 10 selected blocks is:

FLR=27.16+0.33 * FSCSTR (1)

The correlation coefficient between FLR and FSCSTR is positive and 0.87.

Using the regression equation, FLR is predicted for the remaining 7 blocks.

Table 2.2 gives true FSCSTR values in col. (6) (copied from Table 1), predicted FLR values, based on regression equation (1), in col. (7) and regression residuals (difference between true and regression predicted) in col. (8). The average squared error (sum of squares of residuals in col. (8) divided by 7) is 37.93.

2.3 Comparison

Based on the 10 randomly selected blocks, GIS technique appears to be better than regression technique since the per cent relative efficiency of GIS compared to regression is (37.93/24)*100% or 158%.

However, it is only for one of $^{17}C_{10}$ possible sets of 10 blocks. So, we repeat the same exercise for another 8 randomly selected sets of 10 blocks. The results are given in Table 3 : Col. (1) gives the set number, Id of randomly selected blocks in col. (2), average squared errors in col. (3) and col. (4) respectively for GIS and regression techniques; in addition, regression intercepts and slopes are given in col. (5) and col. (6) along with correlation coefficient in col. (7). With high correlation coefficient values (0.74 to 0.87) across the 9 random selections, it was expected that regression technique would be good, but GIS did better in most cases, 8 out of 9 cases.

On an average, based on 9 replications, the relative efficiency of GIS technique is (29.48/19.06)*100 = 155% as compared to regression technique. Moreover, one does not need any auxiliary variable to apply GIS technique.

3. Methodology

3.1 GIS technique

For using GIS technique, a digitized map is needed. Using

data on atleast 10 points a surface is fitted (see appendix) over the entire study area, including areas with data and without data. Predicted value at a location is the height of the surface at the location (see appendix).

We are treating centroids of blocks as the representative of blocks, both for fitting the surface and also for prediction.

After digitization of the district of Nadia, along with its 17 block boundaries, we used GIS package for computation of the centroids (see Fig.1). Data for blocks are then attached to respective block centroids. Then a surface is fitted over the district. Basic GIS assumption is that the neighbouring blocks influence the study variable more than the far away blocks. Of several GIS techniques, we used ordinary kriging (see appendix) for fitting a surface (see Fig.2 for fitted surface using set 1). A predicted value of a block is the height of the surface at its centroid. The fitted surface always passes through the true values used for fitting the surface.

3.2 With auxiliary information

A linear regression equation is fitted with the help of ordinary least squares method based on blocks for which both the study variable and the auxiliary variable are available. The predicted intercepts and predicted slopes for the 9 sets are given in col. (5) and col. (6) of Table 3 respectively. The predicted value of a block is the corresponding intercept plus the slope multiplied by the corresponding value of the auxiliary variable of that block. For this method no digitized map of the study area is necessary; also, we do not need to attach data to the centroids.

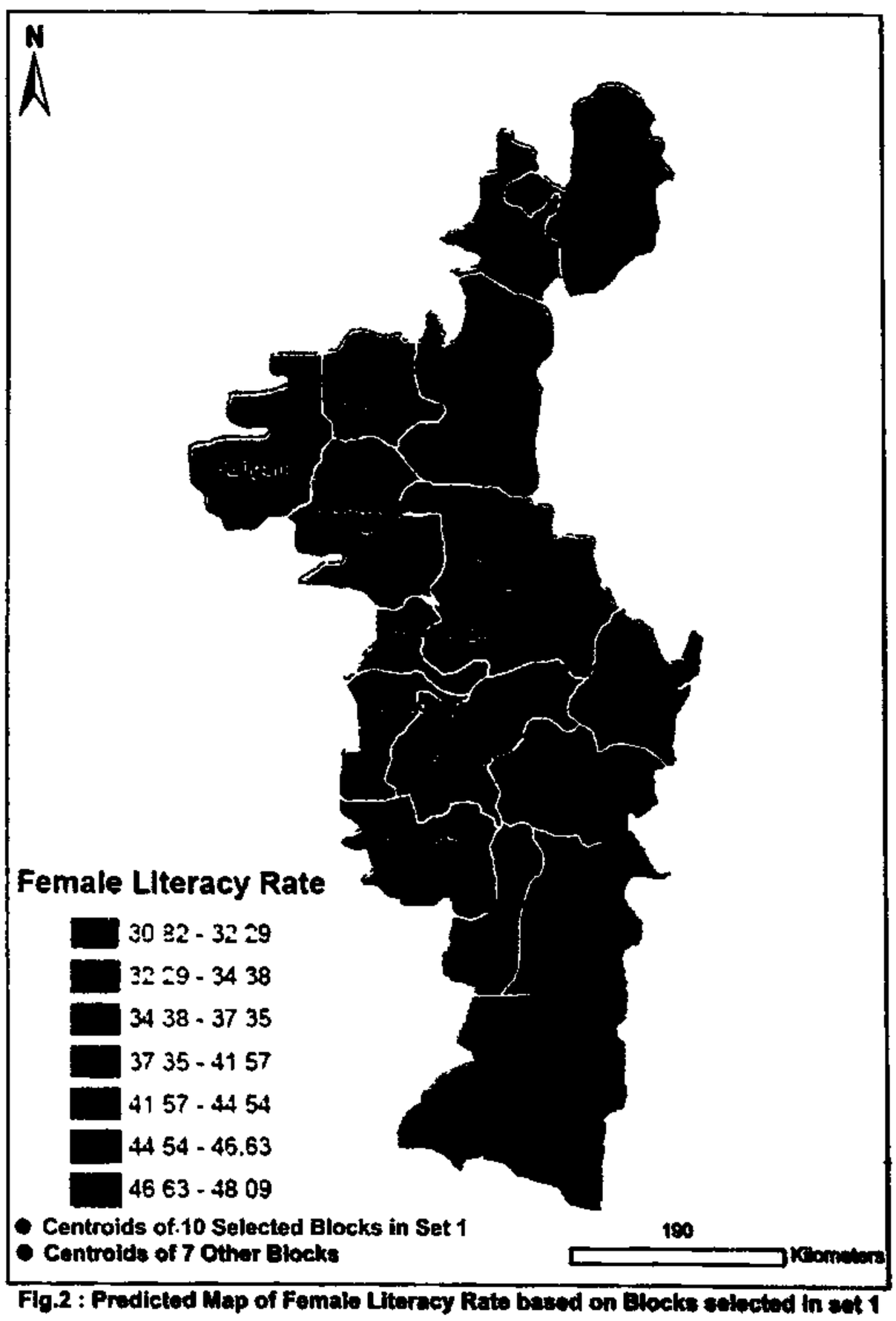

Fig.2 : Predicted Map of Female Literacy Rate based on Blocks selected in set 1

Table3 : Comparison of Error Rates in 9 Randomly Selected Sets of 10 of 17 blocks

Set No	Randomly Selected Blocks with Identification No	Average Squared Error *		Regression		Correlation Coefficient
		GIS	Regression	Intercept	Slope	
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	3,4,5,8,9,10,11,12,13,16	24.00	37.93	27.16	0.33	0.87
2	1,2,3,4,7,8,9,11,13,15	12.64	33.62	26.93	0.34	0.84
3	1,4,7,8,9,11,12,13,14,15	15.45	26.08	29.21	0.33	0.77
4	1,3,5,6,7,8,9,10,15,16	19.19	31.39	24.42	0.41	0.86
5	2,3,6,7,8,9,11,14,15,16	13.88	14.79	24.14	0.46	0.80
6	1,4,5,7,8,10,12,13,15,17	16.78	28.79	26.31	0.39	0.85
7	1,2,3,4,6,7,10,12,13,15	7.42	33.44	23.88	0.42	0.85
8	2,6,7,8,10,12,13,14,16,17	23.28	11.97	23.10	0.48	0.75
9	2,4,6,7,8,9,10,11,13,17	38.91	47.28	26.96	0.30	0.74
Average of Average Squared Error		**19.06**	**29.48**			

*** See table 2.2 for computation of Average Squared Error for Set No 1.**

3.3 Error in prediction

Error is the difference between the true value and the predicted value (see Table 2.2). Average squared error is an average of the 7 squared errors.

4. Appendix

Let (x, y, z) be the three coordinates of a point in 3-dimension. Further let h be a function defined for every point (x, y) in the X-Y plane, and z = h(x, y). As (x, y) varies over X-Y plane, z = h(x, y) gives a surface over X-Y plane. $z_0 = h(x_0, y_0)$ is the height of the surface from the point (x_0, y_0) on the X-Y plane.

For example, let x and y be the latitude and longitude of a location on earth. Let z be the altitude (height from sea level) at the location (x, y). For a given location (x_0, y_0) on earth, $z_0 = h(x_0, y_0)$ is its height from sea level. Similarly, z could be temperature or precipitation etc. The function h is usually unknown; it is not possible to find the altitude/temperature/precipitation at a location just with the knowledge of its latitude and longitude.

The problem of fitting of a surface is essentially a problem of estimating the function h with n known (x, y, z) values, i.e., with known z values at a number of known locations (x_i, y_i), i = 1, 2, ... , n. The problem of fitting of a surface is much more complex than standard regression problem since the z-values are correlated across locations. There are several methods for fitting of a surface, one of which is ordinary kriging.

Ordinary kriging has two distinct tasks: quantification of the spatial structure (the correlation across locations) of observed z-values at known locations and computation of predicted values at locations where z-values have not been observed. Quantification of spatial structure is known as variography that fits a spatial dependence model to the observed z-values. Then the variography model is used, together with observed z-values, for prediction.

Data given are latitudes and longitudes of n locations (x_i, y_i), i = 1, 2, ..., n for which n z values are known. In ordinary kriging, one z-value is estimated for a location of interest; then another z-value is estimated for another location of interest; z-values are estimated one at a time. This way all z-values are estimated for all locations (x, y) over a geographic area of interest, yielding a surface over the geographic area of interest. Like every other method of fitting a surface, ordinary kriging is very computer intensive. Thompson (1992)[4] illustrates a step-by-step estimation of one z-value.

References

[1]Qiang Cai, (2004): " A Methodology For Estimating Small-Area Population By Age And Sex Based on Methods of Spatial Interpolation and Statistical Inference", Department Of Geography, the University Of Iowa, Iowa City, IA 52242, www. ucgis.org/ ucgisfall2004/studentpapers/files/cai.pdf

[2]Patrick Heady, Phil Clarke, Bruce Mitchell, (1998): "GIS an Aid to Small Area Estimation: Sorting and Linking Survey Data for Statistical Modelling in the presence of Alternative Boundaries", Statistical Commission and Economic Commission for Europe Conference of European Statisticians, Work Session on Geographical Information Systems (Ottawa, Canada, 5-7 October 1998), Working Paper No. 9, Submitted by Office for National Statistics, United Kingdom, www.unece.org/stats/documents/1998/10/gis/9.e.pdf

[3]District Primary Census Abstract, Nadia District, (1997): Directorate of Census Operations, West Bengal, Census of India 1991, Series –26 , West Bengal , Part XII – B, District Census Handbook, West Bengal ,Nadia, Village & Townwise Primary Census Abstract

[4]S K Thompson, (1992): Sampling , Wiley