Incomplete Adaptive Cluster Sampling Design

Chang-Tai Chao, National Cheng-Kung University
Steven K. Thompson, The Pennsylvania State University
Chang-Tai Chao , Department of Statistics, School of Management,
National Cheng-Kung University, Tainan, Taiwan 70101 (ctchao@stat.ncku.edu.tw)

Key Words: Adaptive Cluster Sampling Design, Graph Sampling, Adjacency Matrix,Design-based sampling, Restricted adaptive cluster sampling, Framed spatial population.

1 Introduction

In spatial sampling, the research area often is divided into small rectangular grids and these grids are considered as the sampling units. Consider an adaptive cluster sampling type of design applied to a framed spatial population with N units, denoted by $1, \ldots\ldots, N$. If the inclusion of unit i can lead to the inclusion of units j which is in the *neighborhood* of i, then we define an *arc* (directed) which starts from i and ends at j, denoted by the ordered pair (i,j). Therefore, we can consider our population as a *graph* which contains a vertex set $V = \{1, \ldots\ldots, N\}$, which is the collection of all units in this population, and an arc set $E = \{(i,j)|i,j \in V\}$, a set of ordered pairs describing the connection between the population units.

$$\begin{aligned} V &= \{1, \ldots\ldots, N\} \\ E &= \{(i,j)\,|\,i,j \in V\} \\ G &= \{V, E\} \end{aligned}$$

Further, we can define an *adjacency matrix* $\mathcal{A}$ to represent this graph, where $\mathcal{A}$ is an N by N matrix and the elements a_{ij} in $\mathcal{A}$ is (e.g. Foulds [1992])

$$a_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{Otherwise} \end{cases}$$

Additionally, we also define $a_{ii} = 0, \forall i \in V$, that is, the diagonal of $\mathcal{A}$ is 0.

In the general setup of graph sampling, a_{ij}'s are considered as random numbers or unknown constants with possible values of 0 or 1. This can be simplified in the framed spatial population because the natural framed population structure. After the neighborhood has been defined in a framed spatial population, since the arc from i to j (and from j to i) exists only if i and j are neighboring to each other, therefore we can conclude that

$$\text{if } i \text{ and } j \text{ are not neighboring,} a_{ij} = a_{ji} = 0, \ \forall i,j \in V$$

Notice that this information can be gathered before any sample has been selected and observed. Therefore, generally most of the elements a_{ij} in $\mathcal{A}$ can be determined as 0 before the sampling procedure.

The definition of in-degree of a unit i is the total number of units from which we can reach i. Therefore, the in-degree of unit i is the column sum of the i_{th} column of $\mathcal{A}$

$$\text{in-degree of } i = \sum_{j=1}^{N} a_{ji} = a_{.i}$$

Similarly, the out-degree of unit i is the number of units which can be reached from i. Therefore, the out-degree of i is the row sum of the i_{th} row of $\mathcal{A}$

$$\text{out-degree of } i = \sum_{j=1}^{N} a_{ij} = a_{i.}$$

Additional information which can also be known after the neighborhood has been defined is the maximum values of the in-degree and out-degree of each i, which are both equal to the number of units in the neighborhood of i.

The out-degree of a unit which is in the final sample of ACS can be determined exactly after the sample has been observed. In ACS, all of the units in the neighborhood of a unit which have been observed and its associated value of variable of interest satisfies C, then all the units in its neighborhood are supposed to be included in the sample as well. The out-degree of the units which has been observed can be determined as

Support for this research was provided by the National Science Foundation, grant DMS-9626102, and the National Institutes of Health, NIDA, grant RO1 DA09872.

$$a_{i.} = \begin{cases} \text{number of neighboring units of } i, & \text{if } y_i \in C \\ 0, & \text{otherwise} \end{cases}$$

For the purpose of constructing a Horvitz-Thompson estimator for the population quantity of interest, on the other hand, the essential element in the graph sampling is not the out-degree but the in-degree of the sampling units. However, the in-degrees of the units that does not satisfy C can not be determined in ACS with the final data. Consequently, the inclusion probabilities of these units can not be determined. Thompson [1990] discarded these units which are not initially selected and used the initial intersection probability instead of the actual inclusion probability to obtain a modified Horvitz-Thompson type unbiased estimator. This technique is effective only if every units in the initially selected networks can be included into the final sample, though.

Also, we can conclude the following lemma by the property of ACS

Lemma 1 *$(i, j) \in E \Leftrightarrow y_i \in C$ and i,j are neighboring.*

2 Incomplete Adaptive Cluster Sampling Design

As in the usual finite population sampling, we consider that the population consists of N units labeled from $1, \ldots\ldots, N$. For each unit $i \in \{1, \ldots\ldots, N\}$, there is a associated value of variable of interest y_i. An initial sample of size $n < N$ is sampled by simple random sampling without replacement, denoted by $s^{(0)} = (s_1^{(0)}, \ldots\ldots, s_n^{(0)})$. Also, the collection of units which are not included in the initial sample is defined as $s^{(0)'} = \{1, \ldots\ldots, N\} \setminus s^{(0)}$. After $s^{(0)}$ has been observed, if any of $y_{s_i^{(0)}} \in C$, then all the units which are in the neighborhood of $s_i^{(0)}$ will be added into the sample and observed. Define the collection of such new sampling units which are not in $s^{(0)}$ as $s^{(1)}$. The same adaptive procedure will apply to $s^{(1)}$ as well to get another set of new sampling units which is denoted as $s^{(2)}$, and so on. Notice that $s^{(i)} \cap s^{(j)} = \varnothing$ if $i \neq j$ and

$$s^{(m)'} = \{1, \ldots\ldots, N\} \setminus \bigcup_{i=1}^{m} s^{(i)}$$

The sampling procedure will be stopped if either of the following conditions happens

1. No new unit can be added into the sample. This is the original stopping rule of ACS.
2. $s^{(M)}$ has been sampled and observed, where $M \geq 0$ is a pre-specified positive integer. That is, even if $\exists i \in s^{(M)}$ such that $y_i \in C$, the units in the neighborhood of i which belong to $s^{(M)'}$ will not be added into the sample.

The number M is the maximum number of *steps* in an IACS. Notice that if $M = 0$, then it is the design which is the same as the conventional design that is used for the initial selection. The possible edge units and networks which would be encountered in this IACS are defined as what has been defined in the original ACS of Thompson [1990].

3 Properties of IACS

In the general graph theory, a *path* is defined as a *walk* from unit (vertex) i_1 to i_n

$$[i_1, \ldots\ldots, i_{n-1}, i_n]$$

such that all the units in this walk are distinct. Consider there is an adaptive adding process which is initiated by unit i_0, that is, $i_0 \in s^{(0)}$ and $y_{i_0} \in C$. Suppose that a unit $j \notin s^{(0)}$ is included into the sample because of this adding process which is initiated by i_0 and $j \in s^{(n)}$. If the walk which starts from i_0 to j under this IACS design is

$$[i_0, i_1, \ldots\ldots, i_{n-1}, j],$$

then this walk is a path.

Lemma 2 *Given the condition above,*

$$[i_0, i_1, \ldots\ldots, i_{n-1}, j]$$

, is a path

Define the *length* of a path is the number of the units after i_0 or before j. For example, the length of the path $[i_0, i_1, \ldots\ldots, i_{n-1}, j]$ is n. In an IACS design with maximum number of steps M, we define the existence of a path as

Definition 1 *A path $[i_0, i_1, \ldots\ldots, i_{n-1}, j]$, $n \leq M$, exists if and only if all of the following conditions are satisfied*

1. *$y_{i_k} \in C$,$\forall k \in \{0, \ldots, n-1\}$ and*
2. *i_k and i_{k+1}, $\forall k = 0, \ldots, n-1$ and i_{n-1}, j are neighboring.*
3. *i_0 is selected into the initial sample.*

A new variable l_{ij} which represents the length between the units i and j as

Definition 2 *If the path $[i, \ldots\ldots, j]$ exists, then l_{ij}= the length of the path from i to j.*

Obviously, $l_{ii} = 0$ and the maximum value of j_{ij} is the pre-specified maximum number of steps M. Also, according to the design in which whenever a encountered unit satisfying C, then all of the units in its neighborhood will be included as well, l_{ij} will be the length of the shortest path which starts from i to j. Notice that this path may not be unique but the value of l_{ij} is. On the other hand, the existence of the path that starts from i to j does not guarantee the existence of the path starting from j to i. However, if such path does existence, then $l_{ij} = l_{ji}$.

Lemma 3 *Under the design, if the path starting from i to j exists and $j \in s$ then*

1. *The path starting from j to i exists as well and,*
2. *$l_{ij} = l_{ji}$*

We can also give the necessary and sufficient condition for a unit j to be included into the final sample, which is denoted by s_{final}, under the IACS design with the maximum number M

Lemma 4 *$j \in s_{final} \Leftrightarrow \exists i \in s^{(0)}$ such that $\exists$ a path which starts from i to j exists and $l_{ij} \leq M$*

In fact, the condition of $i \in s^{(0)}$ is rather redundant since it is one of the necessary conditions for a path which starts from i exists. A unit which is not initially selected can be included into the final sample due to more than one initially selected units. Define the collection of units which can be included if unit i is initially selected as $s_{(i)}$. Obviously, $l_{ji} \leq M, \forall j \in s_{(i)}$ if the path starting from j to i exists. Also notice that for $i, j \in s$, $s_{(i)}$ and $s_{(j)}$ are not necessarily to be disjoint, but we may like to arrange our initial sample in such way that

$$s_{(i)} \cap s_{(j)} = \varnothing, \forall i, j \in s, i \neq j$$

in practice.

4 Inclusion Probability

The inclusion probabilities for the units in the same network are the same in the original ACS (e.g. Thompson [skt1990]) but different in IACS. One can imagine that the border units will have smaller inclusion probabilities than others. In fact, the inclusion probability in IACS needs to be considered separately for each unit which is included in the final sample. For an IACS design with the maximum number of steps M, the necessary and sufficient condition for a unit to be included into the final sample is given in Lemma 4.

Assume that the probability for a unit i to satisfy C is p_i,

$$P(y_i \in C) = p_i$$

where p_i can be a given constant population parameter or an estimated value. For the purpose to simply the following discussion, we assume that p_i's are independent probabilities. Also we assume the probability for a unit i to be included into the initial sample is q_i.

$$P(i \in s^{(0)}) = q_i$$

For example, if the initial design is a simple random sampling without replacement and the initial sample size is n, then $q_i = 1 - \binom{N-1}{n} / \binom{N}{n}, \forall i$. Under the design and the assumed population model, we can then calculate the inclusion probabilities of each unit in the final sample.

There are some units in the initial sample of which the inclusion probabilities can be determined exactly without using the assumed model. If an initially selected unit satisfying C, the the inclusion probability of this unit can be determined by the following theorem

Theorem 1 *If an initially selected unit i satisfying C, then the inclusion probability of i, denoted by π_i, is*

$$1 - \binom{N - n_i}{n} \Big/ \binom{N}{n} \tag{1}$$

where n_i is the number of units which belong to $s_{(i)}$ and satisfy C.

Although we also can determine the exact inclusion probabilities of some other units by observing the final sample, however, it is not necessary to state all these miscellaneous cases.

The number of possible paths from i to j is a finite integer, denoted by n_{ij}. For the purpose to simplify the notation, define $\overrightarrow{ij}_k$, $k = 1, \ldots n_{ij}$ to represent the possible n_{ij} paths which start from i and end at j. Notice that the length of all possible $\overrightarrow{ij}_k$'s are equal to l_{ij}. The following theorem then follows immediately from Lemma 4 with this notation $\overrightarrow{ij}_k$, notice that one of the necessary condition for $\overrightarrow{ij}_k$ to exist is $i \in s^{(0)}$.

Theorem 2

$$\begin{aligned} & P(j \in s_{final}) \\ =& P(\exists i \textit{ such that } \exists k \in \{1, \ldots, n_{ij}\}, \\ & \overrightarrow{ij}_k \textit{ exists under the design}) \end{aligned}$$

We can then calculate the inclusion probabilities for every units $j \in s_{final}$ by the following discussion.

Consider there are n_{ij} possible paths $\overrightarrow{ij}_k$, $k = 1, \ldots, n_{ij}$ for each $i \in \{1, \ldots\ldots, N\}$ as

$$\begin{array}{ccccc} \overrightarrow{1j}_1 & , \quad \cdots \quad , & \overrightarrow{1j}_k & , \quad \cdots \quad , & \overrightarrow{1j}_{n_{1j}} \\ \vdots & & & & \\ \overrightarrow{ij}_1 & , \quad \cdots \quad , & \overrightarrow{ij}_k & , \quad \cdots \quad , & \overrightarrow{ij}_{n_{ij}} \\ \vdots & & & & \\ \overrightarrow{Nj}_1 & , \quad \cdots \quad , & \overrightarrow{Nj}_k & , \quad \cdots \quad , & \overrightarrow{ij}_{n_{Nj}} \end{array}$$

Let A_{ijk}, where $i = 1, \ldots, N$ and $k = 1, \ldots, n_{ij}$, represent the event such that the path $\overrightarrow{ij}_k$ exists under this IACS design with the maximum number of steps M, then the inclusion probability of j is

$$P(j \in s_{final}) = P(\bigcup_{i,k} A_{ijk}) \qquad (2)$$

and Equation 2 can be calculated by the additive theorem in probability theory.

In order to apply the additive theorem to Equation 2, we need to know the probabilities of all the single events and the intersections of all the possible combinations of A_{ijk}'s. Under the initial design and the population model described in this section, the probability for the path $\overrightarrow{ij}_k$ to exist, under the IACS design with the maximum number of steps M, can be obtained by the following lemma. Notice that the initial design is a conventional design, therefore, the initial inclusion probabilities $q_i, \forall i \in \{1, \ldots\ldots, N\}$ are independent from p_i, the probabilities to satisfy the condition of interest, $\forall i \in \{1, \ldots\ldots, N\}$.

Lemma 5 *The probability for a path* $\overrightarrow{ij}_k = [i, i_1, i_2, \ldots\ldots, i_n, j]$ *to exist is*

$$P(A_{ijk}) = \begin{cases} 0 & \textit{if } l_{ij} > M \\ q_i \cdot p_i \cdot p_{i_1} \cdots p_{i_n} & \textit{if } l_{ij} \le M \end{cases} \qquad (3)$$

The probability to satisfy the condition of interest which depends on the population model can be a given constant or estimated value. However, if a unit i has been observed, then $\forall i \in s_{final}$

$$p_i = \begin{cases} 1 & \text{if } y_i \in C \\ 0 & \text{if } y_i \notin C \end{cases}$$

Similarly, the probability for two different paths $\overrightarrow{ij} = [i, i_1, \ldots\ldots, i_n, j]$ and $\overrightarrow{i'j} = [i', i'_1, \ldots\ldots, i'_{n'}, j]$ to exists simultaneously is

$$\begin{aligned} & P(A_{ijk} \cup A_{i'jk'}) = q_{ii'} \cdot \\ & P(i,\, i' \text{ and all of the intermediate units in} \\ & \overrightarrow{ijk} \text{ and } \overrightarrow{i'j}_{k'} \text{ satisfying } C) \end{aligned} \qquad (4)$$

where $q_{ii'}$ is the probability of both i and i' are included in the initial sample. For example, with an initial design of simple random sampling without replacement and if $i \neq i'$,

$$q_{ii'} = 1 - \binom{N-2}{n} \Big/ \binom{N}{n}$$

The probability for more than two paths which can reach j exists simultaneously can be obtained by the similar principle in Equation 4.

4.1 Example

Consider a population of Figure 1, the variable of interest y_i is a indicator variable with values of 0 or 1. This indicator variable indicates whether a cell, which is the sampling unit, satisfies some prespecified criterion C or not. For example, it can be used to indicate the presence $(y_i = 1)$ or absence $(y_i = 0)$ of certain object in each cell. The neighborhood of a unit is defined as the four adjacent units on the direction north, south, east and west. The condition of interest is $C = \{y : y = 1\}$. That is, if the indicator variable of a initially selected unit is 1, then the units in its neighborhood will be included into the sample as well. Also, the adding process will be stopped after $s^{(2)}$ has been included into the sample. In order to keep our discussion simple, only one unit (5,5) was selected into the initial sample. The observed value of unit (5,5) is $y_{(5,5)} = 1 \in C$, therefore all the units in (5,5)'s neighborhood are included and observed. After the adaptive sampling procedure stopped at the second step, we have

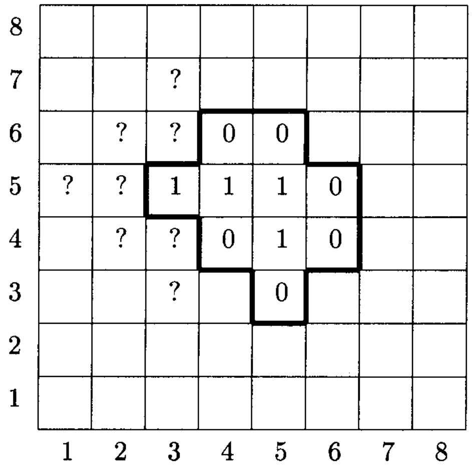
$$\begin{aligned} s^{(0)} &= \{(5,5)\} \\ s^{(1)} &= \{(5,4), (5,6), (4,5), (6,5)\} \\ s^{(2)} &= \{(5,3), (6,4), (4,4), (3,5), (4,6)\} \end{aligned}$$

and the associated observed values are

$$\begin{aligned} \mathbf{y}_{s^{(0)}} &= \{1\} \\ \mathbf{y}_{s^{(1)}} &= \{1, 0, 1, 0\} \\ \mathbf{y}_{s^{(2)}} &= \{0, 0, 0, 1, 0\} \end{aligned}$$

The observed units are blocked by the bold line as which in Figure 1.

Figure 1: A simple population



8								
7			?					
6		?	?	0	0			
5	?	?	1	1	1	0		
4		?	?	0	1	0		
3			?		0			
2								
1								
	1	2	3	4	5	6	7	8

Table 1 shows the actual inclusion probabilities of all the units which are included into the final sample. The simulated inclusion probabilities are simulated out of 100,000 populations of which the values of the observed units are fixed as what has been observed, and the values of the unobserved units are simulated under the assumed population model of which $p_{(i,j)} = 0.3 \, \forall i,j = 1,\dots,8$. The edge units are marked by the $*$ sign.

Table 1: The actual inclusion and simulated inclusion probabilities of each unit in the final sample of the population in 1 under the IACS design and the population model with $p_{(i,j)} = 0.3 \, \forall i,j = 1,\dots,8$. The initial sampling design is a simple random sampling with size 1.

	Inclusion Probability	
Unit	Actual Value	Simulated Value
(5,5)	0.06250	0.06216
(5,4)	0.04688	0.04651
(4,5)	0.07656	0.07625
(3,5)	0.06994	0.06899
(5,3)*	0.06994	0.06993
(5,6)*	0.07848	0.07721
(6,5)*	0.07848	0.07897
(6,4)*	0.06286	0.06305
(4,6)*	0.07708	0.07717
(4,4)*	0.09270	0.09180

5 Adjacency Matrix

The motivation of the idea to find all the possible paths which can reach a unit in order to calculate its inclusion probability is to avoid the complexity which may be caused by an intuitive approach. An intuitive way to obtain the inclusion probability of a unit i is to find the possible number of units which can lead to the inclusion of i, if they satisfy the condition of interest C. Denote this number to be m_i. Due to the uncertainty of the unobserved units and the design, there would be some different possible values of m_i. If there are H possible numbers of m_i, denoted by m_{ih}, $h = 1,\dots,H$, of units which can lead to the inclusion of i, then the inclusion probability of i can be written as

$$P(i \in s_{final}) = \sum_{h=1}^{H} P(m_{ih}) \cdot \left[1 - \binom{N - m_{ih}}{n} \Big/ \binom{N}{n}\right]$$

where N is the population size and n is the initial sample size. And we also need to find the probability of $P(m_i)$, It can be really tedious to find all the possible m_i's and the associated probabilities.

In Figure 1, all the units which are marked by the "?" sign are the unobserved units which can possibly contribute to the inclusion probability of unit $(3,5)$. The possible number of units which can lead (3,5) to be included into the final sample ranges from 4 to 12. We also have different combinations of units which can lead to the same number of $m_{(3,5)}$. For example, there are 10 different possible combinations of unobserved units which can make the number $m_{(3,5)}$ to be 6. The process will easily become too complicated to finish while the number of the maximum steps M increases.

On the other hand, the approach which proposed in Section 4 can be assisted by the adjacency matrix to calculated the inclusion probability with an efficient and systematic method. Also, this method will still be fairly simple even with a large M.

Define an adjacency matrix $\mathcal{A} = \{a_{ij}\}_{i,j=1,\dots,N}$, after the final sample has been collected, such that if $i \in s_{final}$ and j in in the neighborhood of i, then

$$a_{ij} = \begin{cases} 0 & \text{if } y_i \notin C \\ 1 & \text{if } y_i \in C \end{cases}$$

and $a_{ij} = 0$, $\forall i,j \in \{1,\dots,N\}$ if i and j are not neighboring. Also, we define $a_{ii} = 0 \, \forall i$. By this definition of an adjacent matrix $\mathcal{A}$, we will then be able to find all of the possible paths which we need in calculating the inclusion probabilities of the sampling units in the final sample. First of all, I shall state a well-known theorem in graph theory with the notations analogous to which in Section 1

Theorem 3 *Let $G = \{V, E\}$ be a graph, where V and E are as which given in Section 1, and $\mathcal{A}$ is the associated adjacency matrix with G. Then $\forall\, k \in Z^+$, $a^k_{ij} = (i,j)$th entry of $\mathcal{A}^k$, a_{ij} is the number of different walks from i to j of length k.*

We can have the following corollary by Theorem 3

Corollary 1 *If $a^h_{ij} = 0, \forall\, h \in Z^+$ and $h < k$, then a^k_{ij} is the number of different paths from i to j of length k.*

We can now construct a general algorithm to search all the possible paths of a unit i by using Corollary 1. Given the maximum number of steps M, we first construct the adjacency matrix $\mathcal{A}$ as what has been defined as above after the final sample has been collected and observed. Then $\forall\, i \in s_{final}$, we can

0. Let $\{i\} = s^{(0)}_i$.

1. Collect all the $j \in V$ such that $a_{ij} = 1$, denote this collection of units as $s^{(1)}_i$. i.e. $s^{(1)}_i = \{j \in V \mid a_{ij} = 1\}$

⋮

k. Let $s^{(k)}_i = \{j' \in V \mid \exists i' \in s^{(k-1)}_i, a_{i'j'} = 1 \text{ and } a^l_{ij'} = 0, \forall\, l < k\}$

Until $k = M$.

With all the $s^{(k)}_i$, $k \leq M$ ready, we can then construct all the possible paths which will be used in calculating π_i, the inclusion probability of unit i. First, we need to find all the

$$[i, i_1, \ldots, i_{k-i}, i_k],\ k \leq M$$

such that $a_{ii_1} = a_{i_1 i_2} = \ldots = a_{i_{k-1} i_k} = 1$ and $i_l \in s^{(l)}_i$, $\forall\, l = 1, \ldots, k$. Then all the possible paths, if they exists, which can reach i in at most M steps will be in the reverse direction of those paths which we have constructed above. The procedure which has been discussed in Section 4 then can be applied for the inclusion probability of i.

6 Discussion

The inclusion probability of the final sample obtained by the procedure in the previous sections can be used to construct a Horvitz-Thompson type estimators. If p_i is a given constant population parameter, then the inclusion probability in Theorem 2 is the exact inclusion probability of a final selected units. Consequently, the Horvitz-Thompsom estimator based on this inclusion probability will be an unbiased one. If p_i needs to be estimated and an unbiased estimator is available, then the Horvitz-Thompson type estimator will be a moment estimator. The properties, such as the unbiasedness and MSE, of this estimator then need to be examined according to different population models.

One possible advantage of using a population model to estimate p_i and the Horvitz-Thompson estimators is that we can make use of some prior knowledge of the population. But, meanwhile, we usually only need to estimate the probability of which y_i belongs to some pre-specified set rather than y_i itself. It might improve the robustness from using a pure model-based estimator if the population model is mis-specified.

A wide class of population models can be applied in this approach. In a practical framed spatial population, whether a unit satisfies the condition of interest usually depends on its neighboring units because of the natural spatial dependence. The assumption of the independent p_i in the discussion of Section 4 will often over-simplify the practical situation. However, with the approach using the idea of *path*, it seems natural to introduce a Markov Random Field model into this framework. Under this situation, a more sensible way to estimate the inclusion probability $P(j \in s_{final})$ might be to estimate the probabilities of the existence of events A_{ijk}'s, as well as the intersection probabilities of different A_{ijk}'s exist simultaneously, but not p_i seperately. And then Theorem 2 can be applied. A well defined Markov random field can help us to describe the dependence between the neighboring units. A further study for combining a practical spatial population model with the approach introduced in this chapter should be conducted in the future.

7 References

Foulds, L.R. (1992). *Graph Theory Applications*, Springer-Verlag, New York.

Thompson, S.K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association* **85**, 1050–1059.