Development of Sampling Procedures Based Upon Satellite Derived Land Cover History for the NSF Digital African Cities Project

Stuart E. Marsh, Thomas K. Park, Barbara A. Eiswerth,
Mohamud H. Farah, Douglas S. Rautenkranz and Barron J. Orr

The Six Cities project relied on a complex sampling scheme in part to ensure that all areas of habitation have a chance of being selected, that sampling could be based on historical land-use change and to insure that we would be able to critically evaluate the sampling strategy later. This sampling protocol allowed us to determine for example, if there is greater variability around sampling points in heterogeneous areas (which are themselves surrounded by points classified differently than they are) than around points in homogeneous areas. This is a methodological concern so we weight our sampling scheme for this project somewhat differently than we would if we were simply relying on the sampling strategy to do research. For the record and so that the data collected can be properly used we describe the full methodology of point selection in this paper.

I. Procedure for Generation of Random Sample Points

The selection of household survey sites was based on criteria initially developed from analyzing remote sensing data of the six cities. Data sets acquired for the study sites consisted of current and historical (1984-1999) SPOT (Multispectral and Panchromatic) images at 20-m and 10-m spatial resolution respectively and Landsat Thematic Mapper data at 30-m spatial resolution. The satellite image data were radiometrically calibrated, corrected for atmospheric variation between image dates and geometrically rectified to one another (yielding a RMS of < 0.5 pixel). The multispectral images were subsequently classified into land-use and land-cover classes using an unsupervised (ISODATA) algorithm. These classes varied between cities but generally included: Open Ground, Grass/Shrub, Roads, Mixed Vegetation, Agriculture, Low Density Urban, High Density Urban, Industrial/Commercial/Service Land-Use, and Water. Resulting multitemporal classifications were used in a change analysis procedure in which matrices of land-use/cover change were created between multiple image dates. In addition, change images were created to identify areas of change/no change between urban and non-urban land-use classes and urban housing density classes. These final maps had a spatial resolution of 20-m x 20-m.

Two sets of random sample points were then generated: (a) sample points falling only on the areas of change from non-urban to urban, and between urban classes (e.g. low density urban to high density urban), and (b) sample points falling on the non-change urban areas. The survey requirement was to have 12 sample points falling in areas of change, and 28 points falling in urban areas exclusive of the change areas, for a total of 40 sample points.

Field personnel conducting the sample may have had to eliminate points that are in unsuitable locations. For this reason, we generated a set containing 84 points for the change areas and 196 points for the non-change areas. For the change areas, the ratio between the size of the set of potential sample points (n=84) and the set of points actually selected (n=12) for the survey is 7. For the non-change areas, this ratio (196:28) is 7, as well. Maintaining the same ratio ensures that a point in the set of potential sample points will have the same probability of being selected as an actual sample point, for both the non-change and change sampling operations. Details are given below.

Random sampling requires that no point have a greater opportunity than any other point to be included in the sample. Consider points in the map of change from non-urban to urban classes and between specified urban classes. Any point falling in one of these change classes could be selected as a sample point during part (A), whether or not it actually is selected. For this reason all such points must be excluded from the image used to generate part (B) sample points. There are several ways of doing this. The approach that we implemented is explained below. A flowchart is appended to this section to show the overall process with the procedures and resulting maps at each step.

Map Preparation

The following sections describe the procedures for generating random points for the areas of change (Part A) and the areas of no change (Part B). It is presumed that the change matrix has been made from the original classified image that had 2 – 5 (depending on the city) urban classes.

The change matrix was first recoded into two images. In each, all change classes but those specified should be recoded to zero (0). This produced the following set of map variables.

Change Image -> C1: non-urban to any of low, medium, or high density urban
C2: any lower density urban to any higher density urban

No-Change Image* -> C1: unchanged low-density urban class

C2: unchanged medium-density urban class

C3: unchanged high-density urban class

*Depending on the city, 2 or 3 classes will be present; Dakar had 5 urban classes.

A. Sample points for areas of change.
A.1. First recode the change image into two separate images with one class each.
A.2. Generate a density map for each of the single-class images showing the degree of homogeneity of change in 3x3 pixel units across the change image. The resulting density image will have 9 classes plus a zero class. Class 1 is the class of locations of center pixels of all 3x3 matrices in the density map having only one pixel (the center of the window only) of the specified change type. Class 2 is the class for which there are two pixels (the center pixel plus one other in the window) of the specified change type in the 3x3 area, and so on.A.3. We expected urban change normally to be homogenous over an area of several pixels. We considered that there is a higher probability of classification error, resulting in mistaken change detection, for cases of density classes lower than 6 (< pixels showing change). For this reason, the C1 and C2 density maps (now D1 and D2 in the flowchart) will be sampled in two density ranges: densities 1-5 and densities 6-9. We then generated more sample points for the 6-9 density range, thus weighting it more heavily than the 1-5 density range. Each density map was recoded into two maps with one density class each: one map with all of the pixels in the original density classes 1-5, and one map with all of the pixels in the original density classes 6-9. For one map (D1), the density classes 1-5 were recoded as class 1, and all other density classes as class 0, which is also the background class. In the other new density map, the density classes 6-9 were recoded as class 1, and the other classes as 0.

Table 1. Creation of Change Density Classes

Density map	Original density class	New class
D1	1-5	1
	6-9	0
D2	1-5	0
	6-9	1

A.4. Random sample points were then generated from the four density images containing pixels from the change image. Random sample points were generated for each image in the following combinations:

Change type	Density Range	Recoded Class	Number of points
C1	1-5	1	12
C1	6-9	2	30
C2	1-5	1	12
C2	6-9	2	30

A.5. The generated points could then be saved as an internal table in Imagine, or exported as an ASCII text file for use with ArcView or Excel.

B. Sample points for areas of no change.
B.1. The steps implemented to generate sample points for areas of no change follow. The no change image was recoded to two or three (depending on the city) separate images representing areas of no change for each urban class present.

B.2. A density map for each single class image using the procedure described in section A-2 was created.

B.3. A total of 196 sample points were generated from which the field teams could extract 28 usable sample points for areas of no change. Once again, we split the density class range into two parts, density classes 1-5 and 6-9, in each image.

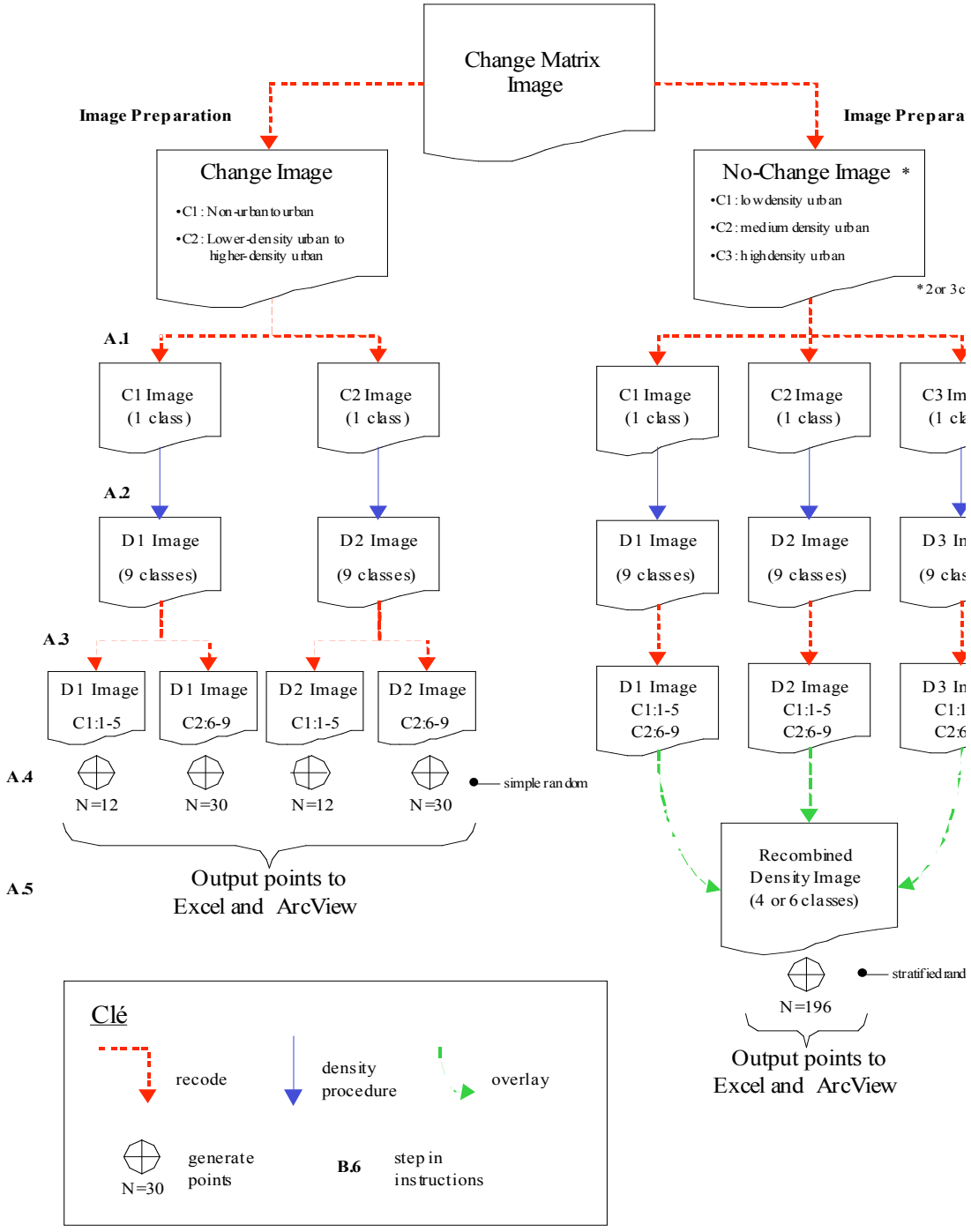
B.4. The sample for the no-change image was then drawn proportionately from among the density classes. These maps were then recombined and produced one map consisting of four or six classes depending on whether you had two or three urban classes in the original image. For example, for an original map with three urban classes the classes would be:

Original Urban Class	Original Density	Recoded Density Map	Class in Recombined Image
high	1-5	1	5
high	6-9	2	6
medium	1-5	1	3
medium	6-9	2	4
low	1-5	1	1
low	6-9	2	2

B.5. Random points were then extracted in a similar manner to section A-4 but following a stratified random sampling procedure based upon area covered by each class in the recombined map.

C. Generate Random Points Map and Table The final step was to generate both digital and hard copy city maps, in a UTM coordinate system. These maps displayed the random sample points overlain on a false color or panchromatic satellite image. An Excel spreadsheet along with the map was provided to the field teams and which included the point ID#, the X and Y coordinates (in UTM), and the density class; the table may also include the original urban class and the neighborhood designation if known. With all points generated, located and their representativity well noted the survey results could then be used to collect representative data on a great variety of subjects and to evaluate the success of the methodology.

The sampling procedures followed in the Six Cities project are outlined in the following flowchart:


Change Matrix Image
Image Preparation
Image Prepara
Change Image
•C1 : Non-urban to urban
•C2 : Lower-density urban to higher-density urban
No-Change Image *
•C1 : low density urban
•C2 : medium density urban
•C3 : high density urban
* 2 or 3 c
A.1
C1 Image (1 class)
C2 Image (1 class)
C1 Image (1 class)
C2 Image (1 class)
A.2
D1 Image (9 classes)
D2 Image (9 classes)
D1 Image (9 classes)
D2 Image (9 classes)
A.3
D1 Image C1:1-5
D1 Image C2:6-9
D2 Image C1:1-5
D2 Image C2:6-9
D1 Image C1:1-5 C2:6-9
D2 Image C1:1-5 C2:6-9
A.4
N=12
N=30
N=12
N=30
simple random
A.5
Output points to Excel and ArcView
Recombined Density Image (4 or 6 classes)
stratified rand
N=196
Output points to Excel and ArcView
Clé
recode
density procedure
overlay
N=30
generate points
B.6
step in instructions

II. Incorporating Local Knowledge into the Stratification Scheme

The urban classifications and the pixel density or the "nine pixel context" provide the strata from which the computer generates a series of geographically located points. These strata provided us with an initial stratified sample reflecting what can be seen in a remote sensing image, i.e. largely building size and building density. This stratification does not include any obviously reliable information on other relevant factors such as population density, ethnicity, poverty/wealth etc. It was up to the local team to incorporate their knowledge of such factors into the sampling frame via an additional level of stratification. This was based upon a few simple rules to avoid introducing bias into the sample:

> Rule 1: All computer-generated points must have a non-zero chance of appearing in the final sample.
> Rule 2: The basic procedure is to weight some subsets of computer generated points more heavily than others by choosing, on the basis of local knowledge, a greater proportion of points from some subsets than from others.
> Rule 3: The criteria that may be considered suitable reasons for differential weighting of subsets of points include the desire to better capture areas of higher population density, areas of greater than normal ethnic diversity or areas of greater than normal poverty. In a related way, if an area is felt to be extremely homogeneous compared to others an argument can be made that fewer points may be needed to capture diversity in that area than in others.
> Rule 4: All criteria for differential weighting must be carefully explained and the numerical ratios for point selection must be carefully noted. Thus a decision might be made to sample an area including 30 computer generated points at a 20% rate (thus selecting 6 points at random from these 30 points) while another subset of points, say 60 points, is to be sampled at a 10% rate - also giving 6 points in the final sample. This would be perfectly acceptable provided that these ratios are carefully noted and that the reasons for the differential sampling are well thought out and recorded.
> Rule 5: Incorporate important social science knowledge into the sample selection procedure via this further level of stratification based on your local knowledge where practicable but do not try to incorporate data that is poorly or only vaguely understood as it will needlessly complicate the procedure. Only incorporate local knowledge for which there is a high level of confidence.
> Rule 6: Remember that the points selected are merely focal points where the interview team will first make a list of the closest 20 households (in terms of walked distance) and then randomly select six for interviews. There does not have to be a household at the exact point selected.

Finally, the interview teams were instructed to avoid these basic errors:

> 1. Do not arbitrarily select, or directly or indirectly, exclude any points. For example, if we were to select five points out of 30 in a given area and then decide one was unsuitable (in a local industrial area) we might be tempted to simply drop it from the sample and choose another point. The best procedure would still be to find the closest 20 households to that point, even though the point itself is in a non-residential area, and to leave the point in the sample. To do otherwise might systematically move the sample away from pockets of industry and thus bias the sample.
> 2. Do not make subsets that second guess the remote sensing data. For example, the remote sensing team may mistakenly designate an area as high density residential but it is inappropriate to move any points out of this category just because on-the-ground reality suggests another class would be better - the classification groups pixels by reflectance and “graininess” and all similar points, from this perspective, need to be sampled as a group to help improve our knowledge and to critique the remote sensing methodology.

3. Once households have been selected they must be interviewed-- we cannot "go next door" because no one is home at the moment when we arrive to do the interview. Doing this would introduce serious bias into the procedure - we might, for example, systematically miss people who work at that time of day or undersample people who tend to be out of their home more frequently. Many similar systematic biases are this easy to introduce but all should be avoided - this requires care in creating the sample. A small reliable sample is in fact one of the primary goals of the methodology.

Abstract

This article discusses the sampling scheme employed by the Six Cities project to ensure that all areas of habitation have a chance of being selected, that we know what that chance is, and that we are able to critically evaluate the sampling strategy after it has been carried out. A weighting strategy that is slightly different from one used only to do research is therefore employed. The article describes a procedure for generating two kinds of random sample points for areas of change and of no change. Finally, a few simple rules for incorporating socioeconomic, demographic, and other relevant information into the sampling frame without introducing bias into the sample are discussed.
Key words: sampling strategies; random sampling; sampling bias; local knowledge; Six Cities project; remote sensing; urban areas in Africa

Résumé

Cet article examine un schéma pour choisir des échantillons qui a été employé par le projet "Six Cities" afin d'assurer que tout les secteurs d'habitation ont une chance d'être sélectionnés, qu'on sait de quoi s'agit-elle, et qu'on peut évaluer critiquement cette stratégie d'échantilloner après qu'elle a été faite. On emploie par conséquent une stratégie de peser des échantillons qui est légèrement différente de celle exigée pour faire des sondages uniquement. Le texte examine une méthode qui est employée afin de générer deux genres de points d'échantillon aléatoires, pour des régions où on trouve un changement, et pour des régions sans changement. Dernier, on examine quelques règles qui nous permettent d'incorporer des données socio-économiques et démographiques, par exemple, dans le cadre d'échantillon sans pour autant en introduir des biais.
Mots clés: les stratégies d'échantilloner; l'échantillonage aléatoire; le biais dans l'échantillonage; le savoir local; le projet "Six Cities"; les secteurs urbains en afrique; la télédétection

Resumen

Este artículo discute el método de muestreo empleado por el proyecto de Seis Ciudades para asegurarse de que todas las áreas de vivienda tienen la misma probabildad de ser seleccionadas, el saber cuál es esa probablidad, y poder evaluar críticamente la estrategia de muestreo después de que se haya competado. Una estrategia de calcular los pesos que es levemente diferente de una usada para investigación solamente se emplea por lo tanto. El artículo describe un procedimiento para generar dos clases de puntos de la muestra escogida al azar para las áreas del cambio y de ningún cambio. Finalmente, se discuten también algunas reglas simples para incorporar socioeconómico,demográfico, y la otra información relevante en el marco de muestreo sin introducir ningún perjuicio en la muestra.
Palabras claves: estrategias de muestreo; muestreo escogida al azar; perjuicio del muestreo; conocimiento local; Proyecto de Seis Ciudades; detección remota; áreas urbanas en África