The Six Cities Project: developing a methodology of surveying densely populated areas using social science assisted and diachronic remote sensing based classification of habitation

Thomas K. Park and Mamadou Baro

Some 15 years ago the first author had the thought that what is done for forests by remote sensing might be adapted to the study of cities to facilitate comparative study of rapid urbanization. This might have significant advantages in poor countries that have limited means to keep abreast of urban growth and the concerns of urban populations. If urbanization could be readily mapped and classified into housing types as well as their rough date of development it would be possible for scholars, municipalities, national governments and NGOs to more readily understand what is occurring in large urban centers. Furthermore, if a consistent and replicable methodology were developed, useful comparative study of urbanization would be feasible at low cost. In 1998, the National Science Foundation awarded researchers at the University of Arizona funding to develop such a methodology in conjunction with local partners in Morocco, Senegal, Mali, Niger, Tanzania and Botswana. At its most simple, the ability to sample based on pixels whose precise coordinates are known past, present and future makes it possible to monitor change spatially and ask questions which otherwise would be quite impossible to answer.

Methodology

While remote sensing readily provides images its key advantages over photography include that each pixel has precise coordinates, that each pixel simultaneously has the potential for information on several wavelengths and that images from multiple periods can be linked and changes automatically recognized due to each image having pixels at the same coordinates. This allows a computer and ground based classification of all pixels into useful categories. Thus, for our purposes, while very high resolution images can sharpen the classification, it is the classes themselves that are of value. To know all the shades of color on a wall would have very little value but to be able to tell the difference between shantytown dwellings and high rise apartments would be of major value. Fortunately, such distinctions are well within the capacity of computer based classification systems (especially if assisted by ground truthing). The construction techniques followed in each country are fairly distinctive but there tends to be significantly more similarity from city to city within a country. This means that the details of a classification turn out to be country, and to some extent region, dependent while the methodology for developing a classification is generalizable.

In this initial development of the methodology, we were concerned both with developing specific classifications for each city and with collecting information to evaluate both the accuracy of the classification and its utility from a social science perspective. Remote sensing based classification is by its nature based on the evaluation of buildings from above. This perspective allows one to consider size, construction material, spatial density and a number of other characteristics. While we felt initially that it was a reasonable starting assumption that the quality of housing would correlate in some significant degree with a variety of socioeconomic indicators,

one of the intents of this initial study was to find out how well this might in fact be the case. It should be noted in this regard that this approach allows one not only to consider the quality of housing but also its location in the city, its approximate time of construction, its proximity to a multitude of urban features and even the general historical development of that pixel and its neighborhood. When these potentials are incorporated into the methodology the likelihood of the methodology being valuable for the construction of a coherent sampling strategy for surveys on a vast array of topics from poverty, access to education, health issues or environmental issues increases enormously.

The core approach then is to create a set of classes which can be used as a sampling frame, to draw a stratified random sample of pixels and to use these pixels as center points for surveys. New points can then be effortlessly generated as needed for future surveys and old points can be returned to equally easily for long term monitoring. Each sample can be counted on to be representative and the details of its generation can be used to extrapolate to the larger city. Thus if a particular habitation class is thought to be quite heterogeneous from an ethnic perspective it can be sampled more intensively if differences in ethnic adaptation to the urban environment are of interest and this weighting can be easily taken into consideration in any extrapolation to the urban population as a whole.

In its basic outline, the methodology developed involves the development of urban classes in each city based on housing type, as perceived via remote sensing images over an approximately 20 year period. The most recent image was ground truthed and used to help classify the earlier images. Once the classification is complete in its basic form, a set of final classes are developed which incorporate the element of time: e.g. a shanty town present in 1980 would be a different class than one appearing only after 1990 and similarly for each basic urban class. A further set of classes were then created marking areas of change or transformation, e.g. from shanty town to low income housing or rural land transforming into shanty town. The final set of classes incorporating all classes were then used as the initial sampling frame. Local partners provided expertise on potential homogeneity or heterogeneity of each class that was used in the weighting of sample points drawn for each class within the final sample.

In order to test the accuracy of the classification itself, some additional factors were used in the selection process: we introduced into the computer sampling algorithm the notion of a sampling window (3 by 3 pixels) that moved over all the pixels and took part of its sample from among those pixels that were in highly homogeneous areas and others in more heterogeneous areas - based on classifications associated with the 9 pixels in the window - the central pixel was the one about which a decision to select or not to select was based on the probabilities specified to the algorithm. The procedure ensured that we had a reasonable number of sample points in homogenous and heterogeneous areas - defined by housing type. While this particular refinement was important in providing feedback on the remote sensing classification it is not important for this paper which concerns itself with another question; namely the utility of the classifications based on habitation as a basis for evaluation of social, environmental, health, and economic issues of concern to the urban populations.

The actual procedure followed was to carry out household level surveys around each selected pixel. For budgetary reasons we confined ourselves to a total of 240 households distributed in

groups of six around 40 selected pixels in each city. The six households were themselves randomly selected from the 20 households located closest to the sample points.

Evaluation of the methodology

The project methodology has been outlined but the original question and hope was that the methodology could facilitate the study of cities and urban planning in general. The core question was whether remote sensing images could be used to help create a stratified random sampling technique for urban areas that would make it possible to tackle a multitude of urban questions via a small sample that was nevertheless highly representative of the greater urban area in that it incorporated large amounts of urban variation.

In an ideal world, we could test the methodology by comparing survey results where we already have copious amounts of data to see how the sample data compares with the larger data set. In the case of the cities we have selected (Marrakech, Dakar, Bamako, Niamey, Dodoma, and Gaborone) this is obviously not possible. We have, for comparison, primarily selective data with no claim to high representativity and the data is often seriously dated. In addition what data is available from the national census is also almost always dated as well and is much too limited from a social science perspective.

In consequence, we cannot directly compare sample data with up-to-date comprehensive data. Instead we need to examine the data collected and see if the categories included in the sampling framework aggregate the data in significant ways or if by contrast the data collected differs only randomly among the strata. *If habitation really corresponds well with a host of other factors then the sample selected using our methodology will capture this data as well as the habitation differences*. In principle we can find out how likely this is by looking first at how well the classification system and the sampling frame captures differences in habitation and then examine how well the same procedure groups differences in other areas.

Thus if the urban classification and the sampling scheme group particular housing or socio-economic data then the data associated with each class will be significantly different than the data in other classes and the data from each sampling point may differ as well in as much as it is reasonable to assume some differences based on geography as well as habitation type. *If the data associated with each of the survey points is significantly homogeneous and significantly different from that around other sampling points then the overall sampling scheme would appear to be well founded – assuming the goal is to capture a maximal amount of variation in a small sample. If in addition the urban classes themselves capture both significant homogeneity within the class and heterogeneity between classes then we will clearly be benefitting from the remote sensing approach to constructing a sample frame*. By contrast if key variables we expect to differ appear similar between urban classes then the remote sensing procedure does not facilitate the construction of a stratified sampling strategy and if the same is true between sampling points then the sampling frame itself would provide little utility.

Taking these points into consideration, we can test the methodology without having comprehensive comparative data. The procedure we have used to do this is to run Kruskal-Wallis tests for selected variables grouped by urban class and by sampling point. As a comparison we have also used the urban "quartier" or ward as a grouping. This simply groups sampling points by

ward to see if there appear to be major distinctions at this level of aggregation . It may be noted that this should not be seen as directly comparable to a more straightforward random sampling scheme stratified only by ward because it remains tied to the remote sensing sampling framework used to select the points.

The Kruskal-Wallis test is useful for quantitative variables and works roughly as follows: it ranks all the data (e.g. from highest to lowest) and then looks at which ranks fall into which groups (as specified in a second variable: e.g. urban classes or wards or sampling points) and uses a Chi-squared test to measure how significant this grouping is. A high probability for the null hypothesis would imply that the grouping is highly likely to derive its values solely by chance while a low probability would imply that it is highly likely the grouping itself is significant. *This statistical analysis will in effect ask two questions of each of the three groupings (ward, urban class, sampling point): are they accurately grouping differences in habitation and are they accurately grouping differences in other areas (e.g. social, public health, economic, environmental).* The baseline surveys carried out by our local partners included information a a great variety of topics but the analysis below provides the statistical results for only a subset of the variables included in the surveys.

The Kruskal-Wallis test is a one way analysis of variance or a non parametric determination of whether the observed differences between the groups is due to random effects or to the specifics of the sampling frame (i.e. the success of the grouping procedure). We have presented the results of the analysis in a series of two tables per city in which the the first (odd numbered) table presents the analysis for housing data and the second (even numbered) table presents the results for the socio-economic variables. The data set has a variety of social science variables in each category as many composite variables can additionally be created from the base variables. In this paper we will restrict ourselves to a subset of variables from each city though we have not limited ourselves to exactly the same variables in each city since the surveys, though following a common format, were not identical and particular variables are useful in making specific points.

An initial look at the housing related variables for Marrakech will illustrate the comparisons we will be considering in the rest of the paper.

Table 1. Kruskal Wallis statistics for Marrakech: housing variables

	Urban Class		quartier		sampling points		
Variable	Chi-squared	Probablity H_0	Chi-squared with ties	Probablitiy H_0	Chi-squared with ties	Probablity H_0	Variable Descriptions
q25	21.805	0.0053	22.64	0.1237	41.227	0.0392	if you rent how much is rent
q29	13.51	0.1408	38.549	0.0033	77.092	0.0002	number of rooms
q29a	72.698	0.0001	86.938	0.0001	113.048	0.0001	average area m^2 of rooms
q31a	30.318	0.0004	54.632	0.0001	76.248	0.0002	size of courtyard
q31	17.356	0.0434	25.125	0.1215	71.776	0.0008	is there a courtyard
q34	36.657	0.0001	82.153	0.0001	104.09	0.0001	# of showers and WC combined

Data Source: Six cities survey of Marrakech, 2001, supervised by Ahmed Belasri, Qadi Ayyad University.

The Kruskal Wallis tests in this table examine six variables (q25, q29, q29a, q31a, q31, and q34) grouped according to three group variables (Urban Class, quartier, and sampling points). The results are displayed in two columns for each combination of test variable and grouping variable. The first column provides the Chi-squared figure and the second the probability that the null hypothesis (that the grouping is insignificant) is true. The most comparable of these figures across variables is the probability so we will focus on this. Variable q25 is highly significant (0.0053) grouped by Urban Class. This suggests that the remote sensing procedure can easily distinguish areas of high rent from areas of low rent and that rent is significantly different in each of the urban classes picked up by remote sensing. Although this variable is also highly significant grouped by sampling point the distinctions seem not quite as clear cut while the sampling points within particular quartiers do not seem to have sufficient commonality among themselves or distinctiveness as a group from sampling points in other quartiers (0.1237).

Variable q29 which measures the number of rooms in the household residence seems highly significant grouped by quartier (0.0033) and by sampling point (0.0002) but only moderately significant when grouped by Urban Class (0.1408). This suggests that location as well as habitation type is a critical determinant of the number of rooms in households' habitation. This is not the case for variable q29a which measures the size of the rooms in question because this is highly significant grouped by any of the three variables (0.0001 in all cases). Variable q31a, which measures the size of the courtyard is similarly highly significant for all three grouping variables. Variable q34 which measures the level of amenities is also significant for all three groupings (0.0001 in each case).

The Marrakech case would seem to suggest that, at the level of housing characteristics, the urban classification system not only does an excellent job distinguishing housing differences but works well as a key component of a sampling scheme intended to capture a large amount of urban variety in a small sample.

At this point in the analysis it is an open question whether this holds up for the other cities but for the moment it may be useful to examine a few social variables to see if they follow a similar pattern.

Table 2. Kruskal Wallis statistics for Marrakech: socio-economic variables

	UrbanClass		quartier		sampling points		
Variable	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Variable Description
q26	43.352	0.0001	31.261	0.0269	64.621	0.0045	how many years has HHhead resided here
q43	39.224	0.0001	69.358	0.0001	117.768	0.0001	how often do you leave the neighborhood
q60b	15.428	0.0798	32.102	0.0214	67.724	0.0021	how many times per year do rural relatives visit
i11_nbr	14.457	0.107	22.415	0.2141	50.552	0.0037	total number of persons in HH

Data Source: Six cities survey of Marrakech, 2001, supervised by Ahmed Belasri, Qadi Ayyad University.

Table 2 provides the results of a similar analysis for four basic survey questions which might not be thought to be intrinsically tied to housing characteristics. As can be seen, for Marrakech all

four variables seem highly significant when grouped by sampling point and the first three are significant when grouped by Urban Class and quartier. Why this might be the case is worth some discussion. The remote sensing classification scheme uses diachronic imagery so that areas which are new in later images are classed as "change" pixels and are separately sampled. Hence new areas of town and new developments in town are separate classes and it makes some sense that new comers to the city may be distinctive in many ways just as households in different parts of the city may vary significantly. Housing type also gets picked up by the classification system and it follows that households in different types and qualities of housing may differ significantly in socio-economic ways as well.

In the Marrakech case this means that both demographic differences (ill_nbr), migration differences (q60b), urban mobility differences (q43), and residential history differences (q26) seem to be captured by the sampling methodology at slightly different levels (as indicated by the chi-squared results). There are presumably commonalities as well as differences that are linked to more complex issues than housing quality, date of arrival, and residential location but our concern here is merely to establish that the sampling methodology seems to capture a great amount of variation along many dimensions with a high level of significance.

Turning to the data from Dakar we can see that housing variables are clearly differentiated by the full sampling strategy but not all groupings are equally effective.

Table 3. Kruskal Wallis statistics for Dakar: housing variables

	Urban Class		quartier		sampling points		
Variable	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Variable Description
siz_crty	17.838	0.0372	114.408	0.0001	114.471	0.0001	size of courtyard
rm_	13.639	0.1358	61.371	0.0053	69.749	0.0018	number of rooms
hous_typ	14.057	0.1203	78.883	0.0001	91.387	0.0001	type of housing
m2_pers	15.991	0.0671	56.696	0.0154	65.525	0.0049	m2 per person

Data Source: Six cities survey of Dakar, 2000, supervised by Magatte Ba, Centre de Suivi Ecologique, Dakar.

Grouped at the level of Urban Class two variables are highly significant (siz_crty and m2_pers) while the other two are marginally significant (rm_ and hous_typ) but all four are highly significant both grouped by quartier and by sampling point. This suggests that for Dakar there is slightly less homogeneity within urban classifications than in Marrakech. It may be that a higher resolution image could produce an improved classification but it is clear nonetheless that the overall sampling strategy is highly effective in making distinctions among habitation variables.

While we are primarily interested in testing the utility of the methodology and rely on the Kruskal Wallis test for some basic indications in this regard it should not be assumed that either of the authors. or any of the six cities researchers, believe that housing should correlate well with all other significant factors. This would be an absurd and unlikely proposition. Instead we maintain merely that if it correlates with quite a few its methodological advantages make it a highly valuable approach to urban surveys. It should be said that other techniques would obviously be preferable for some types of research; thus for example if one were looking at political networking some sort of networking methodology would be clearly preferable. Our position is simply that for many purposes our methodology will provide a better basis for the study of issues of relevance to

a broad part of the urban population.

Table 4. Kruskal Wallis statistics for Dakar: socio-economic variables

	Urban Class		quartier		sampling points		
Variable	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Variable Description
pers_hh	5.739	0.7657	48.909	0.074	52.129	0.077	# persons in HH
rev_tot	18.498	0.0298	49.527	0.0661	59.489	0.0201	total monthly income
inc_cap	14.295	0.1122	41.3	0.2501	48.208	0.1483	income per capita
exp_mth	19.139	0.024	84.198	0.0001	90.071	0.0001	total expenses per month
exp_pers	27.211	0.0013	103.911	0.0001	110.127	0.0001	expenses per capita per month
ratio_pc	15.284	0.0834	56.297	0.0168	60.705	0.0146	ratio of producers to consumers

Data Source: Six cities survey of Dakar, 2000, supervised by Magatte Ba, Centre de Suivi Ecologique, Dakar.

At the level of socio-economic variables, the situation is, as might be expected, more complex. In Dakar, the urban classification does not do a good job distinguishing household size (0.7657) perhaps because household size varies with more than house quality and the range of household size is considerable. Nevertheless when quartier or sampling point are used, household size seems to be significantly linked to both. The variables presented in Table 4 include summations or transformations of base variables. Oddly, income per capita is marginally better explained by Urban Class than by other groupings though it is still only marginally significant (H_0 = 0.1122). The other transformed variables (total monthly income, total monthly expenses, per capita expenses and the ratio of producers to consumers in the household) are highly significant over all three groupings.

Table 5. Kruskal Wallis statistics for Bamako: housing variables

	UrbanClass		quartier		sampling points		
Variable	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Variable Description
h307	11.708	0.0085	79.21	0.0001	136.809	0.0001	type of habitation
h310	7.305	0.0628	54.421	0.0003	77.645	0.0002	number of rooms
h311	8.854	0.0313	44.186	0.0034	71.535	0.0011	av m^2 per room
h312	2.217	0.5285	40.598	0.0092	76.012	0.0004	# persons per room
h314	12.988	0.0047	88.925	0.0001	129.321	0.0001	size of courtyard
h315	5.668	0.1289	112.338	0.0001	144.575	0.0001	number of WC
h317	41.337	0.0001	139.091	0.0001	184.157	0.0001	number of showers
h318	34.279	0.0001	139.844	0.0001	193.837	0.0001	# showers and WC combined
h320	7.886	0.0484	97.334	0.0001	145.491	0.0001	# of faucets

Data Source: Six cities survey of Bamako, 2000, supervised by Sadio Traore, CERPOD, Bamako.

The situation in Bamako is little different in as much as all nine housing related variables listed are highly significant at both the grouping by quartier and by sampling point. Four of the six are even highly significant grouped at the Urban Class level. The two which are not (h312 and h315) the number of persons per room and the number of WC suggest that there are both greater differences in population density and less predictable provision of amenities than might would be

the case if external classification of housing were predictive. That both variables are highly significant at the other two levels of grouping suggests that location in the city explains much of this variation.

Table 6. Kruskal Wallis statistics for Bamako: socio-economic variables

	Urban Class		quartier		sampling points		
Variable	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Variable Description
e409	19.325	0.0002	36.679	0.0183	56.665	0.0046	annual payment for garbage
e435	14.794	0.002	69.905	0.0001	84.572	0.0001	# times rural relatives visit per year
e506	9.132	0.0276	96.473	0.0001	107.849	0.0001	Expenditures energy
d605	17.437	0.0006	110.705	0.0001	134.34	0.0001	expenditures
m10	5.378	0.1461	37.944	0.0186	53.254	0.0637	# persons in HH
m166	0.176	0.9814	31.794	0.081	49.427	0.1224	# of years in principal occupation
m211	4.781	0.1885	42.685	0.0052	64.314	0.0065	monthly income
m256	9.021	0.0351	15.582	0.4104	45.223	0.028	annual income

Data Source: Six cities survey of Bamako, 2000, supervised by Sadio Traore and Moise Balo, CERPOD, Bamako.

The variables dealing with socio-economic issues provide a complex picture. Five of the eight presented are highly significant grouped by Urban Class while seven of the eight are significant grouped by sampling point and by quartier. The number of persons in the household (m10) is not particularly significant grouped by Urban Class while the number of years the household head has spent in his/her principal occupation is completely randomly associated with Urban Class (0.9814) yet mildly significant (0.1224) when grouped by sampling point and highly significant when grouped by quartier (0.081). Monthly income is not really significant grouped by Urban Class (0.1885) while it is highly significant grouped by quartier (0.0052) and sampling point (0.0065). This suggests that in Bamako socio-economic status is more closely tied to location than to type of habitation or length of time in the city.

At the same time, the urban classification easily picks up such things as connectedness to the rural area (e435 with probability of 0.002) and access to or payments for garbage collection (e409 with H_0 probability of 0.0002). Both variables may be linked to age of the residential development or they could both be associated with quality of the housing. Low quality housing might be inhabited by recent immigrants or people with only marginal urban occupations. The high significance of the expenditure variables (e506 and d605) also suggests that costs are associated with the urban classification though the significance of associations become extremely high (0.0001) in the other two groupings.

Niamey is divided into three rather distinct communes which vary considerably both in terms of time of construction and in terms of their residents. The newest (Commune III) for example is across the Niger from the other two and houses the university and associated residences while the oldest (Commune I) includes the government residential areas and their associated housing. Thus location seems to be particularly important even at the macro level. Our study in fact suggests that simple things like housing size and composition vary widely between the three different communes.

Table 7. Kruskal Wallis statistics for Niamey: housing variables

	Urban Class		quartier		sampling points		
Variable	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Variable Description
H307	8.906	0.0306	71.133	0.0001	94.264	0.0001	type of habitation
H310	15.372	0.0015	63.953	0.0001	79.171	0.0002	# of rooms
H311	16.99	0.0007	113.08	0.0001	143.119	0.0001	Av m^2 of rooms
H312	10.432	0.0152	50.533	0.0012	78.513	0.0003	Av # of persons per room
H314	11.577	0.009	54.15	0.0004	105.771	0.0001	Area of courtyard
H315	5.257	0.1539	27.328	0.073	38.904	0.1557	number of WC
H318	2.841	0.4168	19.645	0.1418	30.308	0.065	Number of WC and showers combined
H320	4.723	0.1932	27.613	0.0682	46.656	0.0352	Number of faucets

Data Source: Six cities survey of Niamey, 2000, supervised by Adamou Nafoga, SAP, Niamey.

In Niamey, five of the eight housing variables are highly significant when grouped by Urban Class. These include type of habitation, number of rooms, average room size, average number of people per room and the area of the courtyard. The number of WC is only mildly significant grouped by Urban Class or by sampling point but is highly significant grouped by quartier. This suggests quartiers differ very significantly in this regard but may be fairly homogenous within the quartier. The more complete assessment of amenities (H418) is completely insignificant at the level of Urban Class but rises to high significance at the level of sampling point. One possible interpretation may be that homogeneity clumps within smaller areas on this dimension.

Table 8. Kruskal Wallis statistics for Niamey: socio-economic variables

	UrbanClass		i4_q quartier		i5_ech points		
Variable	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Variable Description
D605_M	24.007	0.0001	122.485	0.0001	144.419	0.0001	mean amount of expenditures
NB_PERS	0.446	0.9314	21.236	0.6248	36.322	0.6366	# of people in HH
E409	7.07	0.0697	28.309	0.134	47.943	0.088	amount of garbage service
E435	20.07	0.0002	48.337	0.0006	71.955	0.0007	# times rural relatives visit per year
H301	1.643	0.6489	38.205	0.033	71.693	0.0015	# of years in habitation
H306	2.456	0.4827	38.655	0.0297	69.834	0.0024	# years HHhead in Niamey

Data Source: Six cities survey of Niamey, 2000, supervised by Adamou Nafoga, SAP, Niamey.

When we examine socio-economic variables we can see readily that Niamey is quite different from Marrakech. The variable measuring number of people in the household in Marrakech was significant (0.107) at the level of Urban Class and highly significant grouped by sampling point. (0.0037). Grouping by quartier was not really significant in Marrakech (0.2141) while in Niamey by contrast, the number of people in the household is not at all significantly linked to Urban Class (0.9314), quartier (0.6248) or sampling point (0.6366). This is highly suggestive and may indicate that in Niamey household size simply does not correlate with the three groupings even if it differs

at the macro level between communes. The high significance at the level of Urban Class of expenditure levels (D605_M with H_0 probability of 0.0001) and the number of times rural relatives visit per year (E435 with H_0 probability of 0.0002) do however suggest that some socio-economic indicators are closely tied to either quality of housing or time in the urban area. It is worth noting as well that five of the six socio-economic variables are highly significant grouped by sampling point and four of the six are highly significant grouped by quartier. Thus spatial and historical based strata are still highly valuable even when basic demography responds to other factors.

It may be noted that Gaborone is by African standards an extremely developed city which offers its inhabitants modern amenities and very high quality housing and urban transportation. Despite the overall high standard of living there are still many poor and a great variety in housing quality. The government of Botswana has also made attempts to shape the urban structure after a socialist ideology such that wealthy neighborhoods are dispersed throughout the urban landscape and overall much less segregated in terms of location than in most African cities.

Gaborone does not appear to be radically different however in terms of the success of our sampling strategy. Six of the nine housing related variables prove to be significant at the level of Urban Class while all nine are highly significant when grouped by ward and by sampling point. The most obvious exception at the level of Urban Class is h312 (H_0 = 0.7903) which measures the average number of persons per room.

Table 9. Kruskal Wallis statistics for Gaborone: housing variables

	Urban Class		ward		sampling points		
Variable	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Variable Description
h307	13.897	0.0163	62.063	0.0002	114.041	0.0001	type of habitation
h310	11.297	0.0458	67.339	0.0001	86.282	0.0001	# of rooms
h311	16.733	0.005	49.712	0.007	104.151	0.0001	Av m^2 of rooms
h312	2.408	0.7903	40.627	0.0581	59.729	0.0231	Av # of persons per room
h314	10.983	0.0517	87.261	0.0001	136.896	0.0001	Area of courtyard
h315	5.713	0.3351	91.266	0.0001	121.323	0.0001	number of WC
h317	9.829	0.0802	92.799	0.0001	120.306	0.0001	number of showers
h318	7.012	0.2197	87.235	0.0001	107.334	0.0001	Number of WC and showers combined
h320	10.462	0.0631	97.776	0.0001	130.544	0.0001	Number of faucets

Data Source: Six cities survey of Gaborone, 2001, supervised by Onalenna Selolwane, Department of Sociology and Musisi Nkambwe, Department of Envrionmental Sciences, University of Botswana, Gaborone.

That h312 is nevertheless highly significant at the quartier level and the sampling point level suggesting that crowdedness varies more by location than by habitation quality or age of development. The high significance of variable h311, h317 and h302 at the level of Urban Class suggests that the remote sensing classification does an excellent job discerning different housing types. That the same variables are extremely significant when grouped by ward and sampling point (0.0002 or 0.0001) suggests that even greater accuracy might be attained in the classification system though the full stratification system easily compensates for this slight crudeness.

Table 10 Kruskal Wallis statistics for Gaborone: socio-economic variables

	UrbanClass		im2 ward		im4 sampling points		
Variable	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Variable Description
hhsum	10.419	0.0642	48.971	0.0084	83.624	0.0001	expenses per HH
sumrev	22.559	0.0004	65.242	0.0001	103.329	0.0001	sum by HH of weekly revenue
totenergy	7.859	0.1642	71.673	0.0001	88.589	0.0001	sum of pula spent on energy per HH
m142	6.613	0.251	35.838	0.1468	52.183	0.1132	actual level of education
m165	15.384	0.0088	47.431	0.003	63.67	0.0076	rev/wk principal occupation
m2316	9.989	0.0455	54.757	0.0018	65.654	0.0086	satisfaction current life

Data Source: Six cities survey of Gaborone, 2001, supervised by Onalenna Selolwane, Department of Sociology and Musisi Nkambwe, Department of Envrionmental Sciences, University of Botswana, Gaborone.

For Gaborone we can examine for the sake of variety a somewhat different set of socio-economic variables, constructed from the baseline data collected. Four of the six are significant when grouped by Urban Class while five of the six are highly significant when grouped either by ward or by sampling point. Variables hhsum, sumrev and m165 (weekly revenue from household head's principal occupation) are highly significant grouped at the level of Urban Class suggesting that in Gaborone housing and income are closely linked. At the same time, despite Gaborone's tradition of intermixing housing of many income levels within the same ward the highly significant linkages of five of these six socio-economic variables both at the ward level and the sampling point level suggests that there are still major income differences between wards. Variable m142 which measures educational level of the household head is clearly less tied to any of the three groupings but it still shows up as mildly significant (0.1132) at the level of the sampling point. This probably means no more than that there are moderately significant differences in educational level that are only weakly linked to location of residence or time in the city.

In Tanzania, we chose to study Dodoma, the official capital but a much smaller city than Dar es Salaam. Dodoma is located in the center of Tanzania and though it is planned for a large population its actual population is only guessed at since it has been some decades since Tanzania has had a census. We thus have no comparable data with which to compare our survey data and the housing due to its distance from the coast and major construction industry is particularly homogeneous. There are some apartment complexes but they are small and like the other housing which varies in size and quality are capped by a tin roof making housing look similar from a satellite's vantage point. This meant that we were able to construct only a minimal set of housing classes even when incorporating our change maps from multiple images. Dodoma may thus be seen as a test or limiting case for the methodology. The rather significant success of the methodology in the Dodoma case despite these problems, as the following tables illustrate, may be attributed to the distribution of sample points across the urban space as much as to the urban classification itself. The computer algorithms ensure that each pixel of urban habitation has a chance of selection and that those chances are spread across the entire urban fabric.

Table 11. Kruskal Wallis statistics for Dodoma: housing variables

	Urban Class		ward		sampling points		
Variable	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Variable Description
H306	3.13	0.2097	68.331	0.1899	57.298	0.071	# yrs in Dodoma
H310	7.996	0.0184	86.587	0.0213	98.037	0.0001	# of rooms
H311	4.799	0.0908	44.759	0.8804	41.671	0.529	Av m^2 of rooms
h313	2.793	0.1552	57.75	0.2398	82.175	0.0002	Is there a courtyard
H314	0.055	0.814	44.199	0.1131	37.135	0.2842	Area of courtyard
H315	8.602	0.0136	70.019	0.0207	70.023	0.0023	number of WC
H316	4.25	0.1194	130.701	0.0001	116.842	0.0001	type of facilities
H317	8.97	0.0113	91.759	0.0006	76.429	0.0009	number of showers
H318	11.851	0.0027	34.147	0.3648	42.677	0.2402	Number of WC and showers combined
H319	0.243	0.8857	93.238	0.003	94.618	0.0001	Is there running water
H320	4.874	0.0874	62.6	0.0039	59.087	0.0035	Number of faucets

Data Source: Six cities survey of Dodoma, 2001, supervised by Elifuraha Mtalo and Manoris v. Meshack, UCLAS, University of Dar es Salaam, Dar es Salaam, Tanzania.

Roughly half (6) of the eleven housing variables presented in Table 11 are significant when grouped by Urban Class. The urban classification scheme was limited to distinguishing only one habitation class by its features with a second habitation class distinguished based on the change map of the city. It might have been possible to improve this by extensive ground truthing but we felt it might be more informative methodologically not to do this. Thus habitation pixels appearing chronologically after the first image became change pixels and were sampled as a second habitation class. Thus the Urban Class variable has only two values (1 and 2). Therefore all significance showing up at the level of urban classification is due to differences in the period of urban development or settlement. Clearly higher resolution imagery might solve the problem of Dodoma's superficial homogeneity. At the same time, the high significance of seven of the eleven variables when grouped by sampling point and the poorer showing of the grouping by ward (only five of eleven are significant) suggests that housing varies dramatically within wards but is fairly homogenous within the small areas surveyed around sampling points.

Table 12. Kruskal Wallis statistics for Dodoma: socio-economic variables

	Urban Class		ward		sampling points		
Variable	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Chi-squared with ties	Probablity H_0	Variable Description
E403	1.581	0.2086	64.298	0.1177	71.717	0.0021	# times per week trash collected
E409	0.857	0.6515	34.944	0.3758	46.309	0.1403	amount for garbage service
E435	2.283	0.3194	49.657	0.4469	48.672	0.1634	# times rural relatives visit per year
E436	5.656	0.0591	63.208	0.0303	47.747	0.1869	length of visits by relatives
H312	13.56	0.0011	74.535	0.1142	64.354	0.0243	Av # of persons per room

Data Source: Six cities survey of Dodoma, 2001, supervised by Elifuraha Mtalo and Manoris v. Meshack, UCLAS, University of Dar es Salaam, Dar es Salaam, Tanzania.

It should not be surprising that with this more simplified urban classification a number of basic socio-economic variables can be found that do not seem to be significant at the level of Urban Class. Perhaps the most significant point to be made from Table 12 is to note that variable H312, which measures the average number of people per room is highly significant at the level of Urban Class and sampling point but only slightly significant when grouped by ward. This would suggest large variability in crowdedness within individual wards and only moderate differences between wards but major differences on this dimension between recent and older urban developments in Dodoma which carry over into the grouping at the level of sampling points.

It is also worth noting that the length of visits to urban relatives by rural relatives (E436: 0.0591) is significantly linked to the distinction between recent and more established urban areas reflected in the Urban Class figures and that this is even more significant at the ward level but not well captured at the level of the sampling point. We might understand this if we first note that variable E435 which measures number of times rural relatives visit is not well captured by any of the three groupings. Some residents have economic activities in both the rural and urban areas in different seasons and so have visitors who come in the off season just as they visit the rural area during the prime agricultural season. This might be seen as suggesting that the diachronic character of this version of Urban Class picks up some of the seasonal variation in ties to the rural area and that some wards, similarly, are more likely to be inhabited by people engaged in such seasonal activities. By contrast, the variations in this variable between sample points are apparently not much greater than within sample points perhaps because too many sample points do not include such people - i.e. the six households they represent are too small to capture this variation.

While only two of the five variables are highly significant when grouped by sampling point it is equally clear that the overall average level of significance for the five variables is much higher at the sampling point level than at either of the other two groupings. No variables are clearly insignificant at the level of sampling point while Urban Class and Ward groupings each have at least two variables (E409 and E435) that seem to have no significant probability of being tied to them. This clearly suggests that there is a complementarity between the urban classification and the sampling strategy based both on that classification and a careful distribution of sampling points across the urban field.

One might tentatively conclude, that in the Dodoma case while the sampling strategy does a good job with the housing variables despite having only change and non-change urban classes this may carry over into less precision at the level of distinguishing socio-economic variables. This is what one would expect if there are clear relationships between quality of housing and a variety of other issues of interest to social science. Thus the Dodoma case, seen as a limiting case for the methodology, seems to suggest that major improvements can be expected from a more expanded set of urban classes. An alternative possibility that only further analysis could confirm is that the overall range in socio-economic indicators is much narrower in Dodoma than in the other cities examined because it has both less urban development and a shorter history as an urban center than the other cities studied.

Specific methodological implications

This project began with the idea that a diachronic remote sensing based urban classification could be the basis of an exemplary stratified sampling technique in urban areas. The preceding analysis fully supports that conclusion across six rather different cities. It further suggests a number of interesting but tentative conclusions. We should note that this project was primarily intended to test a methodology, had a budget of less than $100,000 per city and relied primarily on 10m and 20m imagery. Better imagery can obviously be used today but it is important to stress that the key advantage of remote sensing is classification and a consistent and precise coordinate system. It is probable that there are diminishing returns to increases in resolution and that current one meter imagery might be close to ideal for classification purposes. Spectacular imagery is pretty but implies a tradeoff in terms of pixels that must be classified and, in the end, amalgamated to a small number of useful classes.

At a minimum the project data allows us to say with some confidence that:

- An urban classification using 10m and 20m imagery is quite effective in classifying residential areas in terms of habitation despite the resolution being too low to provide picture quality imagery of the habitations.
- The greater the variety of habitation types the easier it is to make defensible distinctions between urban classes and the more urban classes that can be created the better the resulting sample can be.
- Each city is different and we must expect some differences even in the same socio-economic variables that prove to be linked to housing variables or urban classifications in each city though there are many commonalities in the cases examined.
- There appear to be fairly consistent linkages between housing quality and revenue and expenditure streams across all socio-economic groups. Yet, there are differences between cities in the degree to which such variables link with urban neighborhoods.
- While, overall, the sampling strategy appears highly successful, the details of the Kruskal Wallis tests suggest strongly that the urban dynamic in each city has significant differences that should give rise to more serious research. Even basic demographic factors appear to differ significantly from city to city.
- This analysis should be seen as barely scratching the surface of the data and efforts should be made to do a variety of comparative analyses as well as to contextualize the major differences between urban dynamics.
- The simplified urban classification available for Dodoma, involving only the distinction between change and non-change pixels, has the advantage of making it clear that this distinction is, in and of itself, a significant one. The classifications in the other cities can of course be reduced to a change and non-change classification and this might be a useful exercise to further test the importance of this distinction but we have already clearly validated the importance of diachronic imagery.

Concluding Remarks

Near Term Goals

We would like to generalize the methodology whose initial development has been funded by NSF to facilitate mapping, monitoring, and planning of urban land-use in other cities. Urban sampling points and urban classes can facilitate household surveys of health, nutrition, economic livelihood strategies, urban infrastructure usage or any other issue of critical social or environmental importance. If the data is then placed in a GIS that also incorporates a variety of digital data such as location of infrastructure and sources of pollution, this will allow the appropriate agencies to track issues that might be relevant to the entire city or ones that might be specific to particular locales or types of habitation. It is well known, for example, that some disease vectors in urban areas can be correlated with habitation types or water sources while many other urban problems are linked to habitation, locale and proximity to such things as garbage dumps and other pollution sources.

Advantages of Remote Sensing and our Methodology

This basic methodology has the clear advantage of making it possible to easily quantify historical urban land use change and to accurately sample the entire urban area. This approach also allows us to create a stratified sample that is of minimum size but is highly representative and thus is very cost effective. The GIS database produced through these techniques can be easily accessed by various agencies from government offices to NGO teams and can be accumulated such that many years worth of surveys can be studied to discern diachronic trends.

Cost, Accuracy, Repeatability, and Representativeness

Many different surveys on different topics can be accommodated using the same urban classes and thus made comparable one with the other. Perhaps as important, by applying a uniform basic methodology in many cities it will be possible to make comparative analyses between cities on similar issues. This is a key emphasis of our approach because it is critical to involve teams of researchers from multiple countries in the analysis of urban data and the tracking of key urban issues related to health, nutrition, and urban livelihood. Far too often good research is done and yet only some of the key conclusions are made broadly known while many useful aspects of the data collected disappear into the personal data banks of researchers or NGOs and are never reexamined by others. This project proposes to make key data available long term for all appropriate uses. We are still in the process of preparing the data from the initial six cities.

Current Cities Funded through NSF

Our NSF funding ($575,000) enabled us to begin the development of this system in six cities in Africa (Marrakech, Dakar, Bamako, Niamey, Dodoma and Gaborone). We have encountered such enthusiasm from our African colleagues that we would like to broaden and strengthen this effort in three ways- 1) extend the study to additional cities in the six countries, 2) add new countries in the near term, and 3) add new (and recently released) high resolution Ikonos 1-4 meter satellite imagery to our analysis. Remote sensing image analysis for urban land-use classification is highly dependent on the spatial resolving power of the satellite system utilized. Though our work was pioneering in many ways, it relied on images at 10 to 20-m spatial resolution, which was

the best available at the time, and our efforts can clearly benefit from the added resolution now available. This new imagery could be used to refine the current urban land-cover classes and these refinements can then be extended to better interpret the historical satellite data from the 1980s.

General Applicability of the Targeting System

It is important to remember both that different issues may be key concerns in different cities and that a subset of issues will be important in many different cities. Thus malaria, AIDS, dengue fever, yellow fever, and other epidemic diseases linked to specific vectors or behaviors are not each uniformly significant across Africa or the Near East but health related behaviors and environmental issues related to health and disease are important concerns everywhere. Similarly, livelihood activities and linkages between nutritional status, consumption behaviors and living patterns are of broad concern in all target cities. This methodology recognizes that today's most urgent concerns may not be tomorrow's most urgent concerns and that the key to targeting any urgent urban issue involves accurate and cost effective assessment. The ability to track key behaviors in all their urban complexity both in the here and now and over time will be critical to any serious attempts to solve the problems of poor urban areas. Most projects have time frames shaped by short-term funding and immediate urgency. By contrast our hopes are to build targeting and monitoring capacity across Africa and the Near East in such a way that comparable long term and useful data can be both collected and preserved for the use of scientists and government planners.

Capacity Building in Africa and the Near East

Capacity building in Africa and the Near East will involve the development of expertise to continue refinement of the remote sensing methodology as well as enrichment of the survey data. Just as important, this capacity building will have to involve funding for cooperative research among scholars from all participating countries. We believe the best way to do this is to dedicate long term funds for these purposes so that participating countries can count on a viable level of funding over the long term and the investments in data collection and analysis can produce long term benefits. In the near term, we hope that capacity building in our participating institutions will allow all groups to contribute the majority of the research, satellite image processing and GIS construction for their own cities: thus devolving these major responsibilities from the U of A to the other institutions.

Acknowledgments

We would like to acknowledge the National Science Foundation for its funding of this project (1998-2003) under a comparative urban initiative and for its support of a project conference in Dakar in January 2003 where an initial version of this paper was presented. We would also like to thank in particular our colleagues in Africa: Ahmed Belasri and Abdellah Bencherifa in Morocco, Magatte Ba and many others at the Centre de Suivi Ecologique in Dakar, Sadio Traore and Moïse Balo in Bamako, Adamou Nafoga and his staff at the Structures d'Alerte Précoce (Cabinet du Premier Ministre) in Niamey, Manoris V. Mechack and Elifuraha Mtalo at University College of Land and Architectural Studies in Dar es Salaam and last but not least Onalenna Selolwane and Musisi Nkambwe of the University of Botswana in Gaborone, We owe thanks as well to colleagues at the University of Arizona, especially Stuart Marsh and Jim Greenberg for their input in thinking through the methodology.

Abstract

This paper provides a statistical evaluation of the methodology of the NSF funded Six Cities Project. The project develops a methodology for surveying densely inhabited areas by processing diachronic remote sensing imagery to create habitation strata or urban classes. These classes become part of a sampling strategy which gives every pixel associated with habitation a specified chance of selection and then draws a representative sample of pixels. These pixels become center points for household surveys which can study a variety of issues including health, environment, livelihood strategies, demographics and household labor, expenditures and income. The methodology lends itself to GIS construction and the generation of data that can be easily compared and can be of maximal use to municipalities, governments, scholars and NGOs. It also provides a long term basis for inexpensive surveys that can have a high claim to reliability and representativity.

Key words: remote sensing, urbanism, survey methodology, National Science Foundation, health, environment, livelihood strategies, demographic, household labor, expenditures, income, Africa, Middle East, Morocco, Senegal, Mali, Niger, Tanzania, Botswana, Marrakech, Dakar, Bamako, Niamey, Dodoma, Gaborone.

Resumé

Cet article pourvoit une évaluation statistique de la méthodologie du projet Six Cities financé par NSF. Le projet développe une méthodologie pour l'enquête des endroits à habitation dense à travers l'usage des images satellitaires à fin de créer des strates d'habitation ou des classes urbaines. Ces classes deviennent une part d'une stratégie d'échantillonnage qui donne à chaque pixel correspondant à une zone d'habitation une chance déterminée d'être sélectionné et procède à un tirage de pixel pour créer un échantillon représentatif. Chaque pixel devient le centre point d'une enquête de ménage qui peut étudier une variété de sujets y inclus: la santé, l'environnement, les stratégies de survie, la démographie, le travail au ménage, les dépenses et les revenus. La méthodologie se prête à la création d'un SIG et la génération des données facilement comparable et à haute utilité aux municipalités, aux gouvernements, aux chercheurs et aux ONGs. Elle pourvoit aussi une base à long terme pour les enquêtes à bon prix qui puissent prétendre à une fiabilité et représentativité élevée.

Mots clefs : la télédétection, l'urbanisme, la méthodologie, l'enquête, National Science Foundation, la santé, l'environnement, les stratégies de survie, la démographie, le travail au ménage, les dépenses, les revenus, l'Afrique, le Moyen Orient, le Maroc, le Sénégal, le Mali, le Niger, la Tanzanie, le Botswana, Marrakech, Dakar, Bamako, Niamey, Dodoma, Gaborone.

Resumen

Este artículo proporciona una evaluación estadística de la metodología del proyecto financiado por NSF de seis ciudades en Africa. El proyecto desarrolla una metodología para examinar áreas de habitación densa a traves del procedimiento diacrónicó de imágenes remotas con el fin de crear estratos de habitación o clases urbanas. Estas clases forman parte de una estrategia de ejemplar que da cada pixel asociado a la habitación una probabilidad especificada de la selección y después selecciona una muestra representativa de pixeles. Estos pixeles se convierten en puntos de centro para las encuestas de casa que pueden estudiar una variedad de problemas incluyendo salud, ambiente, las estrategias del sustento, labor y demográficas de casa, y los gastos e ingresos. La metodología se presta a la construcción y la generación de los datos que se pueden comparar fácilmente y pueden estar de uso máximo a los municipios, a los gobiernos, a los investigadores académicos, y a los ONGs. También proporciona una base a largo plazo para los reconocimientos baratos que pueden tener una alta pretensión a la confiabilidad y a la representatividad.

Palabras claves: detección remota, urbanismo, metodología del reconociementos, National Science Foundation, salud, ambiente, estrategias del sustento, demográficas, labor de la casa, gastos, ingresos, África, Medio-este, Marruecos, Senegal, Malí, Niger, Tanzania, Botswana, Marrakech, Dakar, Bamako, Niamey, Dodoma, Gaborone.