West Africa Spatial Analysis Prototype Exploratory Analysis (WASAP)

Cluster Typing Procedures

Demographic and Health Surveys
Macro International Inc.

DHS

West Africa Spatial Analysis Prototype
Exploratory Analysis
(WASAP)

Cluster Typing Procedures

Shea O. Rutstein

Macro International, Inc.
Calverton, Maryland, USA

February 2000

This paper is one of a series of exploratory analyses conducted under the West Africa Spatial Analysis Prototype Project. Funding for this project was provided by the United States Agency for International Development's Regional Economic Development Services Office for West and Central Africa.

I. Introduction

A. The WASAP project

The West African Spatial Analysis Prototype is an attempt to gather existing demographic data collected through the Demographic and Health Surveys of countries in the West Africa region, identify their geographic coordinates and place them within a geographic database and analysis system (geographic information system—GIS). These data are then to be integrated with non-demographic data for geographic areas, such as climate, agriculture, etc. to provide a more complete understanding and context to changes in demographic phenomena (births, deaths, migration, population distribution), and the health and welfare of the population living in West Africa. This understanding will help to orient, plan, monitor and evaluate programs to improve the health and welfare of the West African region. Beyond the existing DHS data, a new series of surveys is being undertaken in the region to monitor changes and to evaluate the impact of programs. However, because of changes in the sampling frames between the earlier round of surveys and the current and upcoming round, the same areal units cannot be used.

A principle intention of the WASAP project is to study the impact of the establishment and improvement of family planning and health facilities on trends in several population and health related statistics, such as use of health services, contraceptive prevalence, and infant and child mortality. Ideally, to evaluate this impact repeated measurements should be done in the same area units noting the change in availability of services and the resulting change in use and outcomes. However, since the first round surveys included in the WASAP project were not originally planned for re-survey, new area units were and will be used in subsequent surveys. There is therefore a need to identify groups of "comparable" area units in earlier and later surveys, in order to track changes over time within these comparable groups. One possibility (and the only one easily open to us at this point in time) is to use the survey information itself to identify comparable groups of areal units. This research note outlines a procedure for producing these comparable areal units and provides a formula for assigning both old and new units to the comparable groups.

II. Data

The data used for this project come from the WASAP GIS data set[1]. These are household-and individual-level data from Demographic and Health Surveys, which have been aggregated for each sampling cluster (ultimate sampling area). The data from all surveys in the WASAP area have been pooled both across countries and across time to form a single data set but with identifiers as to both country and survey. The aggregated data typically are in the form of percentages of the households and population in each cluster that has a certain characteristic (e.g. percent of households with electricity, percent of women 15-49 years of age with completed primary education).

[1] The WASAP data includes DHS surveys conducted in West Africa between 1988 and 1996, the West Africa Long Term Perspective Study database (WALTPS), and the Consultative Group on International Agricultural Research (CGIAR) International Center for Research in Agroforestry (ICRAF) in Nairobi.

III. Process

The design of the classification process is fairly simple, but somewhat complicated to implement. The first step is the selection of classificatory characteristics by which areas will be distinguished from one another. These characteristics should be those that are likely to remain stable over the time between surveys and will not be the subject of trend analysis. The next step is to form an index from the different characteristics. The third step is to divide the index into more or less homogeneous groups and then assign each area unit to its corresponding group based on the index value.

Complications arise in the selection of the appropriate classificatory variables, given that 7in many areas the values of the classificatory variables may change with time and because the same variables have not always been used in all the surveys. Thus a minimum set of classificatory variables needs to be determined, which are included in all surveys. A further complication is how to handle areas that do not have a specific piece of information, even though it was to have been collected in the survey (i.e. missing data). Much of the areal information is in the form of the proportion of households with a given characteristic. If information on the characteristic was not recorded for all the households in the area, then it may not be possible to calculate the proportion of households with that characteristic.

Regarding index formation and classification, there are several approaches. One approach is to use "k-means or hierarchical clustering", a statistical technique that minimizes the statistical "distance" (sum of differences) between data points, based on the selected classificatory variables. This technique would seem to be ideal for the task except for a large drawback: the process needed is to be able to classify areas in future surveys. The available hierarchical clustering techniques do not provide means of classifying areas not included in the analysis.

A simple ad-hoc index can be formed from the various classificatory variables, but the results are likely to vary by researcher because the weights assigned to each variable are arbitrary. A better method is to form an index with statistically assigned weights.

The approach that is taken here is to use the principal-components methodology of factor analysis to generate the weights. In the factor analysis model, there is an assumption that there are basic underlying factors (principal components), which cannot be directly measured, and which are partially represented by measurable characteristics. By combining these characteristics, the underlying factors can be estimated. The principal components are formed by linear combinations of the measured variables with coefficients (factor weights) assigned according to the correlation of each variable to the underlying factor.

There are problems with this approach when not all data points (areas in our case) have a value for all the variables. As noted above, if the variable is not used for a large group of areas, then that variable cannot be included in the analysis. However, if only a relatively few areas do not contain values (i.e. missing data), then these areas can be discounted in the estimation of the index weights either by exclusion or assignment of the mean value of all areas.

Even after estimating the value of the index weights, there remains the problem of estimating the index value for areas with missing data. One possible approach would be to use the mean for all areas. However, this approach has the drawback that most areas with missing data are unlikely to be "average" areas. Another approach, the one adopted here, is calculate an index score based on the valid data for areas with missing values and

then to re-scale the index score according to the ratio of the potential score for the non-missing items and the potential score for all items.

IV. Results

A total of 3,345 clusters were used from DHS surveys conducted between 1986 and 1996. Cluster types were calculated separately for urban and rural areas. Two algorithms are examined for cluster typing, a one-factor procedure and a two-factor procedure.

A. Classificatory variables

The variables selected for classifying areas are

- Type of area of residence (urban-rural)
- Proportion of female-headed households
- Proportions of households
  - with electricity
  - with water piped into the residence
  - using a public tap or well
  - with their own or a shared flush toilet
  - with natural (dirt, sand, dung) flooring
- Proportions of the household population
  - under 15 years of age
  - 65 years of age or over
- Proportions of women 15-49 with
  - no education
  - 1-3 years of education
  - 4-6 years of education
  - 7 or more years of education
- Proportion of women 15-49 who are able to read
- Proportions of currently married women 15-49 with
  - no education
  - 1-3 years of education
  - 4-6 years of education
  - 7 or more years of education
- Proportion of women 15-49 who currently work
- Proportions of women 15-49 whose occupation is
  - professional, technical, clerical
  - in sales and service
  - skilled or unskilled manual labor
  - agricultural
- Proportion of women 15-49 who
  - were never married or living in a consensual union
  - are currently married or living in a consensual union
  - were formerly married or living in a consensual union

These variables were chosen since they are available in almost all the surveys and are likely to remain fairly stable over the time between surveys (up to 15 years). Additional characteristics of areas beyond those included in the DHS would have been useful, such as proximity to the coast, rivers, major roads, amount of rainfall, type of agriculture, etc. However, they were not available in the WASAP data set at the time of conducting this analysis. When such data become available, the procedures outlined below should be re-implemented to include them.

It is apparent that some of the classificatory variables are closely related to one another, such as education of all women and of currently married women. This level of association does not invalidate the procedures used and using all the variables provides

additional information compared to omitting some of these variables. However, using closely related variables does not provide as much additional information as would using other non-related variables.

Certain characteristics, which at first glance would appear to be useful for classification, can not be included as classificatory variables. They may be the subject of the impact analysis (e.g., fertility rates, mortality rates, immunization rates and nutritional status), or were not included for almost all the countries, or were thought not to be stable enough over the likely period between surveys. Thus, possessions, such as radios, televisions, vehicles, etc. were excluded as classificatory variables because they are likely to change over time.

B. The one-factor procedure for classifying areas

Because the urban-rural distinction is so basic factor scores are calculated within each area separately. Table 1 shows the means, standard deviations, factor score coefficients and correlations with the first two factors for each of the component variables. In the one-factor model, the score for factor 1 alone is used to define classifications separately for urban and rural areas, based on quintiles within each area. This procedure yields ten cluster types. While the goal of the classification procedure is not necessarily to rank the clusters by socio-economic status, the higher values of the factor 1 score tend to indicate a higher socio-economic status, which is not unexpected given the classificatory variables used.

For urban areas, the maximum theoretically possible factor score was 5.78 while the minimum was –4.03. This was divided into 5 cluster types by taking quintiles of the scores: less than 3.15 for type 1, 3.15 to 3.72 for type 2, 3.72 to 4.36 for type 3, 4.36 to 5.10 for type 4, over 5.10 for type 5. For rural areas, the maximum theoretically possible factor score was 10.49 while the minimum was –1.40. This was divided into 5 cluster types by taking quintiles of the scores: less than 0.55 for type 1, 0.55 to 0.82 for type 2, 0.82 to 1.37 for type 3, 1.37 to 2.44 for type 4, over 2.44 for type 5.

Certain countries did not have all the variables necessary for a full factor score, and their scores were re-scaled. The 1988 Togo survey did not ask about electricity; the 1986 Liberia survey did not ask about electricity or type of flooring; the Nigeria survey in 1990 did not ask about type of water supply; the 1996 Benin survey did not ask about type of toilet/latrine; and the surveys in Ghana in 1988 and Mali in 1987 did not ask about women's occupation.

Using the factor score coefficients as weights, clusters added from future surveys can be classified into one of the 10 types (5 urban and 5 rural).

Table 2 shows the distribution per country, survey and type of area of the cluster types. While the overall number of clusters of each type is approximately the same in each area, the distribution of clusters by type for each country differs markedly. For example, Nigeria has 52 type-five and 14 type-one urban clusters while Niger has 4 type-five and 37 type-one urban clusters.

Table 3 shows some results according to cluster type and area. It is interesting to note that with a one-factor model there is a natural order of results. Thus the use of modern methods of contraception among married women rises from 3.5% to 16.8% in urban areas and from 0.4% to 8.6% in rural areas as the factor score quintile increases. The other variables also follow in order as well. These results show that the one-factor typing of clusters is essentially socio-economic.

C. Two-factor procedure

A second approach, which may be thought to allow greater distinction between clusters, is to use a second factor in the classification. Factor analysis has as its basis the identification of underlying factors which are not directly observable but which are correlated with observable variables. Principal components methodology extracts these factors in a way that they are unrelated to each other. Thus in theory, the addition of a second factor would increase the amount of information available for classification. In the application here of a two-factor model, the two factors were generated separately for urban and rural areas. Within each area, clusters were classified into four groups, being above or below the median value on each factor. The table below presents the median cutoff points for both factors in each area, after re-scaling for missing variables:

Type of area	Factor 1	Factor 2
Urban	4.01	4.98
Rural	1.03	13.98

Clusters are then assigned to one of eight types depending on the area and whether they score above or below the median on each factor.

Table 4 shows the results of the two-factor cluster typing. This two-factor procedure produces lopsided results for some countries, indicating that adding a second factor may be picking up country characteristics in addition to cluster characteristics.

Table 5 shows some substantive results according to the two-factor typing. Note that there is no natural order of clusters in this procedure since it is arbitrary how the cluster type numbers are assigned when two factors are used.[2] In the table the rows have been arranged in order of increasing use of modern contraception. Once this is done, the other variables essentially follow the pattern. The range of variation within each type of area is reduced because of the fewer number of cluster types and because for these variables the two-factor method does not appear to distinguish clusters better than the one-factor method.

V. Conclusion

While data available in the WASAP DHS data set can be used for cluster typing, there are many limitations for their use in determining trends according to cluster type. Because the DHS I surveys did not contain education information for adults and children on the household schedule and because many of the surveys in the WASAP data set were conducted during that phase, that information can not be used for cluster typing.

Other initially attractive survey items, such as household possessions, should not be used for time-trend comparisons, since changes in their values, which is expected over time, would lead to misclassification of clusters, and obviously potential analysis variables similarly cannot be used.

Omission in specific surveys of other items commonly used, such as electricity and women's occupation, may also lead to inaccurate cluster typing since some form of compensation needs to be employed, such as assumption of mean values or re-scaling of indexes.

For cluster typing, a preferable post-hoc procedure is that of Hierarchical or K-way Cluster Analysis, but this procedure cannot be used pre hoc, as is the goal here.

[2] Possibilities include (low-low, low-high, high-low, high-high) or (low-low, high-low, low-high, high-high).

However, using comparable survey questions and the same generated factor score coefficients, a broad impact and trend analyses can be made. Care must be used, however, to ensure that there are enough sampling areas of each type for the results to be significant.

Table 1. Cluster classificatory variables, mean, standard deviation, factor score coefficients and correlation with factors

WASAP Variable	Description	Mean	Standard Deviation	Factor Score Coefficient		Factor Correlation Coefficient	
				Factor 1	Factor 2	Factor 1	Factor 2
	Urban Areas						
WHEADP	Proportion of female-headed HH	21.290	15.591	.049	.126	.360	.348
WELECTRP	Prop. W/electr. in the cluster	39.420	35.408	.096	-.085	.701	-.235
WPIPEDP	Prop. W/piped water in residence	24.973	29.847	.080	-.118	.581	-.327
WTAPP	Prop. Using public tap or well	54.224	30.614	-.066	.110	-.481	.304
WTOILETP	Prop. w/own or shared toilet	12.823	22.222	.087	-.102	.631	-.282
WFLOORP	Prop. HH w/natural flooring	16.695	27.765	-.067	.141	-.490	.389
WHHAG0P	Prop. below 15 years	43.939	6.861	-.057	.004	-.412	.011
WHHAG65P	Prop. over 64 years	2.955	2.819	-.020	.043	-.145	.120
WEDUCW0P	Prop W 15-49 w/0 year of educ.	45.337	27.422	-.123	-.121	-.896	-.336
WEDUCW1P	Prop W 15-49 w/1-3 yrs of educ	6.833	7.323	-.026	.243	-.187	.672
WEDUCW4P	Prop W 15-49 w/4-6 yrs of educ	18.757	13.131	.043	.175	.313	.485
WEDUCW7P	Prop W 15-49 w/7+ yrs of educ.	29.075	25.237	.119	-.030	.865	-.082
WEDUMW0P	Prop CMW 15-49 w/0 year of ed.	51.353	29.324	-.117	-.122	-.852	-.337
WEDUMW1P	Prop CMW 15-49 w/1-3 yrs of ed	6.517	8.820	-.015	.236	-.108	.652
WEDUMW4P	Prop CMW 15-49 w/4-6 yrs of ed	17.414	14.960	.049	.159	.356	.440
WEDUMW7P	Prop CMW 15-49 w/7+ yrs of ed.	24.722	25.586	.111	-.035	.806	-.095
WLITERP	Prop. of women 15-49 who reads	47.274	26.734	.124	.062	.904	.171
WWORKP	Prop of women currently work.	51.525	23.521	.017	.122	.127	-.205
WOCCUP1P	Prop in professional/technical	4.651	6.021	.071	-.074	.514	-.021
WOCCUP2P	Prop in sales & service indust	38.583	16.587	-.024	-.008	-.173	-.091
WOCCUP3P	Prop skilled/unskilled manual	7.694	8.340	.021	-.033	.155	.494
WOCCUP4P	Prop in agricultural field	6.169	13.204	-.034	.179	-.247	.336
WNMWP	Prop never married	26.284	15.661	.089	.038	.646	.105
WCMWP	Prop currently married	65.621	16.902	-.089	-.085	-.648	-.236
WFMWP	Prop of formerly married women	8.090	7.660	.015	.111	.110	.307
	Rural Areas						
WHEADP	Proportion of female-headed HH	15.214	16.476	.085	.011	.661	.029
WELECTRP	Prop. w/electr. in the cluster	3.355	11.930	.055	-.153	.428	-.416
WPIPEDP	Prop. w/piped water in residence	1.816	7.957	.031	-.184	.243	-.499
WTAPP	Prop. using public tap or well	58.144	39.061	-.045	-.069	-.350	-.187
WTOILETP	Prop. w/own or shared toilet	1.030	5.902	.040	-.176	.311	-.477
WFLOORP	Prop. HH w/natural flooring	60.036	36.050	-.086	.096	-.675	.262
WHHAG0P	Prop. below 15 years	48.139	5.848	-.020	.010	-.154	.027
WHHAG65P	Prop. over 64 years	4.254	2.965	.003	.010	.020	.028
WEDUCW0P	Prop W 15-49 w/0 year of educ.	73.443	28.583	-.123	-.032	-.964	-.086
WEDUCW1P	Prop W 15-49 w/1-3 yrs of educ	5.894	7.602	.053	.233	.413	.634
WEDUCW4P	Prop W 15-49 w/4-6 yrs of educ	9.683	12.548	.085	.082	.662	.224
WEDUCW7P	Prop W 15-49 w/7+ yrs of educ.	10.981	19.634	.105	-.097	.820	-.263
WEDUMW0P	Prop CMW 15-49 w/0 year of ed.	75.942	27.536	-.120	-.041	-.940	-.110
WEDUMW1P	Prop CMW 15-49 w/1-3 yrs of ed	5.770	8.362	.051	.230	.401	.623
WEDUMW4P	Prop CMW 15-49 w/4-6 yrs of ed	8.826	13.005	.081	.081	.632	.221
WEDUMW7P	Prop CMW 15-49 w/7+ yrs of ed.	9.464	18.661	.098	-.100	.767	-.270
WLITERP	Prop. of women 15-49 who reads	18.188	22.995	.116	-.037	.911	-.100
WWORKP	Prop of women currently work.	58.905	29.312	.025	.172	.199	.466
WOCCUP1P	Prop in professional/technical	.887	2.815	.044	-.109	.348	-.297
WOCCUP2P	Prop in sales & service indust	22.173	18.179	.001	-.114	.007	-.309
WOCCUP3P	Prop skilled/unskilled manual	6.118	10.629	.003	-.069	.021	-.188
WOCCUP4P	Prop in agricultural field	37.397	27.595	.018	.241	.143	.654
WNMWP	Prop never married	12.789	11.331	.075	-.044	.586	-.119
WCMWP	Prop currently married	81.658	14.285	-.095	.008	-.743	.020
WFMWP	Prop of formerly married women	5.553	7.310	.070	.053	.544	.145

Table 2. Distribution of cluster type and type of area by country and survey, one-factor procedure

Type of place of residence		Cluster Type					Total
		1	2	3	4	5	
Urban	Burkina Faso 1993	26	32	29	13	10	110
	Benin 1996	26	18	22	18	8	92
	Ctrl. Af. Rep. 1995	26	34	33	13	2	108
	Cote d'Ivoire 1995	17	34	35	35	25	146
	Cameroon 1991	10	3	3	22	42	80
	Ghana 1988	1	1	4	22	42	72
	Ghana 1993	2	6	20	44	78	150
	Liberia 1986	61	29	8	3	1	102
	Mali 1987	8	17	19	14	2	60
	Mali 1995	52	35	16	15	0	118
	Nigeria 1990	14	9	16	41	52	132
	Niger 1992	37	28	26	10	4	105
	Senegal 1986	12	18	20	12	8	70
	Senegal 1992	12	32	30	29	29	132
	Togo 1988	4	13	28	18	3	66
	Total	308	309	309	309	308	1543
Rural	Burkina Faso 1993	54	34	24	7	1	120
	Benin 1996	7	27	44	25	5	108
	Ctrl. Af. Rep. 1995	0	6	48	63	6	123
	Cote d'Ivoire 1995	3	8	30	39	20	100
	Cameroon 1991	9	12	4	14	30	69
	Ghana 1988	1	4	3	19	51	78
	Ghana 1993	3	18	18	69	142	250
	Liberia 1986	0	0	4	15	35	54
	Mali 1987	22	29	34	3	0	88
	Mali 1995	90	57	28	7	0	182
	Nigeria 1990	32	28	15	41	50	166
	Niger 1992	77	36	16	1	0	130
	Senegal 1986	26	53	33	7	0	119
	Senegal 1992	35	41	26	22	2	126
	Togo 1988	1	7	33	28	18	87
	Total	360	360	360	360	360	1800

Table 3. Substantive results by cluster type, one-factor procedure

Type of place of residence	Cluster Type	Factor Score	Prop women 15-49 Muslim	Prop of W. 20-49 w/birth by 20	Mean number of CEB to all women	Prop total women using a contra-ceptive method	Prop total women using modern contra-ception
Urban	1	2.8	62.0	61.6	3.6	8.4	3.5
	2	3.4	57.5	58.6	3.1	14.0	7.1
	3	4.0	47.7	54.5	2.7	18.7	10.2
	4	4.7	36.1	47.3	2.5	24.5	13.0
	5	5.6	13.7	36.8	2.0	34.0	16.8
	Total	4.1	43.6	51.8	2.8	19.9	10.1
Rural	1	0.4	82.9	67.2	4.1	5.4	0.4
	2	0.7	71.8	64.1	3.9	6.4	0.8
	3	1.1	48.3	61.0	3.7	7.5	1.8
	4	1.8	19.2	57.8	3.5	11.0	3.9
	5	3.2	5.8	58.0	3.2	18.6	8.6
	Total	1.5	44.5	61.6	3.7	9.8	3.0

Table 4. Results of two-factor cluster typing

	Urban Areas				Rural Areas			
	1	2	3	4	5	6	7	8
Burkina Faso 1993	36	24	39	11	67	6	38	9
Benin 1996	11	4	41	56	55	50	0	3
Ctrl. Af. Rep. 1995	2	1	75	30	0	0	24	98
Cote d'Ivoire 1995	48	36	22	40	1	23	14	62
Cameroon 1991	5	19	9	47	1	9	23	36
Ghana 1988	3	47	1	21	7	71	0	0
Ghana 1993	7	67	7	69	1	61	28	160
Liberia 1986	84	4	13	1	1	53	0	0
Mali 1987	33	23	1	3	78	10	0	0
Mali 1995	41	10	56	11	65	7	98	12
Nigeria 1990	21	40	7	64	69	69	0	28
Niger 1992	58	15	24	8	68	2	56	4
Senegal 1986	31	17	12	10	18	3	82	16
Senegal 1992	33	44	25	30	39	18	50	19
Togo 1988	6	1	20	39	14	34	3	36
Total	419	352	352	420	484	416	416	484

Table 5. Substantive results by cluster type, two factor procedure

Type of place of residence	Cluster Type	Factor 1 Score	Factor 2 Score	Prop women 15-49 Muslim	Prop of W. 20-49 w/birth by 20	Mean number of CEB to all women	Prop total women using a contra-ceptive method	Prop total women using modern contra-ception
Urban	1	3.2	4.4	68.6	59.7	3.4	10.2	5.6
	3	3.3	6.0	47.6	59.7	3.1	14.9	7.0
	4	4.8	5.8	22.3	46.0	2.4	25.9	12.3
	2	5.2	4.4	34.9	41.3	2.3	29.2	16.1
	Total	4.1	5.2	43.6	51.8	2.8	19.9	10.1
Rural	1	0.6	13.2	75.4	63.7	4.0	6.1	0.6
	3	0.6	14.5	72.4	66.1	3.9	6.4	0.7
	4	2.1	15.2	15.1	59.7	3.5	12.9	4.7
	2	2.4	12.8	18.7	56.9	3.4	13.9	6.1
	Total	1.4	13.9	44.5	61.6	3.7	9.8	3.0