Evaluating sanitary quality and classifying urban sectors according to environmental conditions

Craig A. Milroy, Patricia C. Borja, Fernando R. Barros and Mauricio L. Barreto

Craig A. Milroy is a postgraduate fellow at the National Security Education Programme (NSEP) and a visiting researcher at the Instituto de Saude Coletiva. He holds masters degrees in environmental engineering and epidemiology, and studies the health and environmental impacts of sanitation.

Address: e-mail: craig@ufba.br

Patricia C. Borja is an environmental engineer and holds a masters degree in urbanization. She is currently a PhD student in urbanization at the School of Architecture (Federal University of Bahia) and works as a researcher and consultant in environmental sanitation.

Address: Rua Felix Mendes, 217 Edificio Casa do Torre, Garcia, Salvador, Bahia 40100-020, Brazil; e-mail: borja@ufba.br

Fernando R. Barros is a physician and holds a masters degree in public health. He is a technical advisor at the State Secretariat of Health of Bahia and conducts epidemiological surveillance of infectious diseases at the municipal, state and national levels.

Address: e-mail: fernanbarros@bol.com.br

Mauricio L. Barreto is a professor of public health epidemiology at the Instituo

SUMMARY: This paper describes how principal components and cluster analyses were used to quantitatively score and rank sanitary conditions in 30 areas of Salvador (Northeast Brazil) and to identify groups of areas with similar environmental quality prior to a programme to improve sanitary infrastructure. In collecting data, street by street, a broad definition of sanitary quality was used, encompassing type and quality of housing, paving, water supply, sewage disposal, drainage and household waste disposal. All variables used in the principal components analysis were coded to reflect the absence of infrastructural elements that contribute to health and environmental quality. Summary variables generated by the analysis were used to score the sanitary quality of each of the areas, and cluster analysis identified four groups of areas, representing high, intermediate, poor and very poor sanitary quality. Higher rates of parasitic infection among children aged 7-14 years of age were found in areas with the worst sanitary conditions, with prevalences increasing progressively from the group with the best sanitary quality to the group with the worst sanitary quality. This kind of analysis provides a method both to appraise the needs of each community (including being able to prioritize areas most in need and identify areas with special sanitation needs) and to score baseline conditions that later allow the impacts of interventions to be assessed.

I. INTRODUCTION

ALTHOUGH THE CONTRIBUTION of sanitary conditions to public health is widely recognized,(1) there is considerable debate regarding exactly how much of a health benefit can be expected from the implementation of services such as water supply, sewage disposal and trash collection.(2) Much of this uncertainty stems from controversies regarding the selection of appropriate indicators for appraising community sanitation needs(3) and of valid methodologies for assessing impacts of sanitation interventions.(4) However, regardless of which approach is used, proper needs assessment and evaluation of the health impacts require an objective strategy to classify the hygienic conditions of the study subjects so that appropriate comparisons may be made among those inhabiting similar sanitary conditions (in order to identify risk factors for infection) or between those inhabiting different sanitary conditions (to investigate the impact of environmental conditions on health). In addition, comparisons must be made using groups that are large enough to achieve adequate statistical power. Unfortunately, aggregating study subjects

according to geographical proximity (e.g. by neighbourhood or drainage basin) may result in misclassification, since a particular area may be more similar (in terms of sanitary conditions) to distant areas than to neighbouring areas.

Ideally, to conduct proper "needs assessment" and "impact evaluation", it is desirable to objectively quantify sanitary conditions with a score or index, both at baseline (in order to benchmark initial conditions or to prioritize areas for services) and after implementation of services (to document any changes in infrastructure effected by the intervention). Associating the change in score with the corresponding change in disease burdens would allow one to determine whether there is an **incremental** impact of sanitation, i.e. how much sanitation infrastructure must be implemented to achieve a certain reduction in parasitism, diarrhoea, etc. However, the health impacts deemed to result from these interventions are seldom, if ever, correlated with quantitatively measured improvements in sanitary quality. In addition, assessing these impacts upon health and environmental quality is complicated by the fact that descriptors of sanitary conditions are often highly correlated (since areas lacking a particular service such as sewage disposal usually lack other types such as water supply and garbage collection) and there is no clear basis for objectively weighing the contributions of these individual services to environmental quality and disease prevention since "sanitary quality" results from the collective contribution of these services.

However, principal components analysis provides an objective means to construct (uncorrelated) quantitative summary indices from highly correlated variables;[5] these indices may then be used to score hygienic conditions and classify study subjects according to type of habitat. With this in mind, the current investigation applied principal components analysis to data describing the sanitary conditions in 30 areas in Salvador (capital of Bahia in Northeast Brazil, and the country's fourth largest city), in order to quantitatively score the environmental conditions in these areas prior to implementation of sanitation services, as well as to identify and form groups of areas with similar infrastructure and sanitary conditions. These areas will receive basic sanitation through the Bahia Azul Environmental Sanitation programme, a multi-nationally funded project which seeks to correct deficiencies in the city's water supply system, raise the level of coverage by adequate sewage disposal from 26 per cent (present coverage) to 80 per cent of the population, and implement systems for the collection, transport and disposal of solid waste. In order to evaluate the effect of these newly implemented sanitation services on child health (e.g. malnutrition, diarrhoea, parasitic infections) and other health indicators, an environmental and epidemiological evaluation was conducted at baseline and will be repeated after implementation of sanitary services.

II. METHODOLOGY

a. Field Evaluation

PRIOR TO COMMENCEMENT of the Bahia Azul sanitation intervention, 30 areas scattered throughout the city of Salvador were selected for evaluation purposes (three areas from each of the following ten drainage basins: Barra, Armação, Calafate, Tripas, Medio-Camarugipe, Cobre, Mangabeira, Lobato, Paripe and Periperi), the majority of which either

de Saude Coletiva (Federal University of Bahia). His research interests include the epidemiology of infectious diseases and malnutrition in the context of social inequality and social and environmental changes.

Address: Instituto Saude Coletiva, Rua Padre Feijo, 29, Canela, Salvador, 40110-170, brazil; e mail: mauricio@ufba.br

The authors would like to thank Dr Maureen Lahiff and Dr Sandy Cairncross for the valuable recommendations relating to the analysis and to the preparation and revision of the text.

1. Huttly, S R A (1990), "The impact of inadequate sanitary conditions on health in developing countries", *World Health Statistics Quarterly* Vol 43, No 3, pages 118-126; also Esrey, S A and J P Habicht (1986), "Epidemiologic evidence for health benefits from improved water and sanitation in developing countries", *Epidemiologic Reviews* Vol 8, pages 117-128; Esrey, S A and S E Burger (1995), "Water and sanitation: health and nutrition benefits to children" in Pinstrup-Anderson, P (editor), *Child Growth and Nutrition in Developing Countries: Priorities for Action*, Cornell University Press, Ithaca, NY; Feachem, R G (1983), "Infections related to water and excreta: the health dimensions of the decade" in Institute of Water Engineers and Scientists, *Water Supply and Sanitation in Developing Countries*, London, pages 25-46; and Esrey, S A, R G Feachem and J M Hughes (1985), "Interventions for the control of diarrhoeal diseases among young children: improving water supplies and excreta disposal facilities", *WHO Bulletin* Vol 63, No 4, pages 757-772.

2. Esrey, S A (1996), "Water, waste, and well-being: a multi-country study",

lacked or possessed only rudimentary forms of sanitation, along with two basins (Barra, Armação) where sanitary infrastructure has existed for some time[6]. These "sentinel areas" had been delineated previously so as to contain 200-300 children under the age of three belonging to the same socioeconomic stratum.[7] A field team comprised of architecture and civil engineering students conducted a detailed evaluation of infrastructure within these areas, using a questionnaire that previously had been tested, standardized and validated in other investigations of environmental quality conducted in Salvador.[8] All streets within the study areas were evaluated in 50-100 metre sections and detailed information was collected on topographical and ecological characteristics, land use, and type and condition of basic infrastructure (including pavement, water supply, sewage disposal, trash collection, drainage and housing, as well as the types of building materials encountered and the physical condition of all infrastructure).[9] Prior to the field evaluation, the team underwent extensive training and a trial evaluation was conducted in an area of Salvador not included in the Bahia Azul study, in order to standardize questionnaire administration and to minimize the possibility for observational errors. Since the objective of this project was to benchmark and classify the external environments of the areas involved in the study, household-level information (e.g. household management of residual solids, water usage, personal hygiene) was not included in the analysis.

b. Definition of Study Variables

In order to calculate sanitation scores and form groups of areas with similar environmental characteristics, it was first necessary to identify categories of infrastructure that contribute to sanitation and to disease prevention. Although the term "sanitation" normally refers to fecal disposal, here the authors use the term broadly to refer to the various categories of urban infrastructure which operate collectively to promote sanitary quality and health (as justified below). As such, within each of the categories that were selected (specifically: habitation, pavement, water supply, sewage disposal, drainage and residual solids management), variables were created to reflect the specific contribution of the infrastructure to sanitation (i.e. whether the street section in question lacked a particular sanitary characteristic deemed to have an impact upon sanitation and disease transmission). For example, the city contains a wide variety of **pavement** types (e.g. asphalt, brick, concrete block); however, since pavement impacts sanitation and health through its ability to prevent human contact with contaminated earth (and as a "protective cover" for soil),[10] the most important determinant of sanitation is whether or not pavement is present. Thus, the various types of pavement should all, in theory, be equally effective in preventing contact between sewage, soil and children. Accordingly, the *pavement* variable used in the principal components analysis was constructed by determining the total percentage of street sections within each area that lacked adequate coverage by man-made paving material. Similarly, the other variables were coded in order to reflect the absence of a particular sanitary characteristic (or the presence of a characteristic indicating poor sanitation).

Table 1 presents the infrastructure categories selected for the analysis and summarizes the contributions to sanitation and health that the variables were coded to reflect. All variables represent the percentage of street sections within an area that lack a particular "sanitary characteristic",

American Journal of Epidemiology Vol 143, No 6, pages 608-622; also Vanderslice, J and J Briscoe (1995), "Environmental interventions in developing countries: interactions and their implications", *American Journal of Epidemiology* Vol 141, No 2, pages 135-144.

3. Kolsky, P J and W J Blumenthal (1995), "Environmental health indicators and sanitation related disease in developing countries: limitations to the use of routine data sources", *World Health Statistics Quarterly* Vol 48, pages 132-138; also Heller, L (1999), "Who really benefits from environmental sanitation services in the cities? An intra-urban analysis in Betim, Brazil", *Environment and Urbanization* Vol 11, No 1, April, pages 133-144.

4. Blum, D and R G Feachem (1983), "Measuring the impacts of water supply and sanitation investments on diarrhoeal diseases: problems of methodology", *International Journal of Epidemiology* Vol 12, No 3, pages 357-365; also Cairncross, A M (1990), "Health impacts in developing countries: new evidence and new prospects", *Journal of the Institution of Water and Environmental Management* Vol 4, No 6, December, pages 571-577; Baltazar, J, J Briscoe, V Mesola, C Moe, F Solon, J Vanderslice and B Young (1988), "Can the case-control method be used to assess the impact of water supply and sanitation on diarrhoea? A study in the Philippines", *WHO Bulletin* Vol 65, No 5, pages 627-635; Briscoe, J, J Baltazar and B Young (1988), "Case-control studies of the effect of environmental sanitation on diarrhoea morbidity: methodological implications of field studies in Africa and Asia", *International Journal of Epidemiology* Vol 17, No 2, pages 441-447; and Briscoe,

Table 1: Summary of infrastructure categories and variables included in the principal components analyses

CATEGORY	FACTOR IMPACTING SANITATION	CORRESPONDING VARIABLE(s)	HOW IMPACT IS MADE
HABITATION (housing type)	(1) Absence of "protected" living environment	*housing*	"Protected" domiciles minimize contact with external environment and thereby reduce exposure to risk
	(2) Absence of external finish	*construc*	Indication of socioeconomic and crowding differences
PAVEMENT	Absence of paved surfaces	*pavement*	Paved surfaces prevent contact with soil/geohelminths
WATER SUPPLY	(1) Absence of public water system	*supply*	Well-maintained systems reduce contact between water and excreta/sewage, provide larger volumes of water
	(2) Presence of factors predisposing contamination (e.g. discontinuous flow, openings or leaks in system)	*continuous, contam, contam2*	Factors such as discontinuous flow and openings/leaks in system predispose contamination and impede hygienic practices
SEWAGE DISPOSAL	Absence of "closed" sewage system	*inadequate, sewage, repair*	"Openings" in system allow sewage to escape, resulting in contamination of public and domestic environments
DRAINAGE	(1) Absence of drainage system	*drainage*	Creates propitious conditions for geohelminth eggs (i.e. moist, humid soil) and mosquitoes
	(2) Existence of flooding problems	*flood*	Flooding brings excreta contaminated wastes into houses, keeps soil moist for eggs
RESIDUAL SOLIDS	(1) Absence of regular solid waste collection	*collection, irreg, trash*	Collection deficiencies lead to the accumulation of solid wastes, which attracts vectors, breeds bacteria and increases exposure
	(2) Presence of stationary solid waste collection points	*dumpster*	Sites where solid wastes are dumped create local accumulation of wastes

therefore, higher values represent worse sanitary conditions.

Habitation (housing type) contributes to sanitation and health by determining the population density of an area as well as the amount of contact that children have with the external environment:[11] dense crowding promotes transmission of disease; in addition, children who live in "protected" environments (e.g. apartment buildings, condominiums or affluent homes) tend to remain relatively isolated from their external environment, while children who live in *bairros populares* (lower-class neighbourhoods) or *favelas* (slums) have more contact with the external environment and with other children in the neighbourhood. Therefore, the variables describing habitation type (*housing and construc*) were defined to reflect whether "protected" (i.e. affluent, less crowded) or "vulnerable" (i.e. densely populated *bairro popular*, *favela*) housing was present. Treated together, these two variables also form a "proxy" for socioeconomic status, since apartment buildings or homes with external finishing represent higher socioeconomic levels whereas dwellings in areas of lower socioeconomic levels often lack external finish. In addition, it was necessary to include variables describing habitation since some of the basic sanitation interventions in Salvador also include construction of

J, R G Feachem and M M Rahaman (1985), "Measuring impact of water supply and sanitation facilities: prospects for the case-control method", WHO document WHO/CWS/85.3, Geneva; see also reference 3.

5. Selvin, S (1995), "Principal components analysis" in *Practical Biostatistical Methods*, Wadsworth Publishing Co, Belmont CA, pages 221-245; also Kleinbaum, D G, L L Kupper and K E Muller (1988), "Factor Analysis" in *Applied Regression Analysis and Other Multivariate Methods* (2nd edition), Duxbury Press, North Scituate MA, pages 595-640;

low-cost housing.

Sewage disposal is important for sanitation and health because its function is to isolate waste water from human contact.[12] Therefore, variables (*inadequate, repair, sewage*) were coded to reflect the existence of "inadequate" sewage solutions, e.g. those that do not effectively isolate sewage (as described in Appendix 1).

Without proper **drainage**, soil is more likely to be exposed by erosion and / or contaminated by sewage.[13] In addition, flooding will carry human wastes into the public and domestic environments, contaminating unprotected earth and facilitating human contact with filth.[14] Furthermore, water which is not properly carried away will leave areas moist and humid, providing propitious conditions for the development of mosquito and geohelminth eggs.[15] The variables *drainage* and *flood* were thus coded to reflect the absence of drainage and presence of flooding problems respectively.

Since adequate amounts of clean water are necessary to practice good personal and domestic hygiene, the quantity and quality of water supply will clearly affect health.[16] In addition, discontinuous supply renders water lines vulnerable to contamination and compels people to store water, a practice which provides breeding areas for mosquitoes and facilitates in home contamination.[17] Therefore, variables relating to **water supply** (*supply, contam, continuous, contam2, water*) were coded in order to reflect the presence of a well-maintained public water system as well as the existence of discontinuous supply and other factors which could compromise water quality.

Finally, without regular **residual solids** collection, trash will accumulate within the domestic and public environment, impacting sanitation and health by providing breeding material for bacteria (e.g. decomposing organic matter, vegetables, meat) and by attracting vectors (rats, cockroaches, flies, mosquitoes, etc.).[18] Variables describing residual solids management (*collection, regular, dumpster, irreg, trash*) were therefore coded to reflect conditions predisposing garbage accumulation, namely, absence of collection, absence of regular collection and existence of stationary collection points / dumpsters (versus regular, door-to-door collection).

Appendix 1 gives details of the criteria for classifying and coding the types of infrastructure encountered in the field evaluation.

c. Statistical Analyses

After defining and coding the variables, principal components were extracted from the associated correlation matrices using the STATA statistical software package (version 5.0).[19] Because all variables are expressed in the same units (e.g. percentage of street sections lacking a particular sanitation characteristic), standardization was not necessary. Three factors were kept in order to bring the total accounted variance to nearly 80 per cent. Factor loadings were then examined to identify heuristic interpretations of the linear combinations.

In order to investigate the potential existence of distinct groups of areas, the unique values of PC(1) – PC(3) for all 30 areas were subsequently evaluated for statistical proximity by cluster analysis (as detailed in Appendix 2).

Finally, the distributions of environmental variables and the prevalence of helminthic and protozoan infections were examined within the four groups in order to assess whether the resultant groupings could

and Johnston, R J (1978), "Principal components analysis and factor analysis" in *Multivariate Statistical Analysis in Geography: A Primer on the General Linear Model*, Longman Group Ltd, London, England, pages 127-183.

6. Barreto, M L, A Strino and M Prado (1997), *Avaliação de impacto epidemiológico do programa de saneamento ambiental da Baía de Todos os Santos (Bahia Azul)*, 2nd technical report, Instituto de Saúde Coletiva (Federal University of Bahia), Salvador (Brazil), 64 pages.

7. Barreto, M L, A Strino and M Prado (1997), *Avaliação de impacto epidemiológico do programa de saneamento ambiental da Baía de Todos os Santos (Bahia Azul)*, 1st technical report, January, Instituto de Saúde Coletiva (Federal University of Bahia), Salvador (Brazil), 60 pages.

8. Borja, P C, A T Elbachá et al. (1994), "Ações de saneamento ambiental em Canabrava (Salvador)" in *VI simpósia Luso-Brasileiro de engenharia sanitaria e ambiental, Florianópolis*, ABES; also Parés, M I and P C Borja (1995), "Plano de intervenção urbana do bairro de Ilha Amarela", MAU / UFBa (trabalho apresentado para a disciplina técnica e prática de projeto-mimeograph), Salvador; and Borja, P C (1997), "Avaliação da qualidade ambiental urbana – uma contribuição metodológica", MAU / UFBa (Master's dissertation, College of Architecture, Federal University of Bahia), Salvador.

9. Barreto, M L, A Strino and M Prado (1997), *Avaliação de impacto epidemiológico do programa de saneamento ambiental da Baía de Todos os Santos (Bahia Azul)*, 3rd technical report, September, Instituto de Saúde Coletiva (Federal

Table 2: Factor scores, eigenvalues and proportions of variance accounted for by principal components analysis of variables describing sanitary conditions

	Factor scores		
	pc1	pc2	pc3
housing	.76	-.31	-.30
construc	.795	-.37	.25
pavement	.885	.15	.06
supply	.555	-.39	.61
continuous	.79	-.35	.04
contam	.776	-.08	.003
inadequate	.75	.07	-.12
repair	.825	.28	-.28
drainage	.83	-.03	-.08
flood	.53	.59	-.47
regular	.30	.75	.38
dumpster	.24	.65	.49
eigenvalue:	5.88	1.95	1.23
total variance explained**	50%	16%	10%

** Summing the individual contributions of these three components shows that 76 per cent of the TOTAL variance was explained by this analysis.

contribute to the analysis of the relationship between environmental conditions and health. Prevalence ratios and chi-square tests for trend were calculated using STATCALC (Epi-Info version 6.0) in order to test observed differences in prevalence between groups.

III. RESULTS

AS DETAILED IN Appendix 1, several principal components analyses were conducted in order to evaluate different combinations of variables. Since the outcomes of these analyses were consistent, representative results from only one of the analyses are presented for the sake of simplicity. Table 2 shows the factor loadings of the first three principal components generated by this analysis as well as the eigenvalues and the proportion of total variance captured by each principal component.

Note that most of the variability in the original data is captured by these three principal components so that they may be used as efficient summaries of the original variables when analyzing data.

a. Grouping Areas using Principal Component Factor Scores

Figures 1 and 2 provide a graphical illustration of the clusters of areas that were identified and Table 3 presents a summary of the resulting groupings. Note that Figures 1 and 2 were not used to determine group

University of Bahia), Salvador (Brazil), 90 pages.

10. WHO Expert Committee (1987), "Public health significance of intestinal parasitic infections", *WHO Bulletin* Vol 65, No 5, pages 575-588.

11. Victora, C G, P G Smith, J P Vaughan, L C Nobre, C Lombardi, A M B Teixeira, S C Fuchs, L B Moreira, L P Gigante and F C Barros (1988), "Water supply, sanitation and housing in relation to the risk of infant mortality from diarrhoea", *International Journal of Epidemiology* Vol 17, No 3, pages 651-654.

12. Esrey, S A, J B Potash, L Roberts and C Shiff (1991), "Effects of improved water supply and sanitation on ascariasis, diarrhoea, dracunculiasis, hookworm infection, schistosomiasis, and trachoma", *WHO Bulletin* Vol 69, No 5, pages 609-21; also reference 1, Esrey, Feachem and Hughes (1985).

13. Kolsky, P J (1992), "Water, health, and cities: concepts and examples" in *Planning for Sustainable Urban Development – Cities and Natural Resource Systems in Developing Countries* (Paper 12), International workshop, 13-17 July, Cardiff.

14. See reference 13.

15. See reference 10.

16. Esrey, S A, H J Collett, M D Miliotis, H S Kuornhof and P Makhale (1989), "The risk of infection from giardia lamblia due to drinking water supply, use of water, and latrines among preschool children in rural Lesotho", *International Journal of Epidemiology* Vol 18, No 1, pages 248-253.

17. Quick, R E, L V Venezel, E D Mintz, L Soleto, J Aporicio, M Gironaz L Hutwagner, K Greene, C Bopp, K Malong, D Chavez and R V Tauxe (1999),

Figure 1

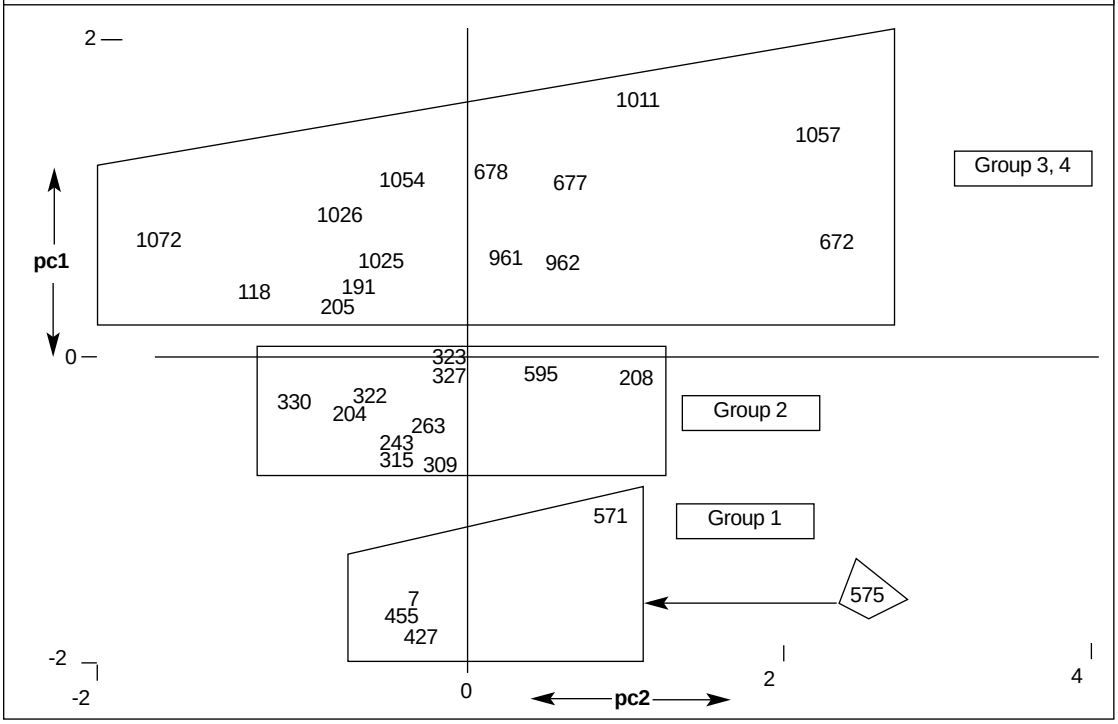


Figure 2

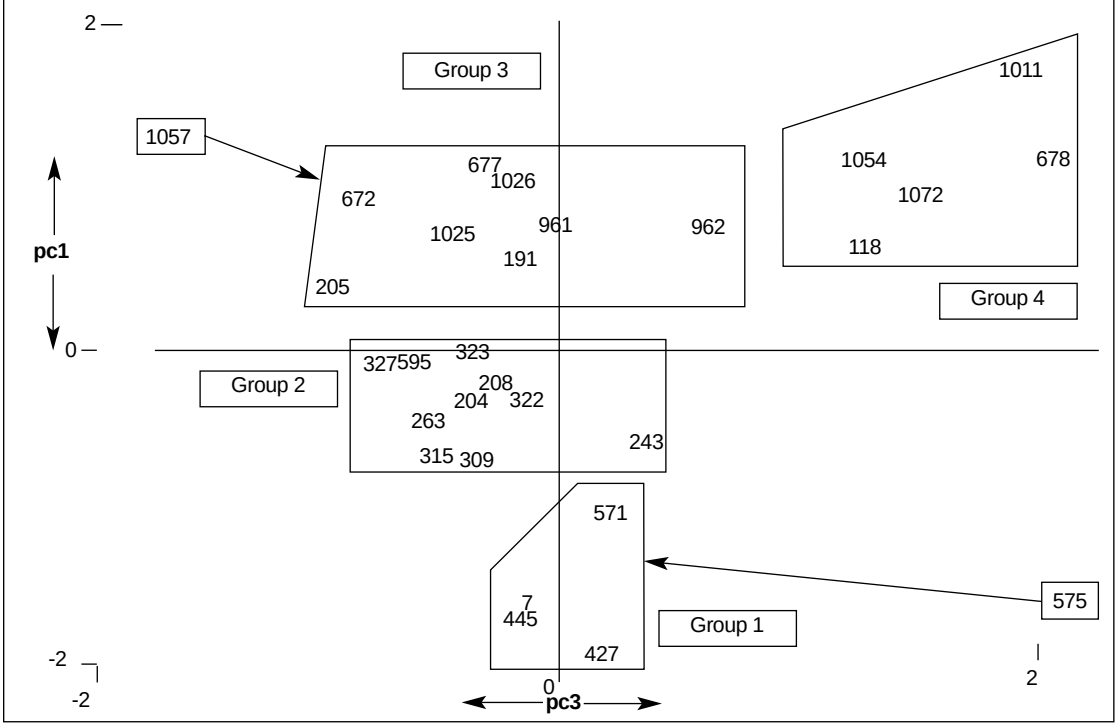


Table 3:	Groupings resulting from principal components and cluster analyses	
	AREA CODE	DRAINAGE BASIN
GROUP #1	7, 428, 445	Barra
	571, 575	Armação
GROUP #2	595	Armação
	208	Lobato
	204, 263, 323	Calafate
	243, 309, 315	Tripas
	322, 327, 330	Medio-Camarugipe
GROUP #3	205	Lobato
	191, 961, 962	Cobre
	672, 677	Mangabeira
	1025, 1026	Periperi
	1057	Paripe
GROUP #4	118	Lobato
	678	Mangabeira
	1011	Periperi
	1054, 1072	Paripe

membership but, rather, as convenient two-dimensional representations of the groupings that resulted from the cluster analysis.

As Figure 1 demonstrates, the first two principal components can be used to distinguish three distinct groups (top, middle and bottom of Figure 1); however, when three principal components are considered (Figure 2), note that the top group subsequently separates into group 3 and group 4.

b. Environmental Characteristics of the Four Groups

Table 4 summarizes (by group) the average values of the descriptive variables.

Examination of Table 4 indicates that groups 1 and 2 represent areas with high and intermediate sanitation levels (respectively) while groups 3 and 4 have low levels. Although area 575 appears as an "outlier", this area differs from the other areas in group 1 only in its method of residual solids management; area 575 has mostly dumpster collection whereas the rest of group 1 has daily door-to-door collection. Similarly, area 1057 appears as an outlier because this area has no formal drainage system while other group 3 areas are partially served.

c. Distribution of Epidemiologic Indicators

Table 5 and Figure 3 present the different prevalences of parasitic infection within the four groups, using data collected in 1997 as part of the epidemiological evaluation of children aged 7-14.[20] As expected, higher rates of parasitic infection are observed in areas with worse sanitary

"Diarrhoea prevention in Bolivia through point-of-use water treatment and safe storage: a promising new strategy", *Epidemiology and Infection*.

18. Moraes, L R S (1998), "Impacto na saúde do acondicionamento e coleta dos residuos solidos domiciliares" in *Proceedings of XXVI Congresso Inter-Americano de ingenharia sanitária y ambiental*; Lima Peru, CD, 10 pages; also Catapreta, C A A and L Heller (1999), "Associação entre coleta de residuos solidos domiciliares e saúde, Belo Horizonte (MG) Brasil", *Pan American Journal of Public Health* Vol 5, No 2, pages 88-96.

19. STATA reference manual, STATA Press, College Station, TX.

20. See reference 6.

Table 4: Environmental Characteristics of the Groups

Category/description	Variable	AVERAGE VALUES			
HABITATION:		**GROUP 1**	**GROUP 2**	**GROUP 3**	**GROUP 4**
% vias** without apartment buildings	*housing*	38	96	99	100
% vias with houses lacking external finish	*construc*	1	22	43	62
PAVEMENT:					
% vias without paved surfaces	*pavement*	2	13	62	67
WATER SUPPLY:					
% vias without public water supply	*supply*	0.6	5	8	33
% vias without continuous water supply	*continuous*	0.3	57	88	98
% vias with factors predisposing to contamination	*contam*	0.9	13	27	33
% without continuous supply OR with factors predisposing to contamination	*contam2*	0.5	58	87	85
SEWAGE DISPOSAL:					
% vias with "inadequaqe/unsatifactory" solutions for sewage disposal	*inadequate*	7.5	31	60	48
% vias with poorly maintained sewage networks	*repair*	1.5	14	41	25
% vias with sewage solutions which will not effectively isolate sewage	*sewage*	8	38	74	62
DRAINAGE:					
% vias without drainage	*drainage*	20	58	82	85
% vias with flooding problems	*flood*	0.4	12	28	8
RESIDUAL SOLIDS:					
% vias without solid waste collection	*collection*	4	52	61	59
% vias without daily waste collection	*regular*	40	54	46	44
% vias without "regular" (i.e. scheduled) collection	*irreg*	4	54	66	65
% vias with stationary collection points	*dumpster*	13	15	16	18
% vias without daily, door-to-door collection	*trash*	16	64	74	71

** "vias" are street sections

conditions, such that prevalence increases progressively from group 1 to group 4. These trends were statistically significant for all types of infection, and particularly for infection by *A. lumbricoides* and *T. trichuris* where differences in prevalence between groups are considerably greater than those for other parasitic organisms (as indicated by both prevalence ratios and chi square results). However, an exception was observed for *G. lamblia* infection, where prevalence in groups 3 and 4 was similar. For infection by *E. hystolitica*, the same trend was observed ($c^2_{\text{for trend}} = 6.1$; p=.013) but the prevalence ratios for groups 2, 3 and 4 included the unit.

IV. DISCUSSION

BECAUSE PRINCIPAL COMPONENTS are linear combinations of the variables and factor loadings (presented in Table 2), an examination of factor loadings provides a basis for forming heuristic interpretations of a principal component's underlying significance:

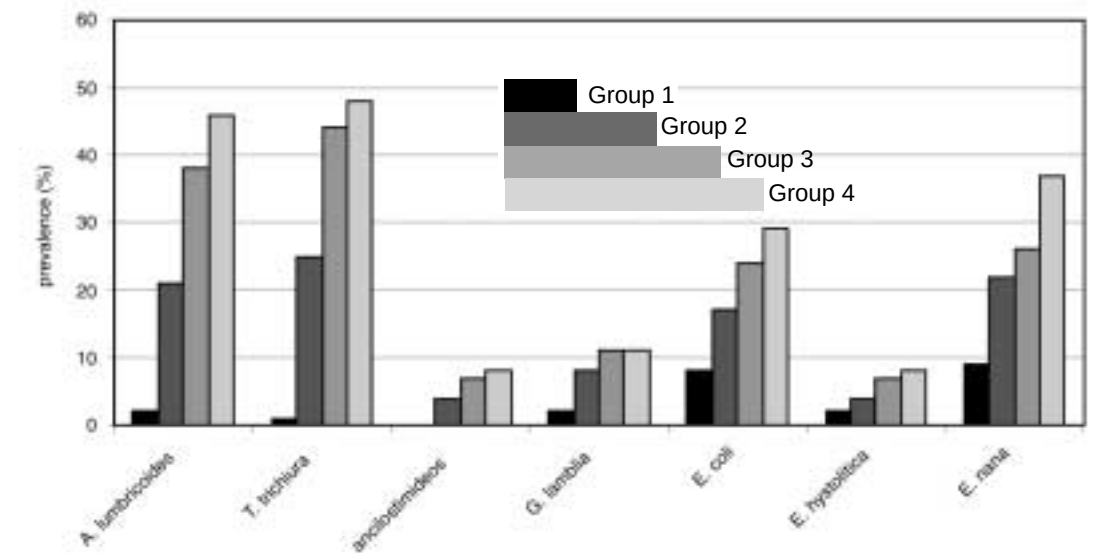
PC(1) = 0.76(*housing*) + 0.79(*construc*) + 0.89(*pavement*) + 0.55(*supply*) + 0.79(*continuous*) + 0.78(*contam*) + 0.75(*inadequate*) + 0.825(*repair*) + 0.83(*drainage*)

Table 5: Prevalence of parasitic infection calculated separately by group, overall prevalence of infection, prevalence ratio, and results of chi-square test for trend

	Group 1		Group 2			Group 3			Group 4				Overall
Parasite	No. of exams	% infected	No. of exams	% infected	prev. ratio** (Conf. Int.)	No. of exams	% infected	prev. ratio** (Conf. Int.)	No. of exams	% infected	prev. ratio** (Conf. Int.)	Chi-square (p-value)*	prevalence (%)
A. lubricoides	90	2.2	507	21	9.4 (2.4 - 37.4)	330	38	16.9 (4.3 - 67)	203	46	20.6 (5.2 - 82)	84.4 (< 0.001)	29
T. trichiura	90	1.0	507	25	22.2 (3.1 - 156)	330	44	40.1 (5.7 - 282)	203	48	43.4 (6.2 - 306)	88.6 (< 0.001)	33
ancilostimideos	90	0	507	4.0	***	330	6.7	1.7 (0.9 - 3)***	203	8.4	2.1 (1.1 - 4.0)***	11.6 (0.001)	5.0
G. lamblia	90	2.2	507	7.5	3.4 (0.8 - 13.7)	330	11	5.2 (1.3 - 21)	203	11	5.1 (1.2 - 21.1)	8.38 (0.004)	9.0
Entamoeba coli	90	7.8	507	17	2.2 (1.0 - 4.6)	330	24	3.0 (1.4 - 6.4)	203	29	3.7 (1.8 - 7.8)	22.5 (< 0.001)	20
Entamoeba hystolitica	90	2.2	507	4.3	1.9 (0.5 - 8.1)	330	6.7	3.0 (0.7 -12.5)	203	7.9	3.6 (0.8 - 15.1)	6.1 (0.013)	5.0
E. nana	90	9.0	507	22	2.5 (1.3 - 5.0)	330	26	2.9 (1.5 - 5.8)	203	37	4.2 (2.1 - 8.3)	13.8 (< 0.001)	25

* chi-square analysis for trend in proportions
** prevalence ratios calculated using group 1 as reference
*** since there was no infection in group 1, prevalence ratio was calculated using group 2 as reference

Figure 3: Prevalence of parasitic infection, by group



+ 0.53(*flood*) + 0.30(*regular*) + 0.24(*dumpster*)

Note that the magnitude of the coefficients assigned to each variable indicates the correlation of the original variable with the principal component, and the square of the coefficient indicates the proportion of the original variable's variance explained by the principal component. For example, variables describing **habitation** (*housing, construc*), **pavement** (*pavement*), frequency and contamination of **water supply** (e.g. *continuous, contam, contam2, water*), **sewage disposal** (*inadequate, repair, sewage*), and **drainage** have large factor loadings in the linear combination

comprising the first principal component. This indicates that the first principal component has a strong positive correlation with each of these variables, and that these variables will strongly influence the PC(1) "score" for each area. Therefore, areas with a higher percentage of streets which lack pavement, well-finished housing, sewage systems, drainage systems, etc. (i.e. less complete and / or poorly maintained basic infrastructure) will tend to have larger values of PC(1). On the other hand, areas with more complete and well-maintained infrastructure will have smaller PC(1) scores. As a result, the first principal component may be interpreted as a rough index of the overall quality of infrastructure, such that the most well serviced and maintained areas will display the lowest PC(1) scores. Furthermore, this "index" may be used as a means of identifying and prioritizing areas which are most in need of "constructed" sanitation services such as pavement, sewage or drainage, or which require infrastructure maintenance.

The second principal component is dominated by variables describing **residual solids management** (e.g. *regular* and *dumpster*), such that PC(2) reflects the frequency and quality of residual solids management within each area:

PC(2) = -0.31(*housing*) – 0.37(*construc*) + 0.15(*pavement*) – 0.39 (*supply*) – 0.35(*continuous*) – 0.08(*contam*) + 0.07(*inadequate*) + 0.28(*repair*) – 0.03(*drainage*) + 0.59(*flood*) + 0.75(*regular*) + 0.65(*dumpster*)

As a result, PC(2) may be used to compare the relative quality of trash removal services within each of the areas. For example, although Table 4 shows that group 1 areas are well served with sanitation as a whole, the PC(2) scores of areas 571 and 575 in the Armação basin indicate that these areas could benefit significantly from improved trash collection. In addition, the variable *flood* also loads on PC(2), such that this factor contains most of the variability related to flooding and can be used to identify frequently flooded low-lying areas in group 3 (e.g. areas 672, 677 and 1057, all of which lie to the far right of Figure 1) that are most likely to benefit from improved drainage services.

Finally, the variable *supply* loads heavily on the third principal component, such that PC(3) appears to represent the degree of coverage by public water systems:

PC(3) = -0.30(*housing*) + 0.25(*construc*) + 0.06(*pavement*) + 0.61(*supply*) + 0.04(*continuous*) + 0.003(*contam*) – 0.12(*inadequate*) – 0.28(*repair*) – 0.08(*drainage*) – 0.47(*flood*) + 0.38(*regular*) + 0.49(*dumpster*)

However, PC(3) can also be interpreted as a "proxy" describing the level of development (since water systems are often the first public service to be implanted in a newly settled area) or as an abstract representation of whether the area is inaccessible to vehicles or built on a slope, for the following reasons:

- the variable *flood* loads strongly on PC(3): areas on slopes are less likely to flood;
- the variable *dumpster* also loads rather substantially upon PC(3): in areas that are difficult to access, residents are normally required to carry their trash to points which are accessible to collectors.

Therefore, in addition to identifying recently inhabited areas in need of water supply infrastructure, PC(3) can also be used to identify areas whose difficult access will require special implementation strategies that are likely to require special designs or incur higher construction costs (e.g. group 4 areas).

Since PC(2) and PC(3) contain both positive and negative factor

weights, they are more difficult to use as indices for ranking because there is no absolute interpretation for higher scores for these two PCs. These factors are thus better interpreted as indicating "tendencies" (e.g. since the variable *supply* loads heavily on PC(3), areas with higher PC(3) scores will tend to have less coverage by public water supplies).

a. Evaluation of Group Differences

Since sanitation infrastructure is normally implemented at the community level, it is useful to identify areas with similar needs in order to design appropriate sanitation solutions. Cluster analysis provides a useful way of identifying groups of areas that could benefit from similar sanitation measures. Evaluating the average values of the environmental indicator variables (for each group) helps to identify essential differences between the groups as well as illustrate the special sanitation needs of each group. Table 4 provides insight into the group differences in environmental characteristics.

Group 1 is comprised of the areas of Barra and Armação (two affluent neighbourhoods in Salvador). An average of 62 per cent of the street sections within these areas have apartment buildings and only 1 per cent of the streets surveyed contain dwellings without external finish. As shown by the values for *pavement, supply, inadequate* and *collection*, an average of less than 10 per cent of the street sections in this group lack basic sanitation / infrastructure and less than 1 per cent of the streets lack continuous water supply. Furthermore, less than 10 per cent of the street sections demonstrate problems with infrastructure maintenance, as evidenced by the values for *contam, repair* and *flood*. Although the comparably high average for *drainage* (20 per cent) indicates a relative lack of drainage, this is due to the influence of area 575, which lacks drainage in nearly 70 per cent of its street sections (whereas 90 per cent of the streets in the other areas in group 1 have drainage). In addition, the apparently high value for *regular* is entirely due to the areas in the Armação basin, which have collection on alternate days while areas in Barra have daily trash collection in over 90 per cent of the street sections surveyed. Therefore, group 1 represents the areas with the most complete and well-maintained sanitary infrastructure in Salvador. It is worth reiterating, however, that the quality of sanitation in Barra is more consistent and well established than that of Armação.

Group 2 consists of areas in the basins of Calafate, Tripas, Medio-Camarugipe (middle and lower-class neighbourhoods) as well as two individual areas, namely, area 595 of Armação and area 208 of Lobato. The smaller average values for the *housing, construc* and *pavement* variables indicate that habitation consists of well-finished houses (i.e. with external finish) on paved streets. Overall, this group of 11 areas is characterized by intermediate levels of sanitation: the values for *inadequate, drainage* and *collection* indicate that an average of approximately half of the streets in this group have adequate solutions for sewage, residual solids management and drainage. In addition, although the areas comprising this group are well-served by public water supply (average *supply* = 5 per cent), this supply appears to be continuous in an average of only half of the street sections (average *continuous* = 57 per cent), such that improvement of water service within the areas of group 2 should be prioritized. Furthermore, examination of the *contam, repair* and *regular* variables reveals that approximately 15 per cent of the existing infrastructure within this group

requires maintenance and that trash collection within this group appears to be lax. Although group 2 represents intermediate sanitary infrastructure quality, the relative "positions" of these areas in Figure 1 show that this group also displays the widest variation in conditions. For example, area 243 has sanitation that is almost equal to the quality of that in the group 1 areas, while areas in the Medio-Camarugipe basin are characterized by sanitation that is almost as bad as groups 3 and 4.

Finally, groups 3 and 4 are comprised of areas in the basins of Mangabeira, Lobato, Periperi and Paripe (which represent the most impoverished areas of Salvador). These two groups are characterized by very precarious sanitary conditions that are significantly worse than groups 1 and 2. As indicated by the very large average values for *pavement, supply, inadequate, drainage* and *collection*, a majority of the street sections in these groups lack basic infrastructure. However, group 3 appears to have greater sewage disposal deficiencies while group 4 appears to have a greater need for water supply. Furthermore, the values for *contam, repair, flood* and *regular* demonstrate that an average of nearly one-third to one-half of the infrastructure that exists within these areas is poorly maintained. As the values for *housing, construc* and *pavement* indicate, habitation within these groups consists primarily of poorly finished (or even ramshackle) houses on unpaved roads. In fact, the neighbourhoods in groups 3 and 4 are primarily slums and are thus much more densely populated and randomly constructed than the neighbourhoods of groups 1 and 2. The provision of sewage disposal in such areas is especially challenging due to the tight spacing of dwellings and must be considered case by case, usually with a slightly different approach for each individual *favela*. Because the average values for *housing, pavement, continuous, contam, drainage, collection, regular* and *dumpster* are very similar for these two groups, the areas of these two groups have similar values for PC(1) and PC(2), and are differentiated primarily by the values for PC(3). According to the heuristic interpretation of PC(3) provided above, areas in group 4 (which have higher PC(3) values), in addition to having a higher percentage of street sections without public water supply, are also more recently inhabited or located in more precarious locations such as slopes / hillsides, and those of group 3 in lower areas. This implies that although both groups are badly in need of basic sanitation infrastructure, implementing the required sanitary measures in group 4 areas may be even more logistically difficult and costly.

Table 3 demonstrates that some basins (e.g. Armação, Lobato, Mangabeira, Paripe, Periperi) contain areas that are more similar (in terms of sanitary conditions) to areas in other drainage basins than to the areas within their own basin. Grouping children by drainage basin would therefore mis-classify the exposure status of children in these areas, since their sanitary conditions are significantly different from those for children in the other areas of these basins.

The relatively large differences in prevalence of parasitic infection between groups demonstrates that subtle differences in sanitary quality can have significant impacts on epidemiological parameters. However, this cross-sectional analysis cannot conclusively determine whether sanitation is the only factor responsible for the observed differences in the levels of parasitic infection nor whether specific deficiencies in sanitary services are likely to be responsible. Therefore, a second field survey will be conducted upon completion of the intervention in order to generate a

second score for each area that will reflect how much the sanitary conditions were altered over the course of the intervention. By associating score changes that result from sanitation implementation with changes in disease prevalence which occur as the result of the intervention, one may estimate how much sanitation infrastructure is required to attain a particular reduction in parasitic infection rates. In addition, since different areas will receive different types and degrees of sanitation from the intervention, it may be possible to evaluate the relative benefits of different types of sanitation infrastructure.

V. CONCLUSIONS

THIS PAPER DESCRIBED how principal components and cluster analyses were used to quantitatively score and rank sanitary conditions in 30 areas of Salvador, prior to implementation of sanitary infrastructure, and to identify groups of areas with similar environmental quality. The analysis included information on both presence **and** quality of infrastructure. Principal components analysis was chosen to address the following objectives of the study.

Quantitative scoring of sanitation. Since principal components analysis calculates summary "scores" from input data (which, in this case, contain information about sanitary infrastructure), these scores may be used as a quantitative representation of the sanitary conditions in an area. It is therefore possible to rank areas in terms of sanitary quality by comparing their scores, as well as to prioritize areas which are most in need of sanitation and to identify areas with special sanitation needs. In addition, by comparing the scores before and after the intervention, one may quantitatively evaluate the changes in sanitary conditions which occurred in an area during the intervention.

Classification of areas according to sanitary quality. Areas that are similar with respect to the variables included in the analysis will have similar principal component scores, and may therefore be grouped. Grouping areas with similar sanitary characteristics increases the precision and power of statistical analyses (since areas may be compared in groups rather than one-to-one). In addition, this approach is also a useful tool to define / design strategies for epidemiological surveillance of population groups; furthermore, subjects may be classified by type of habitat, which reduces the probability of mis-classifying the "sanitary status" of their neighbourhood.

Prediction of epidemiological impact. Presenting the prevalence of "sanitation-related" diseases according to level of sanitary quality (rather than by geographic proximity) provides a more accurate portrayal of the epidemiological profile of infection. Furthermore, one may estimate the health impacts associated with the implementation of sanitation infrastructure by comparing the changes in the prevalence of infection that occur during and after implementation of sanitary services with the changes in sanitary score that result from the intervention.

The strong association between group membership and disease prevalence indicates the validity of the procedure used to define variables, calculate scores of sanitary quality, and identify groups, and demonstrates that it is a useful method for determining priority areas for sanitation intervention. Note that all variables used in the analysis were coded according to a specific strategy designed to document the absence of

particular infrastructure characteristics which contribute to sanitary quality and disease prevention. Although this analysis was applied in order to quantify sanitary conditions, principal components analysis is a general technique which may be used to appraise or classify urban areas according to almost any criteria (i.e. economic, ecological, socioeconomic, population type, etc.), identify or prioritize areas in need of other types of development services or qualitatively evaluate the impact of interventions. However, the success of developing "scores" (and the subsequent grouping of similar areas) depends on the creation of an appropriate conceptual framework to organize the information that is to be analyzed and on the definition of indicators whose variability will be sufficient to distinguish groups of areas. As such, the objectives of the analysis must be carefully considered at the outset and rational criteria for the inclusion of variables should be established prior to analysis.

Appendix 1: Description of criteria for coding the variables used in principal components analyses

Note that some variables were defined by combining the information from two other variables. For example, although the capacity of a system to isolate sewage was described by the following two variables:

- inadequate, which describes the specific type of sewage disposal and indicates the presence of sewage systems that are incapable of effectively isolating sewage (because of inappropriate design); and
- repair, which indicates whether the sewage system in question has maintenance problems which would allow sewage to escape (i.e. presence of openings, clogs, etc.), regardless of whether the system is appropriately designed,

these two variables may also be merged in order to create one variable (sewage) which summarizes the capacity of the sewer system to isolate sewage. When the sewage disposal information enters the analysis as one "unified" variable (*sewage*), one may identify the total number of street sections within a particular area which contain sewage solutions that do not effectively isolate excrement, but one may not distinguish whether this inability results from the design of the system or from maintenance problems. On the other hand, when the information about sewage disposal enters the analysis as two separate variables (e.g. as *inadequate* and *repair*), one may differentiate areas with a high percentage of "inadequate" sewage solutions from those with a high proportion of sewage system maintenance problems; however, one may not distinguish the total number of streets with sewage solutions which effectively isolate waste water. As such, there is some trade-off involved in defining and choosing variables for the principal components analysis. Clearly, some experimentation is warranted in order to ensure robust results. However, information must never be duplicated in two variables, since this would bias the correlation matrix. Therefore, when the information from two variables was unified, double counting of observations was avoided.

This strategy of "variable combination" (i.e. combining information from two variables in one unified variable) was also applied in the **water** and **residual solids** categories. In each case, the effect of this strategy was evaluated by conducting one principal components analysis using the two variables and a separate analysis using the "unified" variable, and comparing the resulting differences in principal component factor loadings and area groupings. Although "variable combination" had a minor influence on component factor loading, the outcome of the grouping process was consistent.

NOTE: the term "via" refers to street sections of 50-100 metres which were chosen as the unit of analysis.

continued next page

Appendix 1: continued

CATEGORY	*VARIABLE name*	VARIABLE DEFINITION	Corresponding infrastructure types (criteria for coding)	Portuguese term(s)
HABITATION	*housing*	% vias without private apartment buildings	Any road WITHOUT either of the following housing types: (1) low-rise apartment building (less than four floors) (2) high-rise apartment building (more than four floors)	RUAS SEM: vertical de baixo gabarito vertical de alto gabarito
	construc	% vias with houses which lack external finish	presence of houses built of: (1) masonry without external cover (2) mud bricks and wood (3) wood (4) other types of building materials OR: streets in which some houses have external covering and some do not	alvenaria sem revestimento taipa madeira outro alvenaria com & sem revestimento
PAVED SURFACES	*pavement*	% vias without paved surfaces	dirt roads without surfacing OR: dirt roads that are partially paved with: (1) asphalt (2) paving stone/brick (3) concrete slab (4) concrete block (5) gravel (6) mortar (walkway/drainage stairway)	terra batida OU parcialmente pavimentado com: (1) asfalto (2) paralelepípedo (3) placa de concreto (4) bloquetes (5) cascalho (6) argamassa armada
WATER SUPPLY	*supply*	% vias without public water supply	absence of piped water supply water supply via clandestine connection water supply from a well	não tem abastecimento clandestino ("gato") poço
	continuous	% without continuous water supply	receive water once per day receive water 3 - 4 times per week receive water 3 - 4 times per month	todo dia 3 - 4 vezes/semana 3 - 4 vezes/mes
	contam	% vias with factors allowing contamination	any one of the following problems: (1) leaks in water supply lines (2) exposed supply lines (vulnerable) (3) destroyed (4) supply pipe in contact with sewage (5) hydrometer in contact with sewage OR: any combination of the above	vazamento rede aflorando no pavimento rede destruída rede em contato com esgoto hidrômetro em contato com esgoto
	contam2	combines information contained in the variables continuous and contam		
	water	combines information contained in the supply and contam variables		

Appendix 1: continued

SEWAGE DISPOSAL	*inadequate*	% vias with "inadequate" and/or "unsatisfactory" sewage disposal	deposit feces in bag and throw away	"balão"
			sewage discharged directly to street	a céu aberto na rua
			disposal to open-air canal or stream	a céu aberto (canal ou riacho)
			box latrine with discharge to street	fossa com disposição a céu aberto
			box latrine with discharge to drainage	fossa c/ disposição a rede drenagem
			sewage disposal via drainage system	rede de drenagem
			disposal to drainage under public stairs	escadaria drenante
			presence of adequate and inadequate	soluções adequadas e inadequadas
			OR: any combination of the above	
			OR: any of the above solutions on street	
			where Bahia Azul system was under	
			construction at the time of the survey	
	repair	% vias sewage systems with maintenance problems	ANY of the following problems:	presença de vazamento
			(1) leaks/openings in pipes	rede aflorando na rua/pavimento
			(2) exposed pipes (vulnerable location)	rede destruída
			(3) destroyed	rede parcialmente obstruída
			(4) partially obstructed	obstruída
			(5) obstructed	outro
			(6) other problem	mais de um problema
			OR: more than one of above problems	
DRAINAGE	*drainage*	% vias without drainage	absence of the following drainage types:	NÃO TEM:
			(1) gutter	canaleta
			(2) open channel/canal/viaduct	canal
			(3) drainage covered by public stairway	escadaria/rampa drenante
			(4) public stairway with drain at the foot	escadaria com deno
			(5) underground drainage pipes	galeria
	flood	% vias with flooding problems	rain causes flooding to occur in:	QUANDO CHOVE:
			street and some of the houses	alaga toda a rua e parte das casas
			street and all of the houses	alaga toda a rua e todas as casas
			part of the street and some houses	alaga parte da rua e das casas
			all of the houses	alaga todas as casas
			street only	alaga a rua

Appendix 1:	continued			
RESIDUAL SOLIDS	*collection*	% vias without solid waste collection	no formal solid waste collection	sem coleta
	regular	% vias without daily solid waste collection	collection occurs: (1) once per week (2) once per month (3) sporadic	 uma vez por semana uma vez por mês esporática
	irreg	combines information contained in the *collection* and *regular* variables		
	dumpster	% vias where solid waste is collected from stationary collection points	presence of communal collection point collection from stationary trash bins, door-to-door and dumpster collection	ponto de lixo caixa estacionada porta a porta e ponto de lixo
		% vias without daily, door-to-door solid waste collection	combines information from *collection*, *regular* and *dumpster* variables	

Appendix 2: Statistical Background

Principal components analysis is a statistical technique which may be used to:

1. Produce univariate summaries of multivariate data, i.e. explain/account for as much of the total variation in the original data as possible with as few principal components as possible. This is done by calculating summary variables (principal components) which have as high a correlation with the original variables as possible. As a result, information from a large number of original variables may be represented by a few summary variables (a process know as "variable reduction").
2. Discover underlying dimensions within complex, highly intercorrelated data (i.e. identify groups of variables which represent sophisticated relationships within the data).

Principal components are constructed by forming unique, weighted linear combinations of the original variables:

$PC(i) = a_1(\text{variable\#1}) + a_2(\text{variable\#2}) + a_3(\text{variable\#3}) + ...+ a_n(\text{variable\#n})$

Although it is theoretically possible to form as many principal components (i.e. linear combinations) as there are original variables, most of the variability of the original variables can usually be captured by the first few principal components so that the original data may be effectively represented by two or three summary variables (principal components). For an analysis incorporating n different variables, the linear combinations which define the first three principal components are described in EQUATIONS(1) – (3):

EQUATION(1): $PC(1) = a_1(\text{variable \#1}) + a_2(\text{variable \#2}) + a_3(\text{variable \#3}) + ...+ a_n(\text{variable \#n})$
EQUATION(2): $PC(2) = b_1(\text{variable \#1}) + b_2(\text{variable \#2}) + b_3(\text{variable \#3}) + ...+ b_n(\text{variable \#n})$
EQUATION(3): $PC(3) = c_1(\text{variable \#1}) + c_2(\text{variable \#2}) + c_3(\text{variable \#3}) + ...+ c_n(\text{variable \#n})$

where a_1, a_2, a_3 ... a_n represent the coefficients (known as "factor loadings") for the first principal component, b_1, b_2, b_3...b_n represent the coefficients associated with the second principal component, etc. Note that all three principal components are constructed with the same variables but that each principal component assigns unique coefficients to each variable. The coefficients assigned to the n variables are

Appendix 2: **continued**

calculated in order to satisfy the following criteria:

1. The first principal component must have the greatest variance (i.e. the values of first principal component will have the greatest variability), the second principal component must have the second largest variance, the third principal component must have the third largest variance, etc.
2. All principal components must be completely uncorrelated.

The process of principal components analysis can be summarized as follows:

[correlation matrix] [weight matrix] = [factor loading matrix]

This method (called the factor-analytic method) determines a weight matrix (W) that is applied to the correlation matrix (R) to obtain a factor-loading matrix (L). This method of principal components analysis first requires the calculation of the TOTAL VARIANCE in the data (i.e. the sum of the variance of the original variables X_1, X_2, X_3, etc):

TOTAL VARIANCE = $S_1^2 + S_2^2 + S_3^2 + ... + S_n^2$ (where S_1^2 is the variance of X_1, etc.)

The first principal component is the weighted linear combination of the original variables which is found to account for the largest amount of the total variability (i.e. has the highest correlation with as many of the original variables as possible); that is, PC(1) is the linear combination of the original variables:

$PC(1) = a_1X_1 + a_2X_2 + a_3X_3 + ... + a_nX_n$, where weights are chosen in order to maximize the quality:

$$\frac{\text{variance of PC(1)}}{\text{total variance}}$$

Therefore, no other linear combination of the Xs will have as large a variance as PC(1). When the Xs are in standardized form (i.e. variance = 1), the total variation accounted by PC(1) is:

$$\frac{\text{variance of PC(1)}}{n}$$ (where n is the number of original variables in the analysis)

Similarly, the second principal component is the weighted linear combination of the variables that is uncorrelated with PC(1) which accounts for the maximum amount of the remaining total variation not already accounted for by PC(1), i.e. the linear combination:

$PC(2) = b_1X_1 + b_2X_2 + b_3X_3 + ... + b_nX_n$

that has the largest variance of all linear combinations **which are uncorrelated** with PC(1). This ensures that PC(1) and PC(2) are orthogonal. In general, the principal component "i" is the linear combination:

$PC(i) = w_{i1}X_1 + w_{i2}X_2 + ... + w_{in}X_n$

that has the largest variance of all linear combinations which are uncorrelated with all of the previously determined i – 1 principal components. In order to satisfy the criteria of maximum variability and zero correlation, the coefficients for the first principal component are calculated using the following system of equations:

$a_1(S_1^2) + a_2(S_{21}) + a_3(S_{31}) + ...+ a_n(S_{n1}) = a_1(L_1)$
$b_1(S_{12}) + b_2(S_2^2) + b_3(S_{32}) + ...+ b_n(S_{n2}) = b_2(L_1)$

$c_1(S_{1n}) + c_2(S_{2n}) + c_3(S_{3n}) + ...+ c_n(S_n^2) = c_n(L_1)$

where L_1 represents the variance of principal component #1, S_{21} represents the covariance of X_1 and

Appendix 2: continued

X_2, etc. To solve this set of equations, an additional constraint must be imposed, namely that:

$a_1 + a_2 + a_3 + ... + a_n = 1$ (i.e. the weights are chosen subject to the restriction that sum $a^2 = 1$)

so that the variability of PC(1) will not exceed the total variability. The coefficients for principal component number 2, number 3, etc. are generated in the same manner.

Groups of variables which are highly intercorrelated in the original correlation matrix will tend to have high factor loadings on the same principal component; this allows the identification and interpretation of "variable groups" and also provides justification for using principal components to "replace" these groups of highly intercorrelated variables in subsequent analyses (i.e. as a "proxy" for the variable group).

DESCRIPTION OF TERMS

PRINCIPAL COMPONENT/FACTOR: a weighted linear combination of the original variables (see equations 1-3 in the section above).

FACTOR LOADING: the coefficients assigned to the variables comprising the factor. These loadings indicate the correlation between the principal component/factor and the original variable. The square of the loading indicates the proportion of the original variable's variance explained by the principal component.

FACTOR WEIGHT: not a correlation but, rather, a weight (usually standardized into Z score form) assigned to each factor used in determining factor scores. Factor weights are usually different from factor loadings although high factor loadings tend to correspond to high factor weights. Therefore, factor weights and factor loadings give similar information but are measured on different scales and used for different purposes: weights are used to compute factor scores and loadings are used to describe correlations. When the variables are standardized or measured in the same units, variables with higher factor loadings and factor weights will have more influence on the overall factor score.

FACTOR SCORE: the specific value of a factor for a particular sampling unit/observation; calculated by substituting the values of the original variables into the factor expression and multiplying by the appropriate coefficient. Since there are 30 micro-areas (i.e. 30 observations) in this analysis, there will be 30 different values for each principal component. The scores for any one component are scaled to Z-score form so that observations with a positive score are above average with respect to that component, while a micro-area with a negative score is below average with regard to that component/factor .

GROUPING AREAS USING PRINCIPAL COMPONENT FACTOR SCORES

Since principal components analyses are constructed by synthesizing information contained in many variables, areas with similar values for two or more principal components can be considered "statistically similar" and will share many common characteristics. Cluster analysis provides a formal method to assess whether statistical proximity exists between two or more observations. In this analysis, cluster analysis was conducted in the following manner:

1. The "statistical locations" of areas were designated by values of PC(1), PC(2), PC(3).
2. The Mahalanobis statistical distances (i.e. Pythagorean distances in 3-D) between all areas were calculated using the Pythagorean theorem in three dimensions (e.g. the distance between areas A and B would be:

$$\text{distance} = \text{square root}\{[PC(1)_A - PC(1)_B]^2 + [PC(2)_A - PC(2)_B]^2 + [PC(3)_A - PC(3)_B]^2\}$$

3. For each area, the Mahalanobis distances to all other areas were ranked, assigning rank = 1 to the closest area, etc., in order to identify proximal areas.
4. "Clusters" of areas were designated by identifying groups of areas which all are mutual nearest-neighbours of each other.

Appendix 2: continued

5. The "centroids" (i.e. average coordinates) of each cluster were located by calculating the average coordinants (i.e. values for PC(1), PC(2) and PC(3)) of the areas comprising the cluster.
6. The distances from each area to all centroids were calculated, again using the Pythagorean theorem in three dimensions.
7. All areas which were located close to the same centroid (and far from other centroids) were considered to belong to the same group.

This process was repeated for each of the analyses conducted, in order to ensure that the groupings were not merely due to chance (i.e. resulting from a lucky choice of variables); the final groupings thus reflect the collective consideration of all principal component analyses conducted.

REFERENCES

Selvin, S (1995), "Principal components analysis" in *Practical Biostatistical Methods*, Wadsworth Publishing Co., Belmont CA, pages 221-245.

Kleinbaum, D G, L L Kupper and K E Muller (1988), "Factor analysis" in *Applied Regression Analysis and Other Multivariate Methods* (2nd edition), Duxbury Press, North Scituate MA, pages 595-640.

Johnston, R J (1978), "Principal components analysis and factor analysis" in *Multivariate Statistical Analysis in Geography: A Primer on the General Linear Model*, Longman Group Ltd., London, UK, pages 127-183.