Designing Surveys of International Migrants: The Needle in the Haystack, or Finding the Right Haystack? A Case Study of Colombians in Ecuador

Richard E. Bilsborrow, Carolina Population Center, University of North Carolina
September 22, 2006

I. The Issue

International migration is taking on increased importance in the world, with the latest (UN, 2006) estimates of the stock of persons living in countries other than those of their birth reaching 195 million, an all-time high. International migration has accelerated since 1990, even without taking into account the substantial effects of changes in borders and creation of new countries. Most of this migration has been south-north though south-south migration and movements of asylum seekers have proceeded apace as well. For the first time in history, the world's major international development financial institutions such as the World Bank, the Asia Development Bank and the Inter-American Development Bank have launched major programs and studies on the flows of international remittances, largely north-south, to determine how they can be more effectively channeled to promote economic development in developing countries. The annual flow of remittances has reached about $200 billion, far surpassing total overseas development assistance, with remittances becoming the principal source of foreign exchange earnings for a number of developing countries. Understanding the mechanisms behind both the initial migration and the remittance sending behavior has therefore become of major interest. The World Bank has launched a program to develop surveys to collect data to better understand international migration and remittances.

However, a major methodological issue is how to collect the data, since international migrants are generally a very small proportion of the population in host countries, and more so if one focuses on *recent migrants* rather than lifetime migrants. It is clearly the former that is of interest for formulating policies. The proposed paper is concerned with the testing of methodologies developed for designing samples to select international migrants, based on the "rare elements" problem in the sampling literature, and the results of an empirical study in one developing country, which also incorporated the use of snowball sampling to attempt to locate the international migrants. Several significant problems were encountered in the data collection, which may have useful lessons for future surveys on international migration.

II. The Case Study: Colombian Migrants, Including Refugees, in Ecuador

Thousands of Colombians have entered Ecuador each year since the late 1990's, due partly to the upsurge in violence accompanying Plan Colombia. Many of these have fled the violence, leaving virtually everything behind, and have sought asylum in Ecuador. The United Nations High Commissioner for Refugees and the Government of Ecuador

(especially the Ministry of Foreign Relations) are very interested in getting better data about these Colombian migrants, which comprise refugees (with applications processed and recognized by the Ministry of Foreign Relations), asylum seekers (who have applied or intend to apply), and others who have applied but been turned down by the Ministry. UNHCR in Ecuador (the local office is known by its Spanish translation, ACNUR).

Existing data available on Colombian migrants in Ecuador is mainly from three sources: (1) admission (border) statistics; (2) the latest (2001) census of population; and (3) data of ACNUR on registered refugees, who are receiving assistance. All of these sources are unreliable. (3) is recognized even by ACNUR as covering only a small proportion of Colombian migrants, those known through administrative records to ACNUR and the Ministry. The census (2) inquired about both place of birth and place of residence five years ago (including country), which makes possible data on the foreign born or resident abroad in 1996, including those born in Colombia or resident in Colombian in 1996. The census led to an estimated population of Colombians of xxx or x% of the total in Ecuador, but less than one per cent of the population enumerated consisted of Colombians who had come in the previous five years, up to the end of 2001. It is known that the census did not enumerate many Colombians, perhaps even most, as they do not have the appropriate documents to be residing and living in Ecuador, as they do. The third data source indicated significant positive inflows of Colombians to Ecuador as far back as the data exist, since 1978, implying an accumulated population of Colombians of 2.3 million persons by the end of 2005, out of a population of 13 million, which is way too high.

The goal of the project was to test innovative methods for collecting data about refugee populations and, for comparative purposes, a control group of other migrants or national populations in Armenia, Ecuador and Sri Lanka. The overall project was funded by UNHCR, and executed by the Netherlands Interdisciplinary Demographic Institute (NIDI), in The Hague. The Ecuador project was designed by Bilsborrow and implemented by the Centro de Estudios sobre Poblacion y Desarrollo Social (CEPAR), based in Quito.[1]

Based on the review of possible data sources above, as of late 2005, it is clear that there was thus no up-to-date sample frame available for selecting a nationally representative sample of Colombians who had come to Ecuador recently, many seeking refuge or asylum from violence. Despite its limitations, the only potentially representative national sample frame was therefore the most recent census of population, carried out in Ecuador on November 25, 2001. We therefore proposed to use this as the principal basis for creating a sample frame for selecting areas where Colombian are concentrated, based on

the prevalence of Colombians who arrived in the 5 year period prior to that census, which essentially refers to the period of calendar 1996 through 2001.

Since the major recent increase in the migration of Colombians to Ecuador due to escalating violence is reported to have begun in 1999, our implicit, and necessary, assumption, is that the flows of Colombians to Ecuador in 1996-2001, as well as their destinations within Ecuador and residence at the time of the census, are similar to the flows and destinations of migrants that have occurred in the more recent period prior to the household survey, which was planned for early 2006. Note that only the year 2001 is a common year, so we would be assuming that the destinations of those arriving in 2001-2006 are similar to those arriving in 1998-2001, if we retain a five-year reference period in the proposed survey. To assist respondents in the survey in recalling dates, we used the date of January 1, 2000, as the cut-off date. Thus we asked households if they had any household member who had arrived from Colombia since January 1, 2000. This means the direct overlap in years between the census and the survey into 2 of the 5 years.

The original goal was to develop a sample covering all the main areas of concentration of Colombians as determined from the 2001 census of population. This would have involved *proportionate sampling* to select a probability sample of households based on selecting provinces, then cantons, then *parróquias* (the lowest level administrative units in Ecuador), and then finally census sectors, based on the proportion of households with Colombians or Colombian-born heads of households (see discussion of sampling methods for studying international migrants in Chapter 6 of Bilsborrow et al, 1997). This would have required an extensive screening operation in *all selected sample census sectors all over the country*, in order to ultimately interview fewer than 1000 households of Colombians. Such an undertaking was judged by UNHCR as far too expensive, which led to a modified proposal to use conduct a survey only in a handful of provinces, and supplementing the use of disproportionate sampling to select areas for the survey, but supplementing this with *snowball or network sampling* to increase the number of Colombians found**.**

The rest of the paper provides the rationale for (a) the selection of study provinces in Ecuador and the use of proportionate sampling to select the ultimate sampling units (census sectors), and of two-phase sampling to select households for interview; (b) the rationale and procedures for the snowball sample portion of data collection; (c) the implementation of (a), with problems encountered, solutions adopted and results; and (d) the implementation of (b), also with problems encountered, solutions adopted, and results. But first we discuss the need for and implementation of sampling procedures appropriate for sampling rare elements, and for snowball sampling as one possible methodology.

III. Sampling Rare Elements

This section will include a discussion of the rationale for using proportionate and disproportionate sampling and two-phase sampling to find and interview recent Colombian migrants, as "rare elements". This will draw on Bilsborrow et al (1997) and

also Groenewold and Bilsborrow (2004, 2006). It will also include a discussion of multiplicity, adaptive, network and snowball sampling, including of the key issue of establishing multiplicity rules for linking households to the main sample. This will draw on Bilsborrow (2006), and references cited therein, such as Sirkin, Goodman, etc.

IV. Implementation of Proportionate Sampling for Finding Colombian Migrants in Ecuador

As noted above, although there is no fully acceptable and up-to-date sampling frame in Ecuador, the best starting point is the 2001 population census. Given the lack of sufficient funding for a national survey, we first determined the main areas where it could be possible to implement a survey, concentrating on provinces (of the 21) with the *highest prevalence* or proportion of Colombians coming to Ecuador in the most recent time period, the five years prior to the census. Data from the latest (November 2001) census of the Ecuadorian Government National Institute of Statistics and Censuses (INEC) indicated that the provinces with the highest prevalence of recent migrants from Colombian are the three northern border provinces of Sucumbios (Amazon), Carchi (Andes), and Esmeraldas (Pacific Coast), followed by the adjoining provinces in the Andes of Imbabura and Pichincha (which has the capital city of Quito). It was fortunate that these contiguous provinces are the top five in migration prevalence since that would help keep transportation/field costs down.

A multi-stage sampling procedure was considered, to sample cantons (similar to U.S. counties, the next administrative level below provinces) and/or parróquias or parishes (the next level below cantons, and the lowest administrative level used in Ecuador). However, it was possible for CEPAR to obtain census data from INEC on the total population and the number of Colombians arriving in the previous 5 years for even lower level units, the census sectors, which would evidently mean the possibility of finding geographic areas with higher concentrations of Colombians than would be the case at higher levels in the hierarchy. Thus if a parróquia had 10% of its population having come in the previous five years from Colombia, then there would be census sectors in that parróquia with more than 10%, as well as others with less than 10%. A sample based on census sectors would then facilitate finding households with recent Colombian migrants, as sought, though it would likely not lead to covering all the cantons nor parróquias in the province.

The final step in creating the sampling frame (the procedures are described in detail in Bilsborrow et al, 1997, Ch. 6 for international migration, and in sampling textbooks, such as Kish, 1965) was then to calculate the prevalence of recent migrants from Colombia for all census sectors in the five provinces. This prevalence was measured as (number of Colombians enumerated in 2001 who did not live in Ecuador in 1996)/(total population of census sector in 2001). Tabulations were run showing the prevalence of Colombians in all census sectors in the five provinces.
Census sectors have on average 120 households in urban areas and about 80 in rural areas. As a result, the number of sectors per province is large, even huge for Pichincha.

We therefore decided *a priori* to exclude all sectors with fewer than 3% Colombians enumerated. This then became the sample frame.

Since the census indicated similar prevalences of migrants in urban and rural census sectors, we also determined a priori to select equal numbers of census sectors in urban and rural areas. Strata were then formed, according to the prevalence of Colombians, with stratum I comprising those sectors with 3-5.9 % Colombians, stratum II comprising those with 6-8.9, III comprising 9-11.9, and IV comprising 12-14.9, with one case of 18 included as well. Proportionate sampling was then used to select urban sectors from the urban list of the five provinces, and separately from the rural list, using systematic sampling. (Procedures will be described more fully in the actual paper.)   Initially, 35 urban and 35 rural sectors were selected, with sampling in proportion to the proportion of Colombians in the census sector. The number of census sectors selected in the sample from the five provinces is indicated as follows (with U=urban, R=rural), in the first set of numbers below (the other numbers are to be described shortly):

Sucumbios	6 U, 13 R	3U, 6R
Carchi	5 U, 13 R	2U, 7R
Pichincha	19 U, 4 R	10U, 3R
Imbabura	3 U, 2 R	2U
Esmeraldas	2 U, 3 R	1U, 1R
TOTAL	35 U, 35 R	18U, 17R

We then obtained from INEC the census sector maps used in the last population census in 2001, which showed major topography including rivers and streams, towns, roads, churches and schools, and isolated dwellings (by dots), which were most useful to locate each census sector. We then conducted a listing operation in the field in each sample sector. This was a major part (and cost) of the fieldwork of the survey, accounting for almost half the total field time. Thus a field team of 4 interviewers and one supervisor would travel to each sample sector and go door to door to determine, and record, each occupied household, noting the number of members and whether it contained any Colombian member who had come to Ecuador since January 1, 2000, or basically in the previous six years. Sketch maps were prepared to permit finding the households occupied by Colombians for interview later, which was usually done in the next day or two. To avoid clustering effects, in the event a census sector had more than 10 households with Colombians who arrived recently, a maximum of 10 was selected by the supervisor using a random procedure, prior to conducting interviews.

(The design and content of the questionnaires will be described in the full paper. The listing operation will also be described further, including problems encountered.)

The plan was to then interview each household sampled using the above procedure. This involved a household questionnaire, administered to the economic head of the household or his/her proxy, and then an individual questionnaire to each Colombian who had come at age 15 or over since 1/1/01. If a household had more than six Colombian adult migrants thus defined, a maximum of six were interviewed, separately.

We were concerned that not enough Colombians would be encountered and interviewed successfully through the procedures above, due to (a) changes in the location of recent immigrants in the 2000-2005 reference period compared to their location in the 1996-2001 census reference period, and/or (b) to deliberate concealment of their status as Colombians during the listing operation, since virtually all the Colombians are residing and working in Ecuador without legal documents. At the same time, we thought that (c) most Colombians immigrants would have a close network of Colombian *paisanos* or friends that they would know and would be willing to point us to, including name and address. By also interviewing them, we could increase the number of total interviews obtained through this supplementary *network or snowball* sample. We anticipated than on average, each individual respondent interviewed would refer us to at least one other Colombian recent migrant. We were also curious about how such snowball sampling would work for finding Colombian migrants.   Snowball sampling had never been used in Ecuador before, as far as we knew, and we had ourselves never had the experience.

So what happened? The short version is that (a) was indeed a problem, as the numbers of households with recent migrants as defined was less than we expected from the sampling design in three of the five provinces, and was equal or more than expected only in the two border provinces of Sucumbios and Esmeraldas. At the same time, we are confident that (b) did not turn out to be a significant problem, mainly because interviewers were very experienced in conducting surveys, sensitive to the concern, and also asked neighbors for confirmation. The biggest disappointment, however, was (c), as most respondents either did not know other recent Colombian migrants living in other households (other than immediate neighbors, who would usually already be in the sample from being in the same census sector), *or* did not want to tell us about them, perhaps to not burden them with answering the long questionnaire. Thus the number of households identified through the snowball procedure was less than one per every four interviews.

(The paper will discuss further a-c, and solutions attempted and results. In addition, I will discuss issues of how the snowball sample might be linked to the main sample via a multiplicity rule so as to keep the total sample a scientific probability sample, which is not customary in snowball samples of this type.)

Because fewer Colombians were found and interviewed than expected in the initial fieldwork, the time of the fieldwork was less than anticipated. And because fewer households were identified for the snowball supplementary sample than expected, that planned last phase of fieldwork was going to take less time than expected. Therefore, it became possible and desirable to draw a supplementary sample of census sectors, from the same original lists, and maintaining the same proportion across provinces and urban-rural sectors. The result is the additional 35 census sectors indicated in the table above, raising the total number of census sectors to 105. (Procedures will be further described in full paper.)

The final result was 534 households interviewed, but only 499 had complete data on all variables, once all data cleaning was completed by CEPAR. The total number of people

was 2028, which included 904 persons who completed individual interviews. Of the 534 total households, 183 were found and interviewed through the snowball procedure. This number was far lower than the 500 or so expected.

V. Conclusions

I will first return to and explain the title of the paper with reference to this survey. Then I will discuss methodological lessons, focusing on how the experience could assist in the design of future surveys of international migrants, whether in developed or developing countries, when there is no adequate comprehensive and up-to-date sampling frame.

References (to be cited)

Bilsborrow, Richard E (2006). "The Design of Samples for International Migration Surveys in Countries of Destination: Methodological Issues and Lessons from NiDi Surveys in Spain and Italy." Washington, DC: The World Bank (unpublished).

Bilsborrow, Richard E., and CEPAR (2006). *The Living Conditions of Refugees, Asylum-Seekers and Other Colombians in Ecuador: Millennium Development Indicators and Coping Behavior*. Netherlands, The Hague: Netherlands Interdisciplinary Demographic Institute, for the United Nations High Commisioner for Refugees.

Bilsborrow, Richard E., Graeme Hugo, A.S. Oberai and Hania Zlotnik (1997), *International Migration Statistics. Guidelines for Improving Data Collection Systems*. Geneva: International Labour Office.

Birnbaum, Z.W., and M. Sirken (1965). Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates. *Vital and Health Statistics*, Series 2, Number 11.

Blair, Johnny (1999). A Probability Sample of Gay Urban Males: The Use of Two-Phase Adaptive Sampling. *The Journal of Sex Research*, vol 36(1): 39-44.

Castles, Stephen, and M. Miller (1998). *The Age of Migration: International Population Movements in the Modern World*. NY: Guilford Press.

Goodman, Leo A. (1961). Snowball sampling. *Annals of Mathematical Statistics*, Vol. 32(1): 148-170.

Goodman, R., and Leslie Kish (1950). Controlled Selection: A Technique in Probability Sampling. *Journal of the American Statistical Association*, vol. 45: 350-72.

Groenewold, George, and R. E. Bilsborrow (2006, In press). Sampling International Migrants: Lessons Learned from the NiDi/Eurostat Study on Migration from West Africa and the Mediterranean Region to the EU. Forthcoming in *Proceedings of European Population Conference on Migration to Europe*, Corrado Bonifazi and Guiseppe Gesano (eds.). Rome.

Kalsbeek, W.D. (2003). Sampling minority groups in health surveys. *Statistics in Medicine*, 22(9):1527-49.

Kalton, Graham (1983). *Introduction to Survey Sampling*. Sage University Paper 35. Sage Publications.

Kalton, Graham, and D.W. Anderson (1986). Sampling Rare Populations. *Journal of the Royal Statistical Society A*, vol. 149 (Part 1): 65-82.

Kish, Leslie (1965). *Survey Sampling*. John Wiley & Sons.

Kramm, E.R., M. Crane, M. Sirken, and M. Brown (1962). A Cystic Fibrosis Pilot Study in Three New England States. *American Journal of Public Health*, vol. 52: 2041-2057.

Lepkowski, James M. (1991). Sampling the Difficult-to Sample. *Journal of Nutrition*, vol. 121(3): 416-423.

Rothbart, George S., M. Fine, and S. Sudman (1982). On Finding and Interviewing the Needles in the Haystack: The Use of Multiplicity Sampling. *The Public Opinion Quarterly*, vol. 46(3): 408-431.

Salganik, Matthew J., and D. D. Heckathorn (2004). Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology*, pp. 1-6.

Schoorl, Jeannette J., Liesbeth Heering, Ingrid Esveldt, George Groenewold, Rolf F. van der Erf, Aline M. Bosch, H. de Valk and Bart J. de Bruijn (2000). *Push and pull factors of international migration: a comparative report*. 3/2000/E/no.14, European Commission/Eurostat. Luxembourg.

Sirken, Monroe G. (1970). Household Surveys with Multiplicity. *Journal of the American Statistical Association*, vol. 65 (239): 257-266.

Sirken, Monroe G. (1998). Network Sampling. In: *Encyclopedia of Biostatistics*. John Wiley and Sons, Volume 4, pp 2977-2986.

Sirken, Monroe G. (1998). A Short History of Network Sampling. *Proceedings of the American Statistical Association*, pp. 1-6.

Sudman, Seymour (1985). Efficient Screening Methods for the Sampling of Geographically Clustered Special Populations. *Journal of Marketing Research*, Vol XXII (February): 20-29.

Sudman, Seymour, and Graham Kalton (1986). New Developments in the Sampling of Special Populations. *Annual Review of Sociology*, 1986, vol. 12: 401-429.

Sudman, Seymour, M. Sirken, and C. Cowan (1988). Sampling Rare and Elusive Populations. *Science*, vol. 240 (20 May): 991-996.

Thompson, Stephen K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, 412: 1050-1059.

United Nations (2006). *International Migration and Development: Report of the Secretary-General*. New York: UN General Assembly, May 18, 2006 (A/60/871).

United Nations (2004). *International Migration and Development*. New York: Department of Economic and Social Affairs.

Wodon, Q., ed. (2004). *International Migration in Latin America: Brain Drain, Remittances and Development*. Washington, D.C.: The World Bank (draft).