The Mexican Migration Project weights [1]

Introduction

The Mexican Migration Project (MMP) gathers data in places of various sizes, carrying out its survey in large metropolitan areas, medium-size cities, small towns and rural villages or *ranchos*. The MMP draws its samples from a vast geography. During the early part of its history, it focused on Western Mexico (mainly the states of Jalisco, Guanajuato and Michoacán.) Later it expanded to include communities located elsewhere, even in Oaxaca or Baja California. The MMP data is not explicitly designed to be representative of Mexico as a whole. It samples more heavily in Western Mexico and its communities are not selected at random.

Once a city, town or *rancho* has been selected, the project directors or the fieldwork supervisor delimit the *survey site*, i.e., the precise area where the survey will be implemented. This is what the project calls a "community". In a city, it is usually a neighborhood or a geographically distinct part of it. In a *rancho,* it tends to be the whole place. The fieldwork team maps the survey site and enumerates all dwellings within it to create a *sampling frame*. Those to be visited are drawn at random from that enumeration. In any given community, MMP fieldworkers typically interview 200 *eligible* dwellings. It is exceptional that the first 200 units drawn from the enumeration will all be eligible: vacant houses and business locales where no one resides, for example, are *non-eligible* for the survey. The project estimates the number of eligible households in the survey site as the total number of enumerated dwellings times the proportion of eligible households among those randomly selected from the frame.

For each community sampled in Mexico, the MMP attempts to sample immigrants who settled in the United States. This is not always possible, especially in the case of communities with only a handful of migrants residing in the US. In the majority of cases, however, the project has been able to supplement the Mexican samples with corresponding US samples —there are US samples for 57 of the 71 communities in MMP71. MMP fieldworkers in the United States rely on field notes and referrals in order to locate immigrants from specific communities who settled north of the border. Hence, the procedure in this case involves neither an enumeration of dwellings nor a random sample. A special computation is necessary to estimate the size of the out-migrant population, as it will be outlined below.

[1] These notes are also applicable to the Latin American Migration Project (LAMP) weights.

Computation of the weights

In surveys that over-sample specific population groups (such as minorities, the unemployed, or the military) weights are essential for the computation of correct descriptive statistics for the population under study. Failure to apply them would lead to naive statements, such as that 50% of the population is unemployed or enrolled in the military. The MMP does not purposely over-sample any particular population group, since households within a community are sampled at random. Hence, MMP weights do not vary by households or individuals. Instead, they are community-specific and sample-specific.

There are two weights for each community: one for its Mexican sample and one for its U.S. sample. A sampling weight is calculated as the inverse of the *sampling fraction*. In the case of the weights for the Mexican communities (variable *mxweight*)[2] the sampling fraction simply is obtained by dividing the number of interviewed households by the estimate of eligible households in the sampling frame. Table 1 shows the computation of the Mexican weights for community 58. The information necessary for these computations is provided by the field supervisor's sampling information worksheet (request a copy of the *MMP & LAMP Field Supervisor Manual* for details.)

Table 1. Computation of Weight in Mexican Sample, MMP71 Community No. 58

total households in the survey area (count)	254
estimate of eligible households (1)	243
visited households	118
eligible	113
interviewed households	**100**
dismissed (2)	10
refusals	3
ineligible	5
vacant dwellings (3)	3
foreign-born head	0
business only (non-residential)	2
does not exist (4)	0
in the USA (5)	0
other ineligible (6)	0

Sampling Fraction (7)	0.41
Weight (8)	2.43

(1) household count * (eligible visited / total visited).

(2) confirmed residential dwellings whose residents could never be found or were unable to respond to the survey in a coherent manner.

(3) empty house or under construction and uninhabited.

(4) incorrectly considered to be an independent household in the original count.

(5) vacant households whose owners, according to neighbors, were "in the United States."

(6) institutions such as a school, a nursing home, a government office, etc.

(7) interviewed households / estimate of eligible households.

(8) 1/sampling fraction.

[2] Variable *doweight* in LAMP data sets.

The U.S. weight, just like the Mexican weight, is calculated as the inverse of the sampling fraction, which results from dividing the number of households interviewed in the U.S. by the total size (number of households) of the U.S. community. Since U.S. samples are constructed through referrals (drawing them at random is financially prohibitive), the population must be *estimated* using some indirect method. The method of choice estimates the U.S. community population as a proportion of the Mexican population for the community under study. The leading question is: how does the number of community *x* U.S. households (those households whose heads were originally from community *x* in Mexico) compare to the number of households *in* community *x* (the Mexican population)?

To answer this question, the project uses data on the children of the household heads in the Mexican sample. The MMP questionnaire gathers data on all household residents as well as those children of the household head who are no longer members of the household. We obtain the proportion we need by comparing the number of children who settled in the United States versus those who left the parental home but stayed in Mexico. Applying that proportion to the Mexican population (the estimate of eligible households within the sampling frame) we estimate the U.S. population (total number of U.S. households) for each community. Table 2 shows, as an example, the computation of the U.S. weights for the same community shown on Table 1.

Table 2. Computation of Weight in U.S. Sample, MMP71 Community No. 58

a)	Total number of children in the home sample, who *do not* live in the household and for whom uscurtrp *is not* unknown *surveypl=1, relhead=3, hhmemshp=1 and uscurtrp ne 9999*	165
Of those selected in (a)…		
b)	Total number of children living in the US *same and uscurtrp=1*	13
c)	Total number of migrant children living in home country *same and uscurtrp=2*	0
d)	Total number of nonmigrant children living in home country *same and uscurtrp=8888*	152
e)	Total number of children living in home country *c + d*	152
f)	Ratio of children in the US to children in home country *b / e*	0.086
g)	Home community sample weight *Take it from field supervisor's worksheet*	2.43
h)	Home community sample size *N for this community*	100
i)	Estimate of eligible households in home community *g * h*	243.237
j)	Estimated number of US households *f * i*	20.803
k)	**Total number of households interviewed in the US** *N for US sample*	10
m)	Estimated US sampling fraction *k / j*	0.4807
n)	**US sample weight: inverse of the sampling fraction** *1 / m*	**2.080**

The actual meaning of MMP weights

When applied, the Mexican weights produce data representative of the area formed by all of the sampling frames. This area is neither “Mexico”, nor “Western Mexico”, nor the combination of states included in the survey. It is not the combination of cities, towns and *ranchos* where the survey was fielded either. It *is* the combined population of all sampled communities. As of today, this sample area is quite large: it includes 71 communities scattered throughout the Mexican landscape and their counterparts in the US. Just as it is not a geographically continuous area, it is not an area fixed in time either. Each MMP community was surveyed once, between 1982 and 1999.

Descriptive analysis

Weighting is recommended when computing descriptive statistics for *analytic* purposes. Otherwise, these statistics would be biased in the direction determined by the characteristics of the population sampled in small towns and *ranchos*, where sampling fractions tend to be higher. Conversely, without weighting, households in urban areas would be underrepresented since their likelihood of being interviewed is generally lower than that of households in a small town or rancho (but see discussion relate to figure 1, in the paragraphs below).

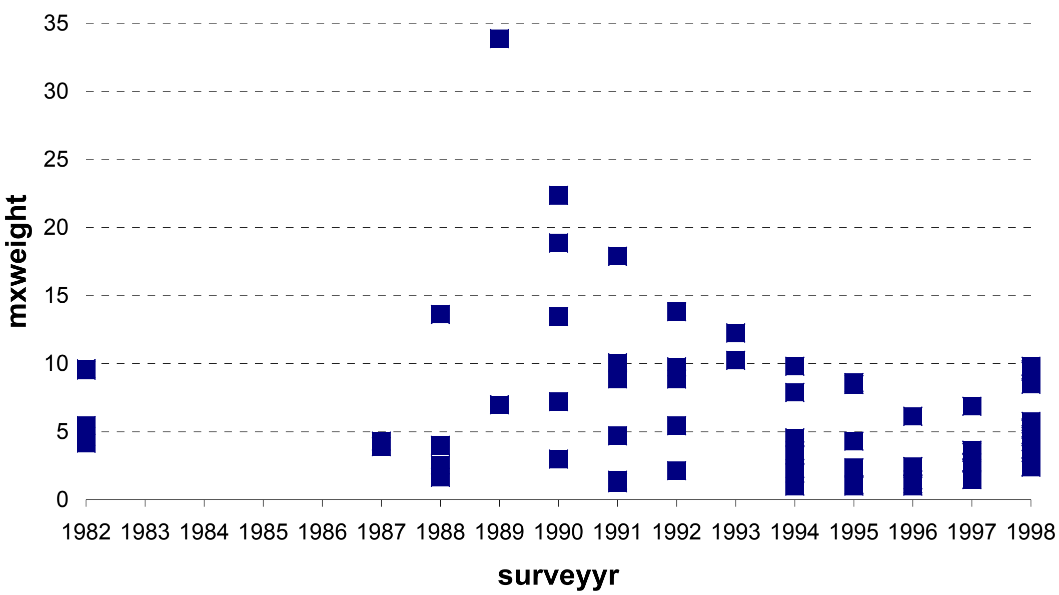
Consider, for example, MMP communities 37, 44 and 45, which are *ranchos* and show the lowest value of *mxweight* –1.00– and communities 9 and 11, urban areas thus presenting the largest value of *mxweight* – 33.88 and 22.36, respectively. Table 3 shows the percentage of male migrants and the average number of U.S. trips, for men age 15 and over, both unweighted and weighted. Because migration is more prevalent in the smaller communities, the unweighted statistics overestimate both the percentage of migrants and the average number of U.S. trips. For the five communities being analyzed, smaller areas are more likely to be fully sampled and more likely to show high sampling fractions, which produce small weights.

Table 3. Males, age 15+, Mexican Samples, MMP71 Communities No. 9, 11, 37, 44, 45:
Unweighed and Weighed Means for Selected Variables

	community					all combined	
statistic	**9**	**11**	**37**	**44**	**45**	**unweighted**	**weighted**
% migrants	44.9	38.7	78.9	53.5	54.3	52.0	43.3
mean of USTRIPS	1.51	1.17	2.37	2.03	2.26	1.81	1.42
MXWEIGHT	33.88	22.36	1.00	1.00	1.00		
N	492	424	289	275	416		

This simple example may be effective enough to make a case for weighting descriptive MMP statistics. Yet some questions arise from the fact that survey sites are delimited rather arbitrarily. The denominator in the computation of weights is the total population of households, which depends on the dimensions of the survey site. Knowing their team capabilities, supervisors of large fieldwork teams are likely to draw larger survey sites than those in charge of small teams. In this sense, weights are somewhat dependent on project resources at the time of putting different surveys on the field. Beyond this inevitable variability, the MMP makes an effort to keep consistent fieldwork procedures and avoid too much unnecessary variability in survey site dimensions. Figure 1 shows that, shortly after the project went back to the field in 1987, these efforts have improved over time, keeping weight variability at relatively moderate levels.

Figure 1
Mexican weights by survey year



Sometimes descriptive statistics are provided for *reporting* rather than *analytic* purposes. The typical situation consists of a table of means and standard deviations that precedes a regression. In such a case, if the regression procedure does not weight the data, it seems reasonable not to weight the data when reporting sample means and standard deviations either. Readers are more likely to be interested in the means and standard deviations of the variables included in a causal model as *they are* – just as they will be used in the model.

The U.S. weights compensate for the varying estimated sizes of the migrant communities in the United States and provide a tool for the measurement of the relative contribution of each U.S. sample to the binational migrant community. The U.S. sample weights do not solve the problem of unrepresentativeness of the U.S. samples, which are constructed through referrals using 'snowball' sampling methods. The researcher may choose to dismiss the U.S. samples altogether on pure statistical grounds. Depending on the research problem, this may or may not limit the scope of the analysis.

Causal modeling

The MMP data is most suitable for the estimation of causal models that intend to shed light on the complexities of the migration process. Massey and Zenteno (2000; Zenteno and Massey 1999, in Spanish) show that the MMP data, although only partially drawn at random, is roughly comparable to Mexican representative national-level data from the Encuesta Nacional de la Dinámica Demográfica (ENADID). Moreover, MMP data is actually superior to ENADID's when estimating causal models, for two reasons. First, the vast amount of quantitative information contained in the MMP database allows for complex modeling. Second, by including data from migrant households located in the United States, the MMP reduces selectivity problems found in ENADID, since the latter only surveys households in Mexico.

As stated by Massey *et al*. (1987: 12-13):

> The ethnosurvey is not, of course, the last word in studying international migration. One is still faced with the issue of generalizability. The ethnosurvey is not a technique for aggregate statistical estimation. Facts and figures computed from ethnosurvey data cannot be easily extrapolated to the rest of Mexico or to the population of Mexican migrants. What the method does provide is a way of understanding and interpreting the social processes that underlie the aggregate statistics. The strength of the ethnosurvey is that it provides hard information so that the social process of international migration can be described to others in a cogent and convincing way.

When the MMP data is used at its best – to estimate causal models– weighting is, in most cases, unnecessary. Regression models assume that the coefficients are the same for everyone in the sample. If that is truly the case, there is no good justification for weighting. If it is not the case, the most likely solution is to re-specify the model, for example including the appropriate interaction terms, or to split it into separate models for different population groups.

Winship and Radbill (1994) show that when the model has been correctly specified and the weights are solely a function of independent variables, then unweighted OLS estimates are unbiased, consistent, and have smaller standard errors than weighted OLS estimates. Thus, the use of sampling weights will produce less efficient parameter estimates than the use of the unweighted sample (p.244 and Table 1). Furthermore, they demonstrate that noticeable differences between the estimates derived using the unweighted and the weighted sample are a good indication that the model is not correctly specified or, alternatively, that the weights are a function of the dependent variable. When the latter is the case, the authors recommend reconsidering the model specification. If it is not possible to re-specify the model, the use of sample weights may be more appropriate, coupled with White's heteroskedastic consistent estimator for the standard errors (White 1980). In the case of the MMP data, the weights are solely a function of the community dummies and the survey place, and there seems to be no need to use the weights for causal modeling.

References

Massey, Douglas S., Rafael Alarcón, Jorge Durand, and Humberto González. 1987. *Return to Aztlan: The Social Process of International Migration from Western Mexico*. Berkeley and Los Angeles: University of California Press.

Massey, Douglas S. and René Zenteno. 2000. "A Validation of the Ethnosurvey: The Case of Mexico-U.S. Migration." *International Migration Review* 34(3):766-793.

Winship, Christopher and Larry Radbill. 1994. "Sampling Weights and Regression Analysis." *Sociological Methods and Research* 23(2):230-257.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Hereroskedasticity". *Econometrica* 48:817-838.

Zenteno, René M. and Douglas S. Massey. 1999. "Especificidad *versus* representatividad: enfoques metodológicos en el estudio de la migración mexicana hacia Estados Unidos." *Estudios Demográficos y Urbanos*: 14(1):75-116.