Short or Long? Uncertanties and Errors in the Measurement of Internal Migration through Population Censuses: the Case of Mexico

Estela Rivero-Fuentes
El Colegio de México
estela.rivero@colmex.mx

Edith Yolanda Gutiérrez Vázquez
El Colegio de México
egutierrez@colmex.mx

In this paper we use data of the 2000 population census in Mexico, where the questions on recent migration were asked both on the short and on the long form census questionnaires, to evaluate the quality of long form estimates. We conduct this evaluation in terms of the probability of emigration and immigration conditional on survival at the end of the exposure period and remaining on the country, for two different units of analysis: the state and the municipality. In our assessment we try to understand to what extent the differences found between the survey and the census are due to the sampling procedure and to what extent they are due to the weighting process. Our results show that survey design is not adequate for capturing the variability of internal migration, and that it captures particularly badly those flows that are small.

Population and housing censuses are one of the most commonly used data sources for the study of internal migration in both developed and developing countries. Their importance relies on the fact that internal migration is oftentimes a rare phenomenon, which implies that its incidence cannot be estimated with precision without large enough samples or a continuous population registration system. This problem is particularly acute in the case of small geographical units and when the interest is to study migration flows determining places of origin and destination.

In its most recent edition of the Principles and Recommendations for Population and Housing Censuses, the United Nations recommend that censuses collect information on either place of previous residence and duration of residence, or on place of residence at a specified date in the past, and that these questions are asked to all the population resident in the country at the time of the census (United Nations 2007).[1] Furthermore, it is recommended that data are recollected for reasonably small geographical units (United Nations 1970).

[1] Similar recommendations have also been issued by the Center for Global Development (Santo Tomas, Summers and Clemmens, 2009)

Because respondents oftentimes have difficulties remembering the places where they lived in the past (especially in very mobile populations), and several localities can have the same name, the questions that are used for recent internal migration are difficult to implement and require intensive training, both for the interviewers and for those who will be coding the questionnaires. This, combined with the fact that population censuses are expensive to carry has lead many countries to choose not to include the questions relative to migration in the main census questionnaire, but rather, to move them to the survey that accompanies the census exercise (also known as the census long form). This approach has been taken in China and in many Latin American countries such as Argentina, Brazil, and Mexico.

It is argued that when the questions of recent internal migration are included in the survey instead of in the main census questionnaire, in addition from saving money the phenomenon will be better measured because the questions will be asked by interviewers that are better trained and supervised. However, the measurement that comes from the survey is from statistical principles less precise, and it may be subject to two sources of error, which may be very important if the patterns of migration have changed in the country since the last census. Estimates of internal migration based on the survey that accompany the census may err because:

1) The sample may not include geographical units that represent adequately the variation of migration regimes in a country, state, or municipality. Commonly, the geographical units that are included in a sample are selected based on a combination of its population size and other sociodemographic characteristics such as its education levels and aggregate income at the time of the last census.[2] These indicators are associated to emigration and immigration, but do not necessarily link directly with all migration patterns. Furthermore, the sociodemographic map of the country (including that of migration) may change between census and census, thus leading some of the new regions of emigration or of immigration to be under-represented in the sample, while older regions that have lost importance may be over-represented.
2) Even if the sample selection did adequately represent the different migration regimes, the fact that the sampling frame is based on the results of the previous

[2] The selection of the geographical units is oftentimes based on indicators at the time of the last census because that is the last observation for which reliable data is available.

census may represent a problem, this time for the sampling weights. This would happen if the distribution of the variables that are being used to make the sample selection (population size, aggregate income; education level; etc.) changed in the intercensal period, and the sample weights were not adjusted for these changes. It is important to notice here that this problem is quite difficult to solve, because it pertains not only to the general weight of each sampling unit, which can be adjusted after the new census estimates are available, but to the relationship of the variables used to select the sample and the variables that will be estimated in the sample, which can be changing over time.[3]

3) In most countries the sample selection is meant to be produce estimates of indicators that have less variability than migration, such as income, fertility, and alphabetization levels.

In this paper we use data of the 2000 population census in Mexico, where the questions on recent migration were asked both on the short and on the long form census questionnaires, to evaluate the quality of the estimates obtained from the long form. We compare the estimates based on the complete census data and on its accompanying survey, in terms of the probability of emigration and immigration conditional on survival at the end of the exposure period and remaining on the country, for two different units of analysis: the state and the municipality –the minimal level of representativity of the census long form. In order to understand to what extent the differences found between the survey and the census are due to the sampling procedure and to what extent they are due to the weighting process, we compare the census estimates with those that are produced from the sample, weighted and unweighted.

This evaluation effort is similar to those taken by Gage (2006), Hough and Swanson (2006) and others when the American Community Survey was being assessed as a replacement of the census long form (see the especial issue of Population Research and Policy Review vol. 25 for an overview of the discussion at the time). However, contrary to the evaluation of the American Community Survey, we conduct our evaluation against the complete census estimation and not against another sample. Furthermore, we focus on migration, a rare indicator –something that to our knowledge had not been done before,

[3] Past comparisons of census and surveys, and how much these are due to sampling weights consider the overall effect of sampling weights, but not how the effect of these may be changing over time for specific questions. See for example Symens Smith (1998), Gage (2006) and Hough and Swanson (2006).

and see how well the sample fares for estimating migration in different conditions. That is, we assess whether the precision of the sample estimates are associated with the level of migration in the geographical unit.

Background: The measurement of recent internal migration in population censuses in Mexico

Of the different possible questions that can be asked in population censuses and its related surveys to measure recent migration, place of previous residence and duration are the more recommendable, since they are the only one that provides the information necessary to estimate migration rates (Xu-Doeve 2008).[4] On the other hand, place of residence at a fixed prior date has the advantage that it relies in only one question, that the information produced is easily manipulable, that the migration interval is clear cut, and that it allows to easily calculate migration streams (United Nations 1970)[5].

Until 2010, the Mexican population census had included at least two questions on migration on its basic questionnaire: place of birth, and place of residence at a fixed prior date.

In 2000, Mexico implemented for the first time in its history with a design where it had both a short and a long census questionnaire. The long questionnaire included 17 household characteristics, and 50 individual characteristics including among others internal and international migration, fertility and infant mortality, presence of disabilities, access to medical services, education level, economic activity and income. The short questionnaire included only 14 household characteristics, and 29 individual characteristics including internal migration, fertility, education level, economic activity and income. The question on recent internal migration, place of residence five years prior to the census –with a level of specification of the municipality-, was included both on the long and on the short questionnaire (INEGI 2000a).

[4] Differently from the United Nations, Xu-Doeve recommends that Population Censuses focus their data collection efforts on the question on duration of residence and do not ask about place of reside at a fixed point in time in the past. He also suggests that, if possible, censuses do not inquire only about the last move. This, he argues, would allow capturing multiple moves and circular migration (Xu-Doeve 2008).

[5] Amongst the drawbacks of the question on place of residence on a certain date are that people may have more difficulties remembering where they lived on a certain date –especially in very mobile populations, than their last place of residence; that multiple moves and circular migration on the reference period go unnoticed (United Nations 1970 and Bell et al. 2002); and that they miss migrants who are born or die during the measurement period (Bell et al. 2002).

The sample was selected to be representative at the level of the municipality and of those localities that had a population of 50,000 or more. The sampling design used was a one-stage cluster design, where complete geographical units were chosen. In other words, complete minimal geographical units (agebs by its Spanish acronym), blocks, or rural localities were selected. All municipalities in the country and all localities that, according to the 1995 population census, had a population of 2,000 or more were represented in the sample. In order to facilitate field operations and supervision, and the posterior estimation of the indicators, all households in the sampled units were given the long questionnaire, whereas the households in the non-sampled units were given the short questionnaire (INEGI 2000b).

In the 2010 census, the recent migration question will only be included in the census long form. However, there has not been, to our knowledge, any evaluation of whether this mechanism is adequate or not to produce the different indicators of internal migration needed in the planning of local needs (for example the probability of emigration and immigration, future number of immigrants and emigrants, and migration flows).

Methodology

Using data from the 2000 Mexican long and short population census forms, we estimate several indicators of recent migration, including:

a) Probability of immigration conditional on survival to the date of the census and remaining on the country[6], at both the state and the municipality level
b) Probability of emigration conditional on survival to the date of the census and remaining on the country, at both the state and the municipality level
c) Probabilities of state to state migration
d) Probabilities of municipality to municipality migration

[6] We decided to estimate conditional probabilities of emigration and immigration rather than rates, using the population at the end of the exposure period, to guarantee consistency in our data (since all the data comes from the same source) and comparability between the different geographical units we would be comparing. Had we calculated rates, we would have had to use population estimates at the beginning of the exposure period (1995), and the data quality of these estimates could have varied between states and municipalities. Using the same data source, however, we are making the same assumption for every unit. That is, that individuals who die and migrate internationally are absent from both the numerator and the denominator. This approach is also favored by Bell et al. (2002).

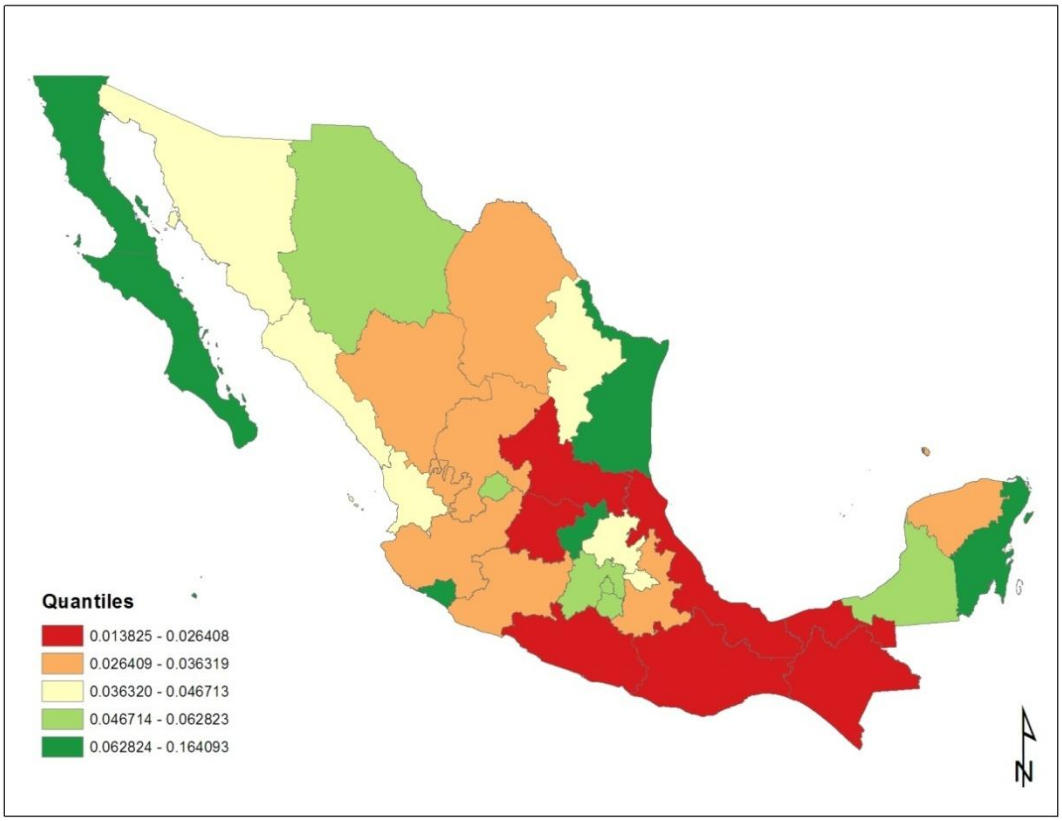
All these indicators are calculated based on the whole census data and on the census long form, using the weighted and the unweighted estimators. We also estimate the standard errors of the estimators based on the census long form, and with these, calculate their 95% confidence intervals. We compare the census-based indicators with both the weighted and unweighted estimators.

Preliminary results

The following maps present the absolute and relative differences between the census results and the weighted and unweighted sample estimations for the immigration and the emigration probabilities. Maps are shadowed[7] in quantiles were the green colors represent the highest levels and the red colors the lowest ones. Fist we analyze the immigration results, we compare the distribution of immigration levels with differences between census and sample indicators to test whether states with the lowest levels are also places with the highest errors. Then we do the same procedure with the emigration dimension.

[7] Colors from www.ColorBrewer.org by Cynthia A. Brewer, Geography, Pennsylvania State University.

Map 1. Mexico, 2000: Immigration probability, Census Estimation

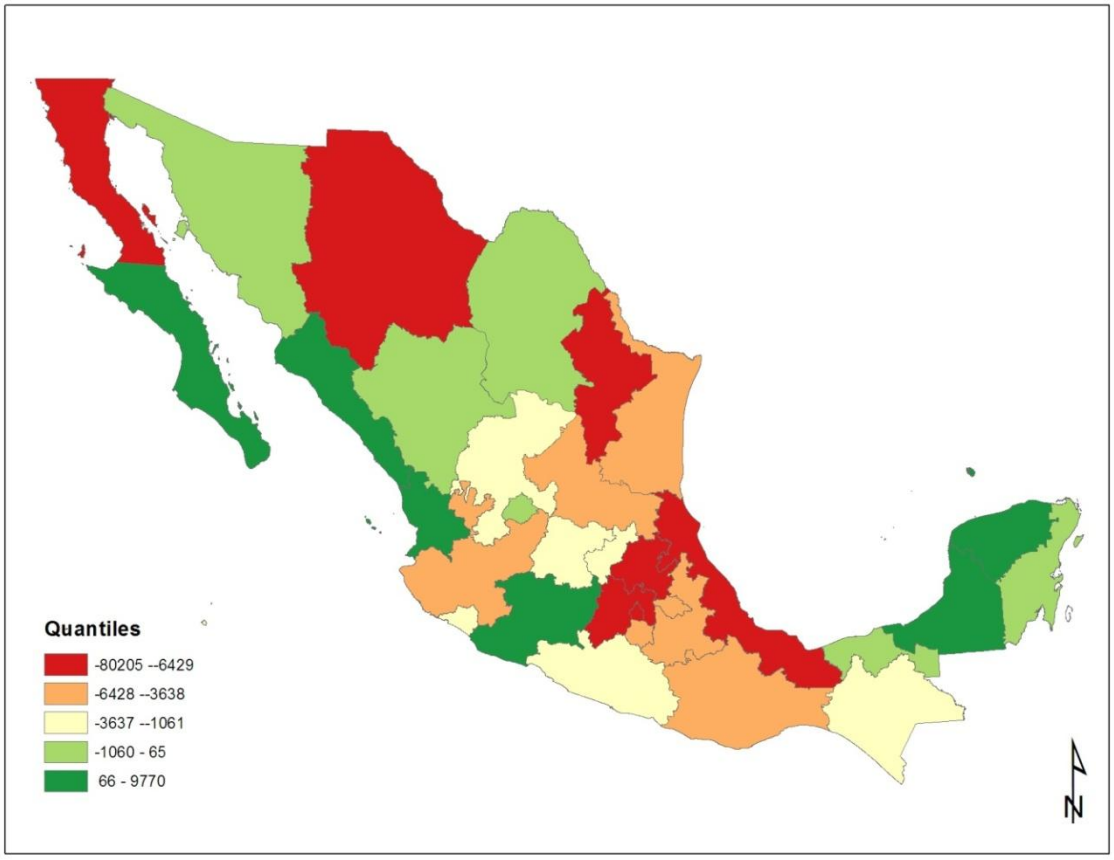


Source: Author's calculation based on the 2000 Mexican Census

In Map 1, we observe the geographical distribution of the immigration probability. It is possible to see that states with the lowest immigration level are concentrated in the Southern Mexico and in the Golf coast, while the highest levels are more distributed across the country: there are three states in the North region, two in the Center and West, and one in the South.

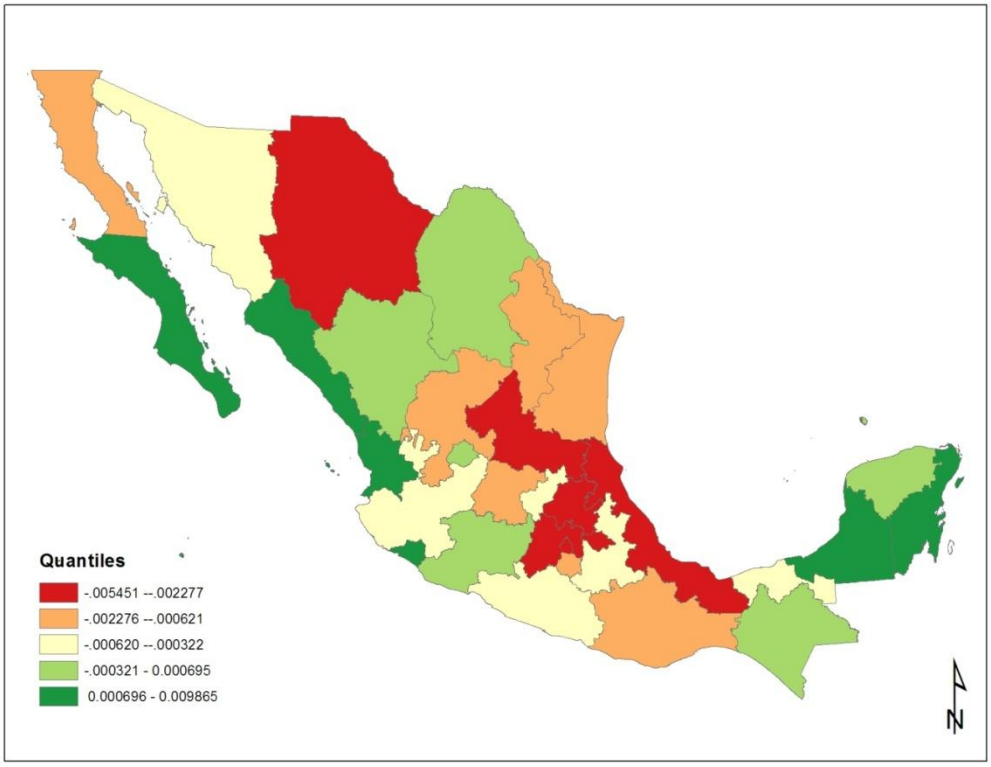
Analyzing the differences between the weighted and the census results, we found that both in absolute and in relative terms, places with the lowest proportion of immigrant population are those where the survey estimates are worst (Maps 2 and 3). This result is corroborated in Graph 1: twenty seven of the thirty two states have levels of immigration smaller than half range (0.08), and eleven of the twenty seven have errors above the 5 percent. Additionally we could see it is more common that the survey overestimates immigration probabilities: almost 60% of the calculations are below zero.

Map 2. Mexico, 2000: Differences in Immigration Volume
Census Vs. Weighted Estimation



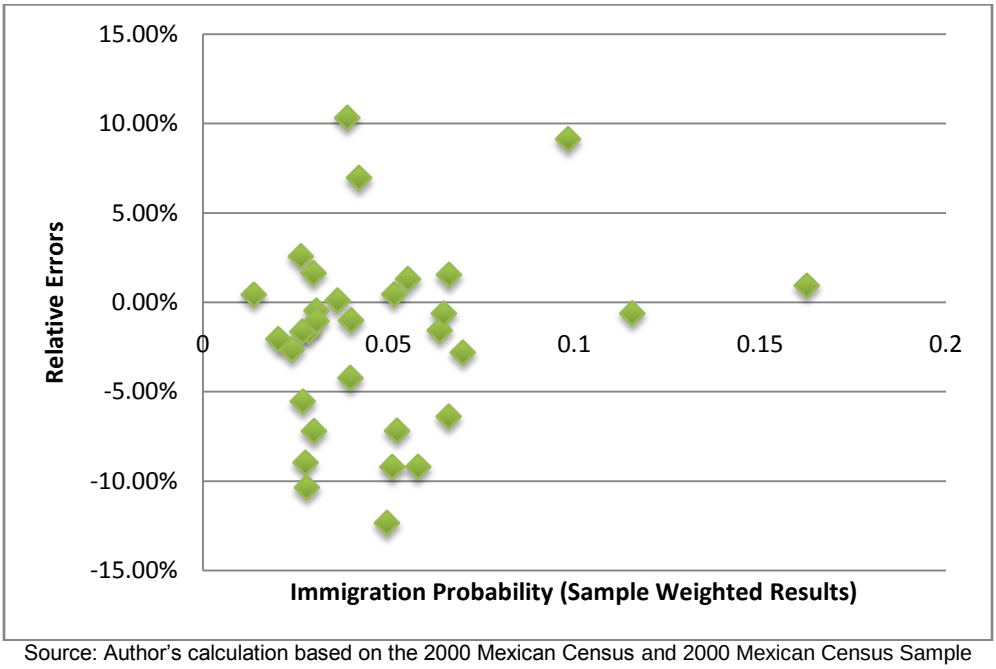
Source: Author's calculation based on the 2000 Mexican Census and 2000 Mexican Census Sample

Map 3. Mexico, 2000: Differences in Immigration probability Census Vs. Weighted Estimation



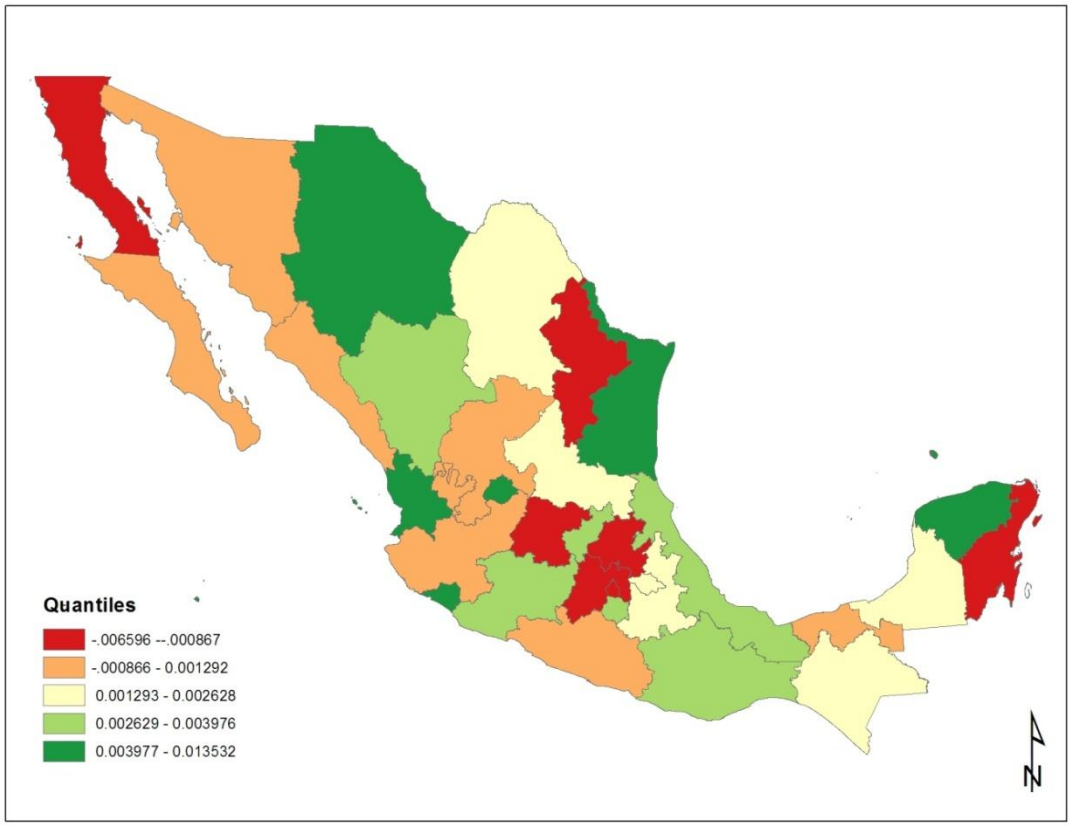
Source: Author's calculation based on the 2000 Mexican Census and 2000 Mexican Census Sample

Graph 1. Mexico, 2000: Relative Errors and Immigration probabilities Census Vs. Weighted Estimation



Source: Author's calculation based on the 2000 Mexican Census and 2000 Mexican Census Sample

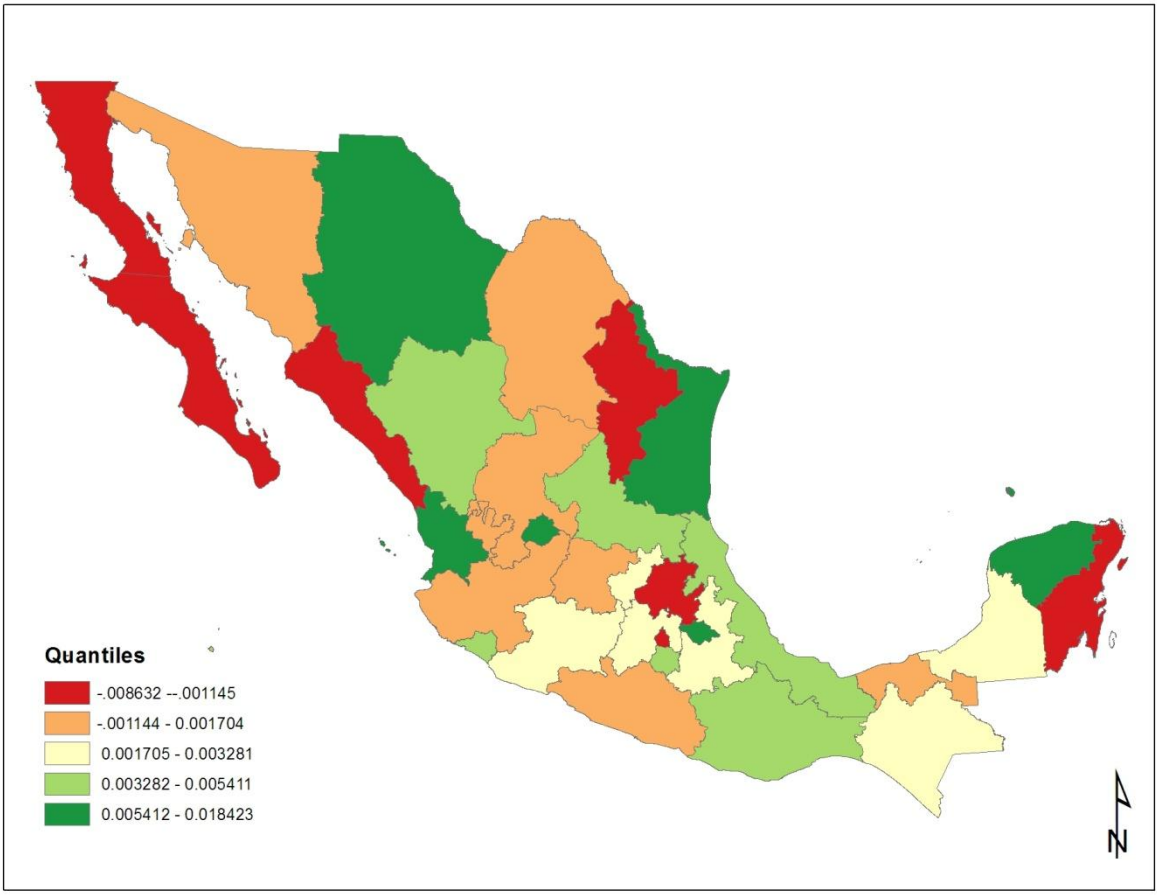
Map 4. Mexico, 2000: Differences in Immigration Probability Census Vs. Unweighted Estimation



Source: Author's calculation based on the 2000 Mexican Census and 2000 Mexican Census Sample

As mentioned before, errors could be due to problems with the sampling procedure and with the weighting process. So, in an attempt to prove if the weights are causing the variations showed before, we map the differences between unweighted results and the census and weighted estimations. Comparing both maps (Maps 4 and 5) with the maps 2 and 3, we realize that places with the largest errors are those where the differences obtained from the analysis using unweighted data are biggest too (negatively or positively). This preliminary analyses leads to the conclusion that estimation issues are due to survey design that is not considering immigration as a factor of sample's precision, as this methodological documentation points out.

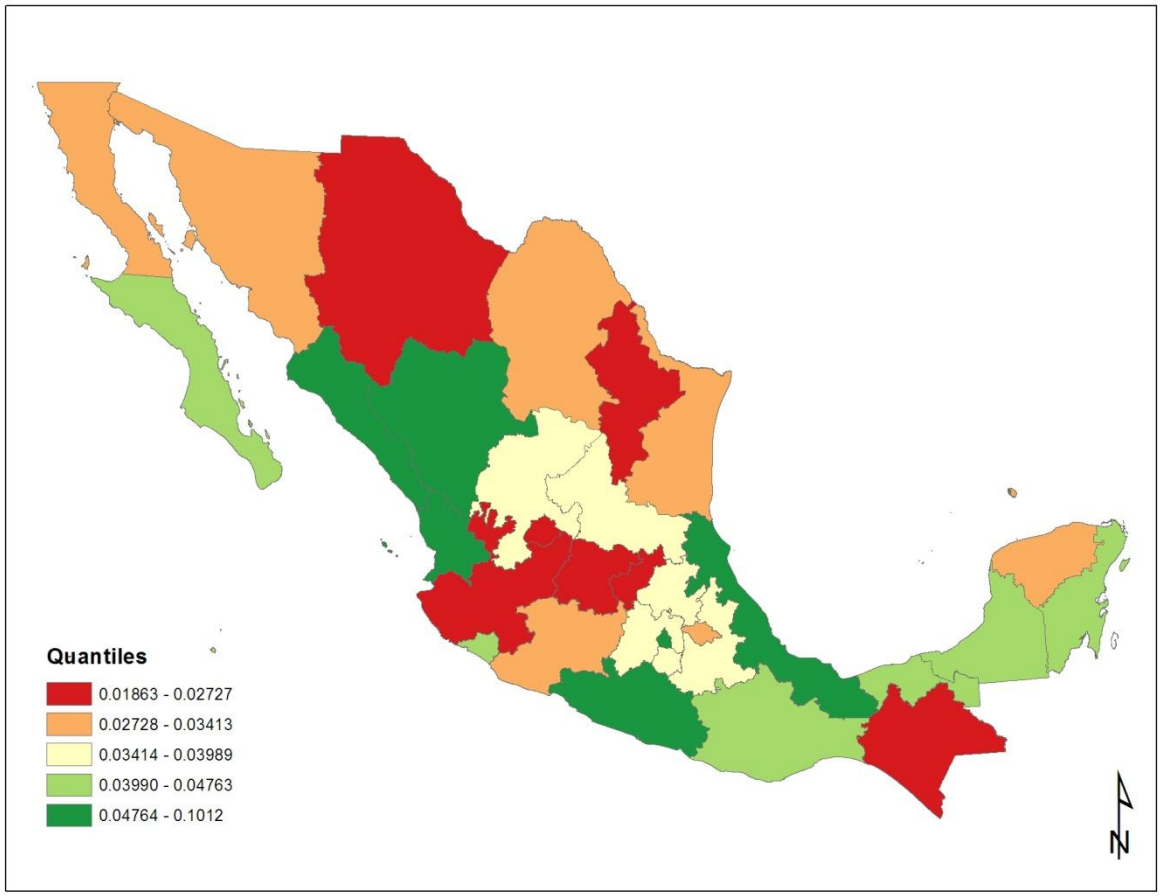
Map 5. Mexico, 2000: Differences in Immigration Probability Weighted Vs. Unweighted Estimation



Source: Author's calculation based on the 2000 Mexican Census Sample

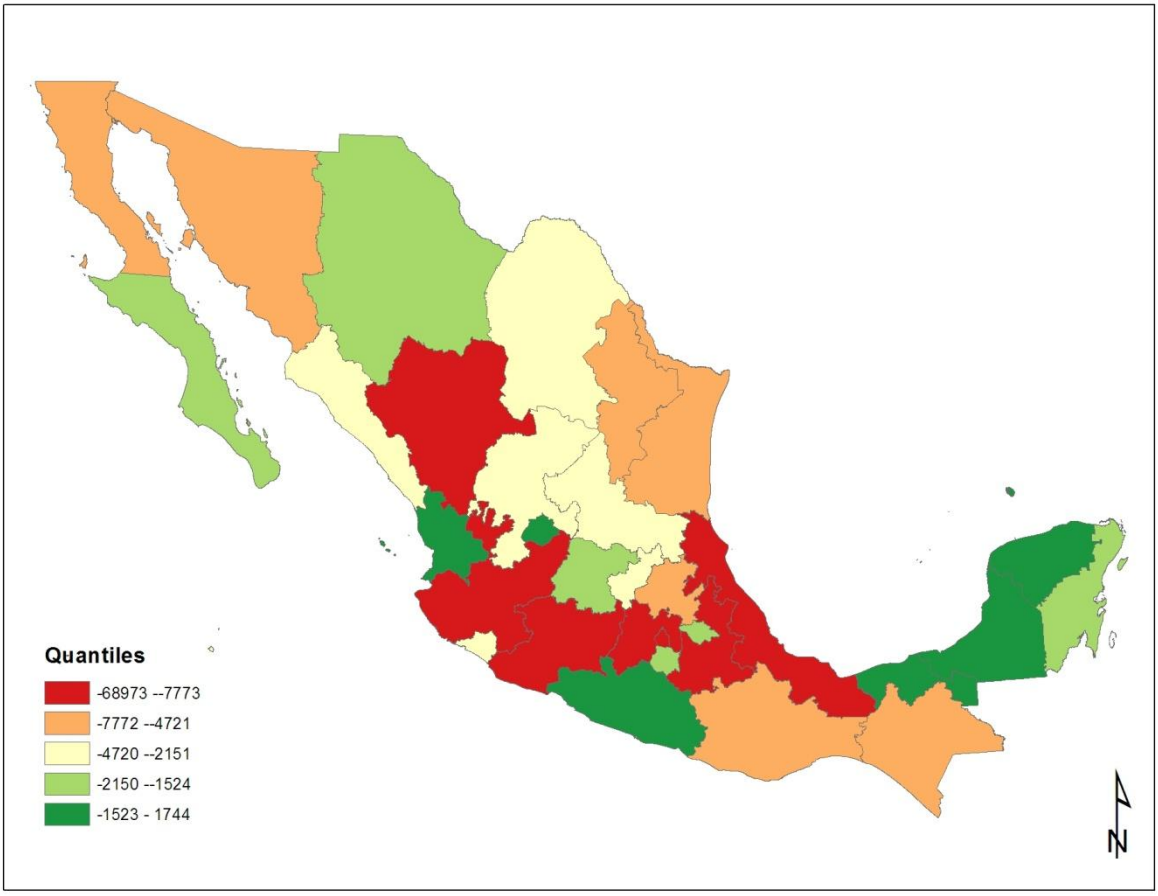
For the emigration probability, patterns are more diverse: states with low emigration are mainly concentrated in the Central-West and states with the largest expulsion of population are in the South-Gulf. Following the methodological strategy exposed in immigration analysis, we map differences between census and weighted sample estimations. Red areas, where emigration it is not so common, have the worst estimations, and the reversing process is happening in areas with low emigration flows. But, two states with high emigration levels have really bad estimations,

Map 6. Mexico, 2000: Emigration Probability, Census Estimation



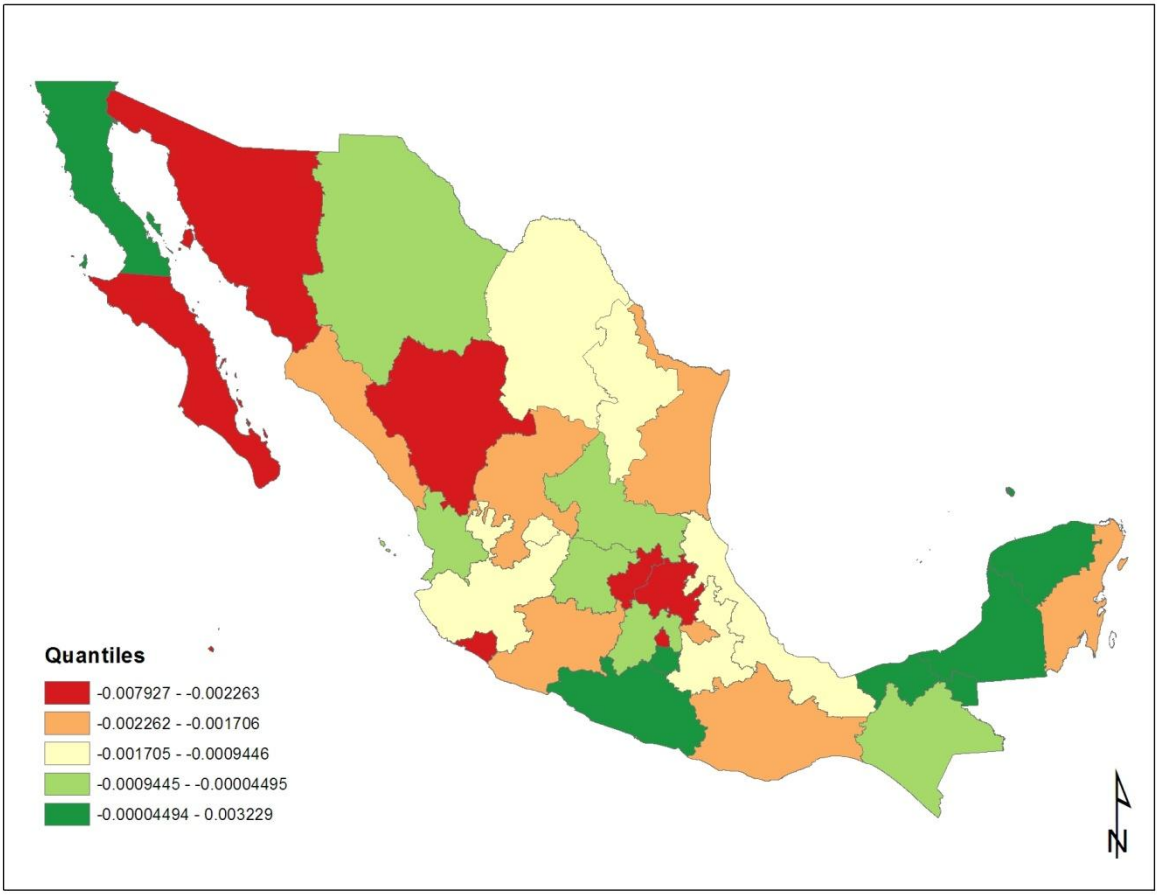
Source: Author's calculation based on the 2000 Mexican Census

Map 7. Mexico, 2000: Differences in Emigration Volume
Census Vs. Weighted Estimation



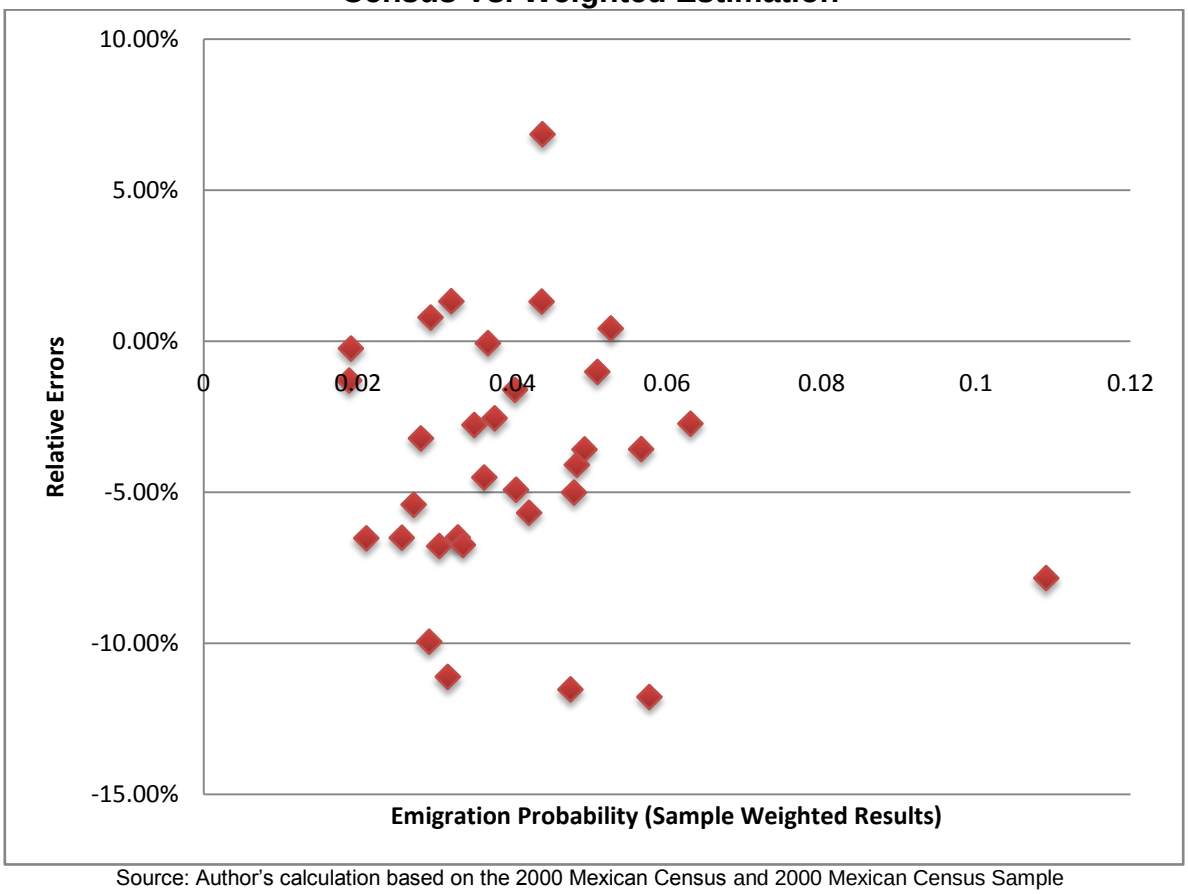
Source: Author's calculation based on the 2000 Mexican Census and 2000 Mexican Census Sample

Map 8. Mexico, 2000: Differences in Emigration Rate
Census Vs. Weighted Estimation



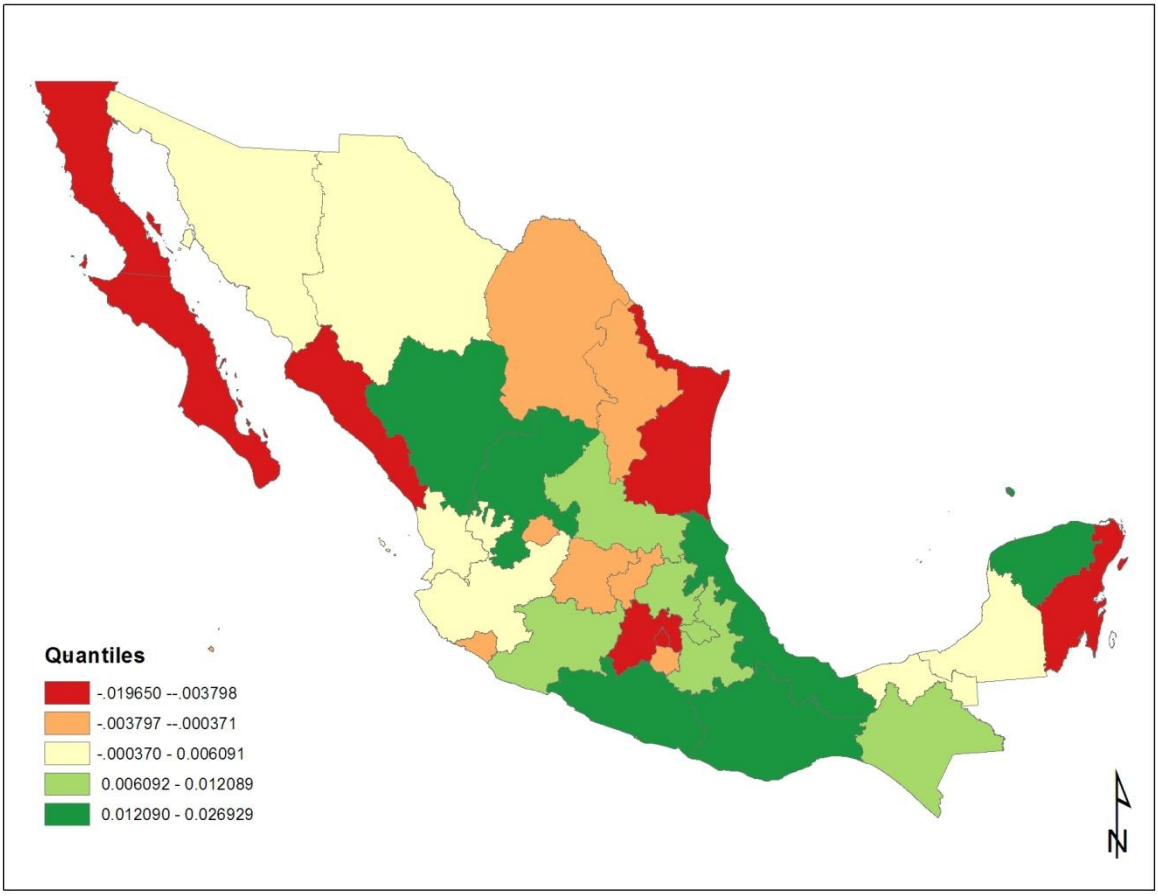
Source: Author's calculation based on the 2000 Mexican Census and 2000 Mexican Census Sample

Graph 2. Mexico, 2000: Relative Errors and Emigration Rate
Census Vs. Weighted Estimation



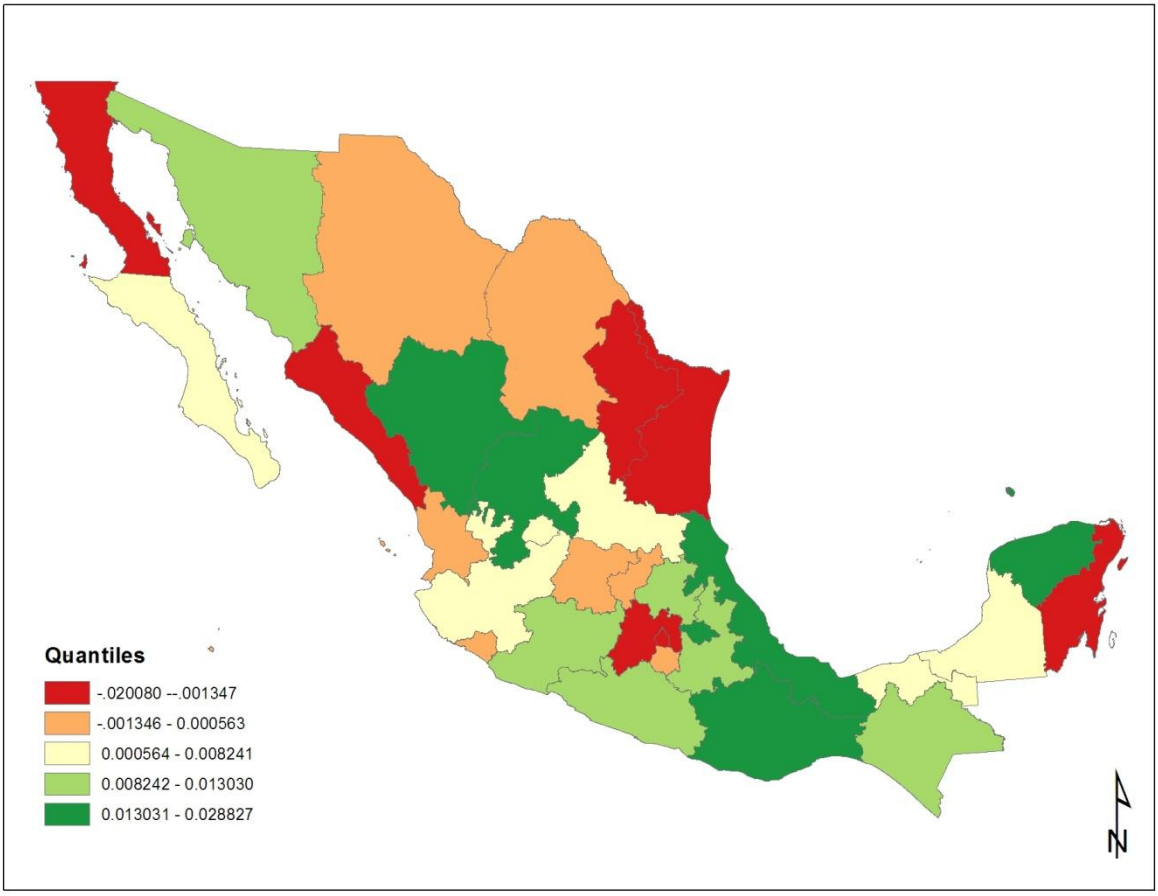
Source: Author's calculation based on the 2000 Mexican Census and 2000 Mexican Census Sample

Map 9. Mexico, 2000: Differences in Emigration Probability
Census Vs. Unweighted Estimation



Source: Author's calculation based on the 2000 Mexican Census and 2000 Mexican Census Sample

Map 10. Mexico, 2000: Differences in Emigration Probability Weighted Vs. Unweighted Estimation



Source: Author's calculation based on the 2000 Mexican Census Sample

From this preliminary results we can conclude that the survey is not an adequate substitute for the complete census estimates; that survey design is not the appropriate for capturing internal population mobility and that it may lead to erroneous migration estimates. Furthermore, we have detected that the direction of the error is not consistent, as flows are sometimes overestimated and sometimes underestimated. For both the weighted and the unweighted estimates, the errors are larger for the states that have smaller migration probabilities. However, the errors also seem to be larger for the unweighted results which seem to imply that most of the errors come from the sampling procedure.

To explore these issues further, in the full extent of the paper we will estimate the standard error of the sampling estimates and calculate confidence intervals, so the census indicators can be properly compared against the survey estimators. We will also make

comparisons at the municipal level and based on state to state flows and regions, and try to relate these to the variables that are used to select the sample. These should shed light, we think, on alternative ways to estimate weights or calculate the sample if the full census is not to be used.

References:

Bell, M; M. Blake; P. Boyle; Duke-Williams, O; P. Rees; J. Stillwell; and G. Hugo. 2002. Cross-national comparison of internal migration: issues and measures. In Journal of the Royal Statistical Society. 165 part 3, pp. 435-464.

Brewer, C. 2010. ColorBrewer, On-line: http://www.ColorBrewer.org , accessed January 18, 2010.

Gage, Linda. 2006. "Comparison of Census 2000 and American Community Survey 1999-2001 Estimates: San Francisco and Tulare Counties, California" . In Population Research and Policy Review, 25: 243-256.

Hough, George C. and David A. Swanson. 2006. "An Evaluation of the American Community Survey: Results from the Oregon Test Site". In Population Research and Policy Review, 25: 257-273.

INEGI (Instituto Nacional de Estadística y Geografía). 2000a. Cuestionarios Censales. www.inegi.gob.mx (consulted online September 15, 2010).

INEGI. 2OOOb. XII Censo General de Población y Vivienda. Base de Datos de la Muestra Censal. Indice de Temas. (XII General Population Census. Data base of the Censal Sample. Topic Index). 111 pp.

Santo Tomas Patricia A.; Lawrence Summers; and Michael Clemens. 2009. Migrants Cont. Five Steps Toward Better Migration Data. Report of the Commission on International Migration Data for Development Research and Policy, Center for Global Development.

Symens Smith, Amy. 1998. "The American Community Survey and Intercensal Population Estimates: Where Are The Crossroads?".US Census Bureau, Population Division Technical Working Paper No. 31.

United Nations. 1998. Recommendations on Statistics of International Migration. *Revision 1*

United Nations. 2007. Principles and Recommendations for Populaton and Housing Censuses 2

United Nations 1970. Manuals on Methods of Estimating Population. Manual VI: Methods of Measuring Internal Migration. Sales No. E.70.XIII.3.

Xu-Doeve, William LJ. 2008. Introduction to the Measurement of Internal and International Migration. ANRC Publishing. The Netherlands. 138 pp.