COMPUTING POPULATION WEIGHTS FOR THE EFH SURVEY*

Carlos Madeira**

I. INTRODUCTION

The Chilean Household Financial Survey (EFH) is a survey about financial assets and debts of Chilean households that has been implemented by the Central Bank of Chile since 2007, in coordination with different survey providers (University of Chile from 2007 to 2009, University Alberto Hurtado from 2010 to 2011, and Ipsos in 2014 and 2017). There are now seven waves of the EFH project, of which four are representative at the national level (2007, 2011, 2014, 2017) and three are representative at the level of the Santiago Metropolitan area (2008, 2009, 2010).

The EFH survey seeks a rigorous estimation of the financial risk incurred by Chilean households. Research applications range include studies of the distribution of households' assets and debt (Central Bank of Chile, 2009), the effects of unemployment cycles on debt risk (Fuenzalida and Ruiz-Tagle, 2009; Madeira, 2014, 2018b), household stress tests (Madeira, 2019b), the exercise of mortgage prepayment or renegotiation options (Madeira and Pérez, 2013), lender repayment priorities (Madeira, 2018a), the motives behind consumer loans (Madeira, 2015), borrowing constraints (Ruiz-Tagle and Vella, 2016), credit default (Alfaro and Gallardo, 2012), the impact of interest rate ceilings (Madeira, 2019a) and changes in real estate prices (Sagner, 2009).

However, in Chile —as in other countries such as Italy, the UK, Spain, Netherlands, and the US—most of the complex financial relationships are concentrated in a small number of upper-income households (Kennickell and Woodburn, 1997; Bover, 2004). The need to cover this small number of households requires the EFH to cover more upper-income households based on a complex sample design. This work examines the challenge of making the EFH sample representative of the 3.85 million Chilean households at the urban national level. I explain briefly how these procedures are built and their importance for data analysis. Finally, I evaluate a set of several "statistical population" methods using the EFH 2007. The method suggested is then applied in a similar way to the EFH 2008, 2009, 2010 and 2011 waves (Central Bank of Chile, 2013). For the more recent EFH waves of 2014 and 2017, a different procedure was applied based on new sample sources (Central Bank of Chile, 2015).

* *I would like to express my gratitude to Sandra Quijada, Jaime Ruiz-Tagle, Rodrigo Alfaro, Natalia Gallardo, and Rodrigo Cifuentes for several useful discussions on the EFH survey, in particular to Sandra for her support with the EFH data and procedures. I also thank the editor Gonzalo Castex and an anonymous referee for their contribution in revising this article.*
** *Gerencia de Investigación Financiera, Banco Central de Chile. E-mail: cmadeira@bcentral.cl*

This paper is organized as follows. Section II explains briefly the most general methods of computing expansion factors for survey datasets and how these can be used to obtain representative statistics and regressions. In section III, I consider a set of post-stratification procedures based on three types of household characteristics: i) their regional area of residence, ii) the income distribution of their urban area (or sub-area) of residence, and iii) the household's income stratum at the Chilean national level. This large set of procedures are then estimated using data from the Chilean population survey (Casen) in 2003 and 2006. When applied to the survey, data expansion factors can be used to compute population weights and obtain representative statistics of any variable at the Chilean national level. Section IV shows how these expansion factors when applied to the EFH 2007 sample allow us to obtain accurate statistics for the income distribution and house ownership among Chilean households. I compare several types of expansion factors in terms of how well they reproduce Chilean national statistics from other datasets, such as age and education. The expansion factors that represent both more accurate statistics and less variance are the ones based on simple information, such as the broad regional area of residence and the income placement of the urban area of residence in terms of the Chilean national population. Finally, section V summarizes the conclusions.

II. GENERAL METHODS TO COMPUTE EXPANSION FACTORS AND POPULATION WEIGHTS

1. What are expansion factors?

Each household in a survey represents a different number of statistically equivalent households. Survey researchers call this the statistical representativeness or "expansion factor" of each observation. For example, if the survey directors go to city A and interview one household in every 1,000, then the expansion factor of city A households should be 1,000. The sum of the expansion factors for all observations in the sample dataset should be equivalent to the target population universe of the survey. If for instance the target population includes all the households of Chile, then the sum of the expansion factors for all observations in the dataset should be equivalent to the household population of Chile. If surveys select respondents with unequal probabilities or in a non-random way, then ignoring the expansion factors implies the estimated statistics cannot be understood as representative of the target population (Neyman, 1934). Suppose A and B are cities of equal size, but the interview rate of A households is only 1 in 1,000 while in B it is 1 in 500. There are 30% of unemployed A people and 15% of unemployed B people. Ignoring population weights gives an unemployment rate of 20% (1/3×30%+ 2/3×15%= 20%). This statistic however is valid only for the set of interviewed people, while for the target population the real statistic is 22.5% (1000/(1000+2×500)×30%+ (2×500)/(1000+2×500)×15% = 22.5%). The "target population" universe is often unknown, due to a lack of an exhaustive listing of all Chilean households. Therefore the size of the target population and its characteristics is usually measured by using another dataset, often a census or other population survey with a larger sample and coverage,

thus making it a reliable approximation of the target population. In the case of Chile there is a Chilean Household Census, which is published every 10 years for the entire population of 3.5 million households, and the National Socioeconomic Characterization (Casen) survey, which is implemented every two years with a sample size ranging from 40 to 80 thousand households. The Casen survey has the advantages of measuring more variables than the Census (such as income) and of being implemented more frequently. The Casen survey also has high response rates (90%), which limits errors from sample bias. For these reasons, the Casen datasets were judged to be the most suitable source of information to estimate the statistical representativeness of the EFH.

2. The inverse selection probability method

There are two main classes of methods to compute expansion factors. The first family is known as the inverse selection design probability method (Lohr, 2009). If every observation in the universe is selected with positive probability, this method is an accurate way of making the smaller sample a statistical representation of the target population. Therefore if observation i in the sample was selected with probability p_i, then the expansion factor f_i is:

$$f_i = p_i^{-1}. \tag{1}$$

A standard example of this method is a random sample where all n sample elements are chosen with equal probability from a population of size N and therefore $fi=N/n$. Some surveys design their samples so as to minimize the estimation variance of a certain statistic (Lavallée and Hidiroglou, 1988; Hedlin, 2000; Rivest, 2002; Kadane, 2005; Horgan, 2006; Fabrizi and Trivisano, 2007). It is important to note that the use of the correct expansion factors still results in consistent estimates for statistics of any other variable in the survey (Neyman, 1934; Kish, 1992); this result requires that the sample selection design be based on exogenous variables, which can be invalid in some cases such as choice-based sampling where observations are selected for example according to whether people took a loan or decided to take part of a clinical trial or suffered from a selective disease such as cancer; therefore this raises issues of optimality and efficiency in the estimates of other statistics, but no problems of asymptotic consistency.

The inverse probability method is usually applied when stratified sample probabilities are carefully chosen before the survey and non-response rates are low. For several surveys, however, non-response rates are high and this may lead to ill-defined statistics. Suppose no households in urban area X were interviewed. Then the selection probability of these households post-survey is 0 and that corresponds to a factor of infinity. Sample design probabilities will give too much importance to a few observations and increase substantially the variance of the statistical methods applied to the survey sample (Kish, 1992).

For these reasons researchers often use alternative methods to estimate the expansion factors for a survey dataset (Kish, 1992). The ideal alternatives seek

to approximate the true inverse selection probabilities by taking into account the most important elements of household heterogeneity, while reducing the effects of sample selection variables less relevant for the outcomes (Kish and Frankel, 1974).

3. Post-stratification methods

The second class of methods are post-stratification procedures (Kish and Frankel, 1974; Lohr, 2009): 1) all the observations in both the sample and the target population are classified in several groups (or strata) according to their characteristics (all the strata are represented in both the sample and target population); 2) the expansion factor of each group in the sample is chosen in a way to match exactly the number of people of that group in the target population. Assume the post-stratified method classifies observations into groups $g = 1,\ldots,G$ and all groups are represented in both the target population and the survey sample. Then the expansion factor for each observation $i \in g(i)$ is:

$$f_i = N_{g(i)} / n_{g(i)} , \tag{2}$$

where $N_{g(i)}$ is the size of group $g(i)$ in the target population and $n_{g(i)}$ is the size of the group $g(i)$ in the survey sample.

If the groups in the post-stratification method include all the relevant variables that affect the selection probability of the sample units, then this method is equivalent to the inverse selection probability method (Kott, 2006a; Lohr, 2009). A good post-stratification method therefore is a reliable approximation for the sample selection probability. For this reason, the choice of groups and variables for the post-stratification procedure should be driven by an explicit theory of sample selection based on the survey design and non-response (Groves and Couper, 1995).

4. Differences between post-stratification methods

Post-stratification methods are usually classified as: 1) full post-stratification methods (Lohr, 2009); or, 2) incomplete post-stratification methods and generalized calibration raking procedures. Full post-stratification methods work by matching the population of each strata in the survey sample with their totals from other surveys (equation (2)). Partial post-stratification is obtained by multiplying the expansion factors that match different variables. Suppose g_h=1,…,G_h represents the strata membership of observation i in terms of variable h. Then the expansion factors of a partial post-stratification procedure using independent variables $h = 1,\ldots,K$ would be:

$$f_i = \prod_{h=1}^{K} N_{g,h(i)} / n_{g,h(i)}. \tag{3}$$

Full post-stratification requires creating mutually exclusive groups for each combination of all the K variables, i.e., $g^*=(g_1,\ldots,g_K)$, and then applying equation 2) as $f_i = N_{g^*(i)} / n_{g^*(i)}$. Partial post-stratification is an easier and simpler approach

than full-post-stratification. However, both methods are only equivalent if all the K variables are independent of each other. If the variables are not independent of each other, then partial post-stratification can create both bias and high variance in the expansion factors (Kozak and Verma, 2006). Therefore full post-stratification is always preferable when feasible.

Generalized raking procedures are a more complex family of post-stratification procedures than the ones described in equation (3). These procedures work by minimizing a distance function of several sample statistics with known statistics of the target population. Such procedures as the Calmar macro are now widely used by national statistics offices in France, Canada, United Kingdom, Italy, Luxembourg, and Spain, among several other countries (Bover, 2004; Crockett, 2008). Calibration methods can be used to create matches with statistics that are known for the whole population, but that are not necessarily specified for each strata in the sample. This is the case because not all surveys collect the same information. For example, Census data elicits only demographic information, while Employment Surveys measure income and occupation information. For this reason, several surveys first create expansion factors using full post-stratification based on Census data and then make partial post-stratification adjustments for other statistics (Deville et al., 1993). In this case, it is important that the first stage of full post-stratification result in expansion factors with a low variance, since the second-stage process of matching more statistics usually increases the variance of the expansion factors even more (Kott, 2006b).

Another problem is how many groups or variables to choose for post-stratification (Wu, 2002; Lohr, 2009). In general, there is a bias versus variance trade-off. Adjusting for more strata can make the expansion factors more volatile and therefore increase the variance of the sample estimators (Kish and Frankel, 1974). For this reason the researcher should choose to match only the most relevant statistics for his survey (Kott, 2006a). Strata should be chosen in a way that individuals inside a group are as homogeneous as possible and new strata should only be added if the individuals belonging to those strata are significantly different from the other strata (Lohr, 2009). Little and Vartivarian (2005) show that if only the relevant variables are included for the post-stratification adjustment, then it is possible to reduce both the bias and the variance of the estimators.

5. Consistency of statistics and regression models

Mean statistics of any variable x can be obtained with standardized expansion factors or population weights $w_i = (f_i / \sum_j f_j)$ (Lohr, 2009): $\text{mean}(x) = \sum_i xi(f_i / \sum_j f_j)$, with f_i denoting the expansion factor of each observation i and $\sum_j f_j$ being the sum of the expansion factors for the whole sample (which should be equivalent to the total represented population). Also, sample mean statistics of x can be computed for a subgroup G of individuals: $\text{mean}(x_G) = \sum_{i \in G} x_i (f_i / \sum_{j \in G} f_j)$. Quantile statistics and other continuous statistics can also be computed in a similar way for the whole population of individuals or any sub-group by using the population weights w_i of each observation i.

Standard-errors and *t*-statistics for estimators using expansion factors are often difficult to derive or involve long expressions (Kott, 2006a; Lohr, 2009). However, if bootstrap is a method that is valid for the specified econometric estimator, then bootstrap is also guaranteed to work with expansion factors (Funaoka et al., 2006). Therefore bootstrap is a safe option for the researcher in most applications (Rao and Wu, 1988). The researcher, however, can account for model and sample uncertainty of the expansion factors when estimating his/her model. This is easily achieved by replicating different expansion factors for each bootstrap sample and obtaining model estimates for each bootstrap draw of the expansion factors (Brownstone and Chu, 1994; Brownstone, 1997).

Unbiasedness of the weighted estimates of a model is not feasible, since weighted estimators involve a multiple of two random variables (the random observations and the weights). However, weighted estimates are approximately unbiased, presenting a bias of order N-1, negligible in large samples (Lohr, 2009). It is possible to estimate regression models without expansion factors and get consistent estimates of the coefficients. However, this result is only valid if the sample selection design is ignorable (exogenous to both the covariates and the unobservable error term). Also, several statistical studies (Nathan and Smith, 1989; Kott, 1991; Särndal et al., 1992) show that weighted estimates are more robust to model misspecification, omitted variable problems and heteroskedasticity. Sample selection design makes nonlinear estimators such as Maximum Likelihood models inefficient. Therefore Maximum Likelihood model variances should be obtained using the Huber-White robust variance matrix.

Even if the regression model is taken to be the real world model, the researcher will only be getting valid estimates for $E(Y|X)$. If the researcher wants to obtain inferences for the whole population, then expansion factors are needed to get valid estimates of $Pr(X)$ and therefore estimate $E(Y)=\int E(Y|X)Pr(X)\partial X$. This requirement of expansion factors is a big concern for welfare analysis.

III. COMPUTING EXPANSION FACTORS FOR THE EFH SAMPLE

1. Classification of urban areas according to population wealth

In order to select a bigger sample of wealthier households, the EFH 2007 survey used two distinct samples (*Centro de Microdatos*, 2008): 1) a sample of 691 households of high income whose addresses were given by the Chilean equivalent of the Income Revenue Service authority (SII); 2) a sample of 3,330 households selected from several urban areas classified in the Chilean Census of Population and Homes, which has a stronger sampling of high-income urban areas. This selection process is similar to the ones implemented in the Survey of Consumer Finances of the US (Kennickell and Woodburn, 1997) and Spain's Family Finance Survey (Bover, 2004).

The 691 households of the SII sample were selected with basis on their reported 2006 taxable income. The distribution of the SII sample was implemented with a

stratification based on the first nine deciles and the top 10 percentiles of taxable income. For the SII sample, the survey provider *Microdatos* built expansion factors by using partial post-stratification in relation to two variables, regions (classified into four groups denoted as "Zones") and three wealth strata (deciles 1 to 5, 6 to 8, and 9 to 10).

The 3,330 Census households were selected from a sample with a high representation of rich urban areas. Urban areas in Chile are classified as counties (known by their Spanish term, "comunas") and smaller sub-areas inside each county are the segments, or "segmentos". The largest counties in Chile were immediately chosen as part of the population from which the EFH household addresses were to be sampled. Then a large set of smaller counties were randomly sampled to be representative of the rest of Chile. The random sampling of smaller counties was necessary, because with a 4,000-household sample it is not possible for the EFH survey to efficiently cover all the 346 counties in Chile. Using 3,764 segments of the Casen 2003, *Microdatos* classified each segment into three wealth types[1]: 1) areas having at least 75% of households of deciles 9 and 10 are classified as type 3; areas are classified as type 2 if at least 75% of their households are of deciles 6 and above; and the remaining areas are classified as type 1. By choosing a higher number of urban areas from types 2 and 3, the EFH was able to over-sample higher-income households. In the final sample of households interviewed by the EFH there are 694 segments: 259 of type 3, 194 of type 2, 225 of type 1, and 16 new segments with no wealth classification. The 16 segments without a wealth classification represent new urban areas created after 2003. After selecting the counties and segments, a given number of households was selected in each segment. The sampling probability of each household i in segment j of county m can be summarized as the multiple of the probability of the county selection ($f_{1,m(i)}$), segment selection ($f_{2,j(i)}$), and the selection of the household in its segment ($f_{3,i}$): $f_{\text{total},i} = f_{1,m(i)} f_{2,j(i)} f_{3,i}$ [2].

The sample design expansion factors were then adjusted for sample non-response across different segments. Finally, the SII and Census samples were jointly combined by multiplying the expansion factors with their respective sample proportions for each wealth strata (types 1,2,3). I will now show that substantial improvements can be made in relation to these expansion factors

---

1 The Casen 2003 was only used to randomly select segments and counties. There was no sampling of the households interviewed in the Casen 2003 survey, therefore it is not possible to link households across both samples.

2 In the Final Report of Microdatos (2008) these probabilities are explained in greater detail. Let h(m) be the wealth type of each county. Let M(k), M(m), and M(j) be the number of households in each region k, county m, and segment j, respectively. Let M(h) be the total number of households of type h = 1; 2; 3 collected by the EFH survey. c(h,k) denotes the number of towns of type h=1,2,3 selected in each region k. The number n(j,m) denotes the number of urban areas inside county m selected for the survey. Finally, g(i,j) denotes the number of households that will be selected for survey inside each urban area j. Note that that some counties are auto-selected (meaning selected with 100% probability), while other counties are randomly sampled among a set of other similar towns. Among the randomly sampled towns in each region, each town m of type h=1,2,3 is selected with probability $f_{1,m}$=c(h(m),k)M(m)/M(k(m)). Then segments inside each county are selected with probability $f_{2,j}$ = n(j,m(j)) M(j)/M(m(j)). Finally, each household i in segment j is selected with probability $f_{3,i}$ = g(i,j(i))/M(j(i)).

by using post-stratification procedures. The reason is that segments in Chile have very heterogeneous population levels; therefore in the final survey sample there were many households with similar characteristics but with very different expansion factors due to their residence in different segments, which increases the variance of any estimator using inverse probability expansion factors (Kish, 1992). Also, since the population of smaller areas such as segments could have a larger measurement error than the estimates for bigger areas such as counties, this implies that expansion factors based on segments may have substantial measurement errors. Considerations of excessive variance and measurement error thus imply that an inverse selection probability expansion factor is unlikely to be optimal for the EFH survey.

2. Linking the Casen and EFH households in different types

To make the EFH representative of Chile we must classify all its households in a similar set of strata as other larger surveys of Chile. The Casen is the most ideal data to build expansion factors for the EFH, since they were used in its sample design and have the same measures of household income. However, these surveys represent the distribution of urban households of Chile in the past. For this reason I update the expansion factors of these surveys to reflect demographic growth between those years and the EFH sample years. Therefore I used the national percentiles of the income distribution of the urban Casen 2006 data and then I updated each percentile for the nominal income growth of each decile between 2006 and 2007 using the Supplementary Income Survey (ESI), which covers over 30 thousand households in the fourth quarter of each year. The EFH households were then classified according to this updated measure of the income percentiles in 2007. The Casen surveys include fewer counties than the total of 346 counties of Chile, so I take into account the non-represented counties by assigning their population to the counties of the same region in the same proportion as their population size. Then I compute the ratio of population of each county in the Casen surveys with their projected population by the Chilean National Statistics Institute (INE) for the EFH sample years, ratio INE-Casen(i) = population INE(i)/population Casen(i).

Since the EFH does not include all the urban areas of the Casen, we must consider that each urban area is representative of a wider population than just itself, including assigning all the non-represented urban areas (counties and segments) in the Casen survey to areas of similar income represented in the EFH, according to their income type (1, 2, 3). For the 16 segments of the EFH not present in the Casen 2003, I use their own sample observations to assign them a type 3, 2, or 1, classification. The expansion factor for these 16 segments will be the same one as for the segments of the same type in the same county. Let $i = 1,..,K$ denote each population group (denoted by segment number and the household strata). Then I obtain a list of all the common groups in the EFH and Casen. The population of each group i represented in the EFH is updated as:

$$population(i)^{*} = population(i) + \frac{population(i)}{\sum_{k=1,k\in S,k\in G(i)}^{K} population(k)} x \sum_{k=1,k\neq S,k\in G(i)}^{K} population(k), \quad (4)$$

where S is the set of groups in the survey. $G(i)$ is the set of groups in the same region with the same type of segment (i.e., 12,3) and household strata (i.e., 1,2,3) as group i. Intuitively, expression (4) implies that the final value of each strata's population, $population(i)^{*}$, represents its direct population value ($population(i)$) plus a component of how much the strata represents the proportion of other strata in the population which did not enter the survey sample.

3. Creating new expansion factors

Since it is not obvious how many strata to use in the EFH, I created a set of different expansion factors and tested which ones made a better fit to the income distribution of Chile. The set of expansion factors is detailed as follows. Each expansion factor is labelled with four digits, expr****. Appendix A summarizes the content and modeling options that were applied to construct each alternative expansion factor.

The first digit classifies the "smoothness" of the factor estimates: it has value 0 if the expansion factor uses the true strata population or 1 for a "smoothed" estimate. The "smoothed" expansion factors expr1*** are obtained by using the mean predicted values of a regression of expr0***. I tried two regression options. The first one was a kernel regression[3] using the median and interquartile range for the household total income of the county and segment of each observation. The second option was a linear regression using the mean years of education, mean number of persons per household, and the 25%, 50% and 75% household income quantiles of their county and segment. Both options yielded similar results, therefore only the second option is reported.

The second digit indicates the level of aggregation of the geographical areas of residence: 6 or "Zones 2" differentiates for two aggregate groups of Chilean regions (*Metropolitan Region* and Other Regions of Chile); 5 or "Zone" differentiates for four groups of Chilean regions (*Northern Region* includes Chilean regions 1 to 4 and 15, *Central Region* includes regions 5 to 8, *Southern Region* includes regions 9 to 12 and 14, and *Metropolitan Region* includes region 13); 4 or Region differentiates for the 15 Chilean regions; 3 or "Provincia" differentiates for each province of Chile; 2 or "County" differentiates for each county of Chile; and, finally, 1 or "Segment" differentiates for each segment of Chile.

The third digit indicates whether the strata consider different types of counties and segments: 0 does not differentiate for counties and segments of different

3 Here I apply the Nadaraya-Watson kernel smoothing method, using a Epanechnikov density function and Silverman's rule of thumb bandwidth (Manski, 1990). Da Silva and Opsomer (2009) applied a similar method to compute alternative expansion factors for the National Health and Nutrition Examination Survey dataset, finding it to be effective in reducing the bias and variance of non-response adjustments to the expansion factors.

wealth types; 1 differentiates for the three different types of segment wealth; 2 differentiates for the three different types of counties; and 3 - differentiates for the 9 different types of wealth of both the county (3 types) and segment (3 types) of residence of the household.

Finally, the fourth digit indicates strata of household income: 0 indicates no differentiation across households in the same area; 1 classifies households according to three income brackets measured by their percentile at the national level (1 to 50, 51-80, 81-100); and 2 classifies households into 18 types according to the national income percentiles (1-35, 36-50, 51-60, 61-68, 69-75, 76-80, 81-85, 86-88, 89-90, 91-92, 93-94, 95, 96, 97, 98, 99, 99.5, and 100).

This classification method results in a "phonebook list" of expansion factors. So for instance, an expansion factor based only on segments would be expr0100, while an expansion factor based on counties and types of segments would be expr0210. A factor based on aggregate zones and types of counties and segments would be expr0530. The expansion factor expr0100 is equivalent to the inverse probability of selection of the unit (IPSU), since it uses the information on the population of all the urban sampling units in the EFH and how they were selected from the Casen urban areas.

The SII sample is highly biased towards high-income households and requires therefore a different statistical treatment from the rest of the sample. Due to the small sample size I decided to use a stratification based on only two aggregate regions ("Zones 2": Metropolitan Region, Other Regions of Chile), but with several income strata (national income percentiles 1-35, 36-50, 51-60, 61-68, 69-75, 76-80, 81-85, 86-88, 89-90, 91-92, 93-94, 95, 96, 97, 98, 99, 99.5, and 100).

The Census and SII samples were then combined in the EFH 2007 to form a joint sample of 4,021 households; therefore, the final expansion factor reflects the proportion each sample has on the dataset:

$$\text{factor}_h(k)^* = \text{factor}_h(k)\ n_h(k)/n(k), \tag{5}$$

where $h \in$ (*SII, CENSUS*) denotes the sample origin of the observation and $k \in$ (*Metropolitan Region*, Other Regions) denotes the geographical area of the observation. $n_h(k)$represents the number of households that sample h has in area k, while $n(k) = n_{SII}(k) + n_{CENSUS}(k)$ denotes the total number of households in the EFH present in area k[4].

4 This adjustment is necessary in order that the average household of both samples represents the same number of Chilean households. The adjustment requires different proportions for the Metropolitan Region and the Other Regions in order to keep the representativity both at the national level and at the level of the Metropolitan Region. This is important, because the EFH waves in 2008 and 2009 were only implemented in the Metropolitan Region. Therefore this adjustment allows for comparability across different EFH waves.

IV. EVALUATING THE EXPANSION FACTORS

Centro de Microdatos provided a set of provisory expansion factors for the EFH survey when information from the Casen 2003 and 2006 were not yet publicly available. It is important to note that the provisory *Microdatos* procedure suffers from three biases. One, in the Census sample it is not taken into account that the richest counties in Chile were chosen to be part of the EFH and therefore that the Chile outside the EFH sample design is poorer than the one that included all. Second, the SII sample is based on only three wealth strata, when the sample selected was highly based on individuals of the highest income percentiles in Chile. Third, it uses partial post-stratification on both the Census and SII samples, so it does not take into account that strata are not independent and that counties are correlated with income. One can easily perceive that all these flaws create a bias in the *Microdatos* sample towards over-counting the number of wealthy households in Chile. These provisory weights create highly biased estimates of the education and income distribution in Chile, with a mean absolute bias for income of 28.3%.

To evaluate the performance of the expansion factors and their variance/dispersion, 2 shows the minimum, maximum, mean, and standard-deviation of each expansion factor option. I also compute their correlation coefficient in relation to the factor expr0602. The factor expr0602 is the one that takes into account the largest heterogeneity in income strata; therefore, the correlation coefficient gives a rough measure of whether the dispersion in each expansion factor is significant for explaining household income.

Table 2 shows that indeed expansion factors based on segments, counties, and even provinces, have a substantial variance, with factors based on segments implying that some households have a statistical representation for just 10 or 15 households, while others represent 15 or 20 thousand. Choosing expansion factors based on segments and counties results in standard-deviations larger than the mean. Creating expansion factors based on aggregate regions of types 5 and 6 instead of segments successfully reduces this standard deviation from above 1,300 to less than 900 and reduces the minimum-maximum range from 15-15,000 to around 50-5,000. However, perhaps the biggest problem for the factor types based on segments, counties and provinces is that their correlation with the income strata population (given by expr0602) is low. By using aggregate regions (such as type 5 or 6) instead of segments one can increase the correlation with the income strata from 40% to above 60%.

To compare how well each expansion factor estimates Chile's income distribution, I compute the average absolute deviations from the estimates of the income distribution of each factor type in relation to the income distribution obtained with factor 0602:

$$\textit{absolute deviation}_b(\text{expr****}) = \frac{1}{K(b)} \sum_{s=1}^{K(b)} |\log(\text{stat}_{\text{expr****}}(s)) - \log(\text{stat}_{0602}(s))| \quad (6)$$

where $b \in$ (*percentiles, deciles, quintiles*), stat(i) $\in$ (*percentile, decile, quintile*) and $K(b) \in (99, 9, 4)$. The top percentile is excluded from the mean absolute deviation statistic, because income is potentially unbounded for the top percentile. Appendix B shows the mean values of absolute $deviation_b$(expr****). Again, it is quite clear that expansion factors based on more aggregate areas (such as 0431, 531, and 0631) are more efficient in estimating the income distribution of Chile. Therefore, the new work improves income estimates and reduces factor variances by accounting for both household type (1,2,3) and segment type (1,2,3). Overall, the expansion factor 0531 shows a good balance between reducing factor variance and providing a good fit of the income distribution of Chile.

1. Monte Carlo performance of the expansion factors

To confirm that the results of appendix B are not the outcome of a lucky one-time draw, I make a Monte Carlo experiment by creating 2,500 bootstrap samples of 3,301 households from the Casen 2003, using the same segments as the EFH. Then I apply the alternative expansion factors, obtaining statistics for the absolute deviations of the percentiles and deciles of each factor in relation to the national urban Casen 2003 income distribution. This exercise does not depend on counterfactual estimations of the income distribution, since the original population is known and fixed.

Table C1 with the sample statistics of the Monte Carlo exercise shows that the mean bias of the provisional CMD weights is large, being around 10.8%, and the worst bias among the available options. Again, expansion factors based on aggregate regions (such as types 5 and 6) work fairly well. In particular, the option expr0531 appears to be quite robust across all Monte Carlo samples and it is also one of the alternatives with lowest estimated standard-error among all the 2500 Monte Carlo samples. I also show the dispersion statistics of the performance of each factor type in tables C2 and C3. The expansion factor expr0531 performs quite well and its top percentile of mean absolute error is still quite below the lowest percentile of mean absolute error for the provisional CMD factors. Therefore, even in the worst 1% scenario, the expansion factor expr0531 would still estimate the income percentiles of Chile with a mean absolute error of only 5.2%. Additional results for exercises based on quintiles of the income distribution are available in the original working paper version of this article (Madeira, 2011).

2. Demographic representation of the EFH

Finally, in table D1, I compare education, age and unemployment statistics of the EFH 2007, using both the provisional expansion factors (CMD) and the new alternative (0531), with those from the Casen 2009 survey and the official National Employment Survey (ENE). The provisional expansion factors based on partial post-stratification underestimate the unemployment rate at the national level and capital region of Santiago, while overestimating the number of people with a post-graduate degree at the national level by 40% and the number of people with college degrees by 17%. For the age distribution the differences between alternatives are smaller, since age across different urban counties does not differ much.

In summary, the new full post-stratification expansion factors (as represented by the alternative 0531) are effective in portraying age, education and unemployment (table 4), the income distribution (appendix B), and reduce the error and variance of statistical estimates in the Monte Carlo exercise (tables C1, C2 and C3).

3. Results for the EFH waves of 2008, 2009, 2010 and 2011/12

This sub-section shows some results of how the alternative expansion factors 0531 worked for the other EFH waves after 2007. The EFH 2007 wave had 4,021 households, of which 82% came from a geographical classification of the urban areas of the Census/Casen 2003 and 18% came from the SII tax records concentrated on the top 10 income percentiles. Due to breaks in the format of the sample sources, the EFH survey changed survey sampling afterwards. In 2008 and 2009 the survey sample was a subset of the households interviewed for the Greater Santiago Employment and Unemployment Survey (EOD) of the University of Chile, with households being classified in terms of three income strata (percentiles 1-50, 51-80, 81-100) and 8 types of counties. In 2010 the sample was from the Chilean SII's list of residential properties. The waves of 2008, 2009 and 2010 were implemented only in the Metropolitan Region and interviewed a final sample of 1,150, 1,190 and 2,037 households each year, respectively. The EFH 2011-12 was again implemented at the national level with 4,059 households, with the sample based on a more detailed characterization of the SII's list of home properties. For the EFH 2011 sample, counties and urban blocks inside each county were classified in different types according to their percentile of median home value, with 3 types of county (percentiles 1-50, 51-80, 81-100) and 7 types of urban blocks (percentiles 1-30, 31-50, 51-80, 81-90, 91-94, 95-98, 99-100). This SII home list of properties includes a total of around 4.25 million residential homes divided across 95,000 urban blocks at the national level, providing a very good coverage of households. Eighty counties and 725 blocks were selected at the national level, with both an original sample and five replacement samples in reserve in case of unit non-response. These EFH waves had a panel sample component of 1,792 households between 2007 and 2011-12 and of 947 households between 2008 and 2009.

For all these EFH waves (2007, 2008, 2009, 2010 and 2011-12) I applied a full post-stratified version of the expansion factor 0531, modifying it to include as additional strata more income types (percentiles 1-30, 31-50, 51-65, 66-80, 81-90, 91-94, 95-98, 99-100), the homeownership status (0 or 1 if owner of the household home) and the mortgage status (0 or 1 if still paying for the main home with a mortgage loan) of the households. Each year the expansion factors were estimated based on weighted averages of the Casen waves of 2006, 2009 and 2011, which comprise over 50,000 households per wave. Furthermore, changes to the values of the population of each county were applied based on the estimates of the Chilean National Statistics Institute (INE) and the national income percentiles were updated based on the nominal income growth per decile available from consecutive waves of the National Income and Employment Survey (ESI, which covers over 30,000 households during the fourth quarter of each year).

Table D2 summarizes the results of the full post-stratified expansion factors method based on the strata for four regional areas, county and block type, household income percentile strata, plus main home ownership and mortgage status. The results show that all the EFH waves have a similar dispersion for their expansion factors, with a minimum and maximum values that are within a ratio of 10 times the median value. Finally, in table D3, I show that using these expansion factors on every EFH wave achieves an appropriate evolution for the different percentiles of household income, unemployment, mortgage and financial asset ownership for the period from 2007 until 2011. Therefore, the chosen expansion factors adequately achieve the goal of showing comparable survey statistics for all the EFH waves.

V. CONCLUSION

This work estimates the expansion factors for the EFH survey, using several full post-stratification options based on the geographical area and the income category of each household. Based on the 2007 wave, a Monte Carlo exercise shows that factors based on aggregate regions of Chile have less bias and variance than those with higher geographic detail such as counties or segments. Since the option 0531 appears to have lower bias and standard-error in a Monte Carlo exercise with 2,500 Monte Carlo simulations, this expansion factor was included in the official EFH 2007 dataset. Furthermore, this expansion factor option was also the basis for the EFH survey waves of 2008, 2009, 2010 and 2011-12.

REFERENCES

Alfaro, R. and N. Gallardo (2012). "The Determinants of Household Debt Default." *Revista de Analisis Económico - Economic Analysis Review* 27(1): 55–70.

Bover, O. (2004). "The Spanish Survey of Household Finances (EFF): Description and Methods of the 2002 Wave." Occasional Paper No. 0409, Banco de España.

Brownstone, D. and X. Chu (1994). "Multiply-Imputed Sampling Weights for Consistent Inference with Panel Attrition." Transportation Center Working Paper 590, University of California, Irvine.

Brownstone, D. (1997). "Multiple imputation methodology for missing data, non-random response and panel attrition." Institute of Transportation Studies Working Paper 97-4, University California Irvine.

Central Bank of Chile (2009). *Encuesta Financiera de Hogares: Metodología y Principales Resultados EFH 2007*.

Central Bank of Chile (2013). *Encuesta Financiera de Hogares: Metodología y Principales Resultados EFH 2011-12*.

Central Bank of Chile (2015). *Encuesta Financiera de Hogares 2014: Principales Resultados*. Central Bank of Chile.

Centro de Microdatos (2008). *Informe Final Encuesta Financiera de Hogares*. Universidad de Chile.

Crockett, A. (2008). "Weighting the Social Surveys." Economic and Social Data Service (ESDS) Government, UK.

Da Silva, D. and J. Opsomer (2009). "Nonparametric Propensity Weighting for Survey Nonresponse through Local Polynomial Regression." *Survey Methodology* 35(2): 165–76.

Deville, J.C., C.E. Sarndal and O. Sautory (1993). "Generalized Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 88(423): 1013–20.

Fabrizi, E. and C. Trivisano (2007). "Efficient Stratification Based on Nonparametric Regression Methods." *Journal of Official Statistics* 23: 35–50.

Fuenzalida, M. and J. Ruiz-Tagle (2009). "Riesgo Financiero de los Hogares." *Economía Chilena* 12(2): 35–53.

Funaoka, F., H. Saigo, R. Sitter, and T. Toida (2006). "Bernoulli Bootstrap for Stratified Multistage Sampling." *Survey Methodology* 32(2): 151–6.

Groves, R. and M. Couper (1995). "Theoretical Motivation for Post-Survey Non-Response Adjustment in Household Surveys." *Journal of Official Statistics* 11(1): 93–106.

Hedlin, D. (2000). "A Procedure for Stratification by an Extended Ekman Rule." *Journal of Official Statistics* 16(1): 15-29.

Horgan, J. (2006). "Stratification of Skewed Populations: A Review." *International Statistical Review* 74: 67–76.

Kadane, J. (2005). "Optimal Dynamic Sample Allocation Among Strata." *Journal of Official Statistics* 21: 531–41.

Kennickell, A. and R. Woodburn (1997). "Consistent Weight Design for the 1989, 1992 and 1995 SCFs, and the Distribution of Wealth." Mimeo, Federal Reserve Board.

Kish, L. and M. Frankel (1974). "Inference from Complex Samples." *Journal of the Royal Statistical Society* Series B (Methodological) 36(1): 1–37.

Kish, L. (1992). "Weighting for Unequal Pi." *Journal of Official Statistics* 8: 183–200.

Kott, P. (1991): "A Model-Based Look at Linear Regression with Survey Data." *The American Statistician* 45: 107–112.

Kott, P. (2006a). "Sample Survey Theory and Methods: A Correspondence Course." Mimeo, National Agricultural Statistics Service (USDA).

Kott, P. (2006b). "Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors." *Survey Methodology* 32(2): 133–42.

Kozak, M. and M. Verma (2006). "Geometric Versus Optimization Approach to Stratification: A Comparison of Efficiency." *Survey Methodology*, 32 (2), 157-163.

Lavallée, P. and M.A. Hidiroglou (1988). "On the Stratification of Skewed Populations." *Survey Methodology* 14: 33–43.

Little, R. and S. Vartivarian (2005). "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology* 31(2): 161–8.

Lohr, S. (2009). "Sampling: Design and analysis." Pacific Grove, CA: Duxbury Press.

Madeira, C. (2011). "Computing Population Weights for the EFH Survey." Working Paper No. 632, Central Bank of Chile.

Madeira, C. and V. Pérez (2013). "Gestión Hipotecaria de las Familias Chilenas." *Economía Chilena* 16(2): 122–33.

Madeira, C. (2014). "El Impacto del Endeudamiento y Riesgo de Desempleo en la Morosidad de las Familias Chilenas." *Economía Chilena* 17(1): 88–102.

Madeira, C. (2015). "Motivaciones del Endeudamiento en las Familias Chilenas." *Economía Chilena* 18(1): 90–106.

Madeira, C. (2018a). "Priorización de Pago de Deudas de Consumo en Chile: El Caso de Bancos y Casas Comerciales." *Economía Chilena* 21(1): 118–32.

Madeira, C. (2018b). "Explaining the Cyclical Volatility of Consumer Debt Risk Using a Heterogeneous Agents Model: The Case of Chile." *Journal of Financial Stability* 39: 209–20.

Madeira, C. (2019a). "The Impact of Interest Rate Ceilings on Households' Credit Access: Evidence from a 2013 Chilean Legislation." Documento preliminar, Central Bank of Chile.

Madeira, C. (2019b). "Household Stress Testing Using Micro Data: Evidence from Chile." Documento preliminar, Central Bank of Chile.

Manski, C. (1990). "Nonparametric Bounds on Treatment Effects." *American Economic Review Papers and Proceedings* 80(2): 319–23.

Nathan, G. and T. Smith (1989). "The Effect of Selection in Regression Analysis." In *Analysis of Complex Surveys*. John Wiley & Sons.

Neyman, J. (1934). "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97: 558–606.

Rao, J. and C. Wu (1988). "Resampling Inference with Complex Survey Data." *Journal of the American Statistical Association* 83: 231–41.

Rao, J. (2005). "Interplay between Sample Survey Theory and Practice: An Appraisal." *Survey Methodology* 31(2): 117–38.

Rivest, L. (2002). "A Generalization of the Lavallée-Hidiroglou Algorithm for Stratification in Business Surveys." *Survey Methodology* 28: 191–8.

Ruiz-Tagle, J. and F. Vella (2016). "Borrowing Constraints and Credit Demand in a Developing Economy." *Journal of Applied Econometrics* 31(5): 865–91.

Sagner, A. (2009). "Determinantes del Precio de Viviendas en Chile." Working Paper No. 549, Central Bank of Chile.

Särndal, C., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

Wu, C. (2002). "Optimal calibration estimators in survey sampling." Statistics Canada International Symposium Series: Proceedings, Statistics Canada.

APPENDIX A

EXPRESSING EXPANSION FACTORS CATEGORIES IN FOUR DIGITS, expr****

1st digit: Smoothness indicator

Options: 0, unsmoothed population value; 1, Kernel smoothed population based on the mean years of education, household size and quantiles (25, 50, 75) of household income for each county and segment.

2nd digit: Regional strata

Options: 1, Segments; 2, Counties; 3, Provinces; 4, Regions; 5, Zones (North, Central, South, and Metropolitan Region); 6, Metropolitan Region and Other Regions of Chile.

3rd digit: County and urban area strata

Options: 0, does not differentiate types of county and segment; 1-3 types of segment according to household income; 2-3 types of county according to household income; 3-9 types of urban areas according to the type of segment (3 types); and the type of county (3 types) of the residential area.

4th digit: Household income strata

Options: 0, no differentiation for household income type; 1, households of three income types according to national level percentile (1 to 50, 51-80, 81-100); 2, households of 18 types according to the national income percentiles (1-35, 36-50, 51-60, 61-68, 69-75, 76-80, 81-85, 86-88, 89-90, 91-92, 93-94, 95, 96, 97, 98, 99, 99.5, and 100).

APPENDIX B

DISPERSION OF THE EXPANSION FACTORS AND CORRELATION WITH INCOME STRATA

Factor	Mean	Standard deviation	Min.	Max.	Correlation w/ 0602 (%)	Correlation w/ IPSU (%)	Absolute deviation of the income distribution relative to 0602 (%)		
							Percentiles	Deciles	Quintiles
602	968	619	70	2,216	100.0	43.2	0.0	0.0	0.0
CMD	982	1315	17	22,054	23.6	46.8	28.3	27.2	26.5
0100/ IPSU	957	1,251	14	15,316	43.2	100.0	9.1	7.9	8.2
200	957	952	27	5,337	41.7	69.6	22.3	21.5	21.4
201	957	1,068	22	8,137	56.3	71.0	3.5	1.4	1.1
1200	957	776	13	5,337	50.2	57.9	23.6	22.4	22.2
1201	957	776	13	5,337	50.2	57.9	23.6	22.4	22.2
210	957	1,208	26	8,279	46.4	87.2	6.7	5.5	5.7
211	957	1,313	10	10,000	47.0	78.4	4.0	2.3	1.7
310	957	999	27	7,358	54.2	76.5	7.9	7.0	7.3
311	957	1,078	10	9,145	57.0	69.9	4.2	2.8	2.3
410	957	901	27	5,337	59.9	73.6	8.4	7.6	7.9
411	957	968	10	5,366	62.6	69.0	3.8	2.5	1.8
430	957	903	27	5,337	60.0	73.4	8.1	7.3	7.8
431	957	970	10	5,366	62.5	68.8	3.7	2.4	1.8
510	957	844	27	5,337	62.5	69.1	10.1	8.8	9.4
511	957	910	27	5,337	66.4	65.0	4.0	2.0	1.3
520	957	691	27	5,337	53.6	55.4	26.1	24.6	24.0
521	957	771	27	5,337	77.5	58.3	4.0	1.6	0.6
530	957	846	27	5,337	62.7	68.9	9.2	8.3	8.8
531	957	912	27	5,337	66.3	64.8	3.9	1.8	1.2
610	957	835	27	5,337	62.9	68.4	10.0	9.1	9.8
611	957	900	27	5,337	67.1	64.3	4.0	2.0	1.3
630	957	837	27	5,337	63.1	68.2	9.5	8.6	9.4
631	957	902	27	5,337	67.0	64.1	3.9	1.8	1.3

Source: Author's elaboration based on Monte Carlo simulations of 3,301 households of the Casen 2003 (segments selected for the EFH).

APPENDIX C

PERFORMANCE OF THE OPTIONS FOR THE EXPANSION FACTOR IN A MONTE CARLO SIMULATION

Table C1

Statistical deviations of the income distribution of each factor type
(2,500 Monte Carlo simulations)

Factor	Percentiles			Deciles		
	Mean bias (%)	Mean absolute deviation (%)	Standard-error of absolute deviation (%)	Mean bias (%)	Mean absolute deviation (%)	Standard-error of absolute deviation (%)
CMD	10.8	11.6	1.7	11.1	11.4	1.8
100	-6.7	7.6	1.5	-6.9	7.2	1.4
200	21.5	21.6	1.5	22.2	22.2	1.5
201	-1.1	3.4	1.5	-1.3	2.9	1.4
1200	22.2	22.3	1.3	23.0	23.0	1.3
1201	22.3	22.4	1.3	23.1	23.1	1.3
210	-8.9	9.1	2.0	-9.0	9.0	2.0
211	-5.4	5.8	1.4	-5.5	5.6	1.3
310	-6.8	7.2	1.7	-6.9	7.1	1.6
311	-3.5	4.1	1.1	-3.4	3.9	1.0
410	-6.1	6.5	1.5	-6.2	6.4	1.5
411	-2.4	3.3	0.9	-2.3	2.9	0.9
430	-6.4	6.7	1.5	-6.5	6.7	1.5
431	-2.6	3.4	0.9	-2.5	2.9	0.9
510	-6.0	6.4	1.4	-6.1	6.4	1.4
511	-2.2	3.3	0.9	-2.0	2.8	0.8
520	24.5	24.5	1.2	25.1	25.1	1.2
521	0.2	3.9	0.8	-0.1	3.2	0.7
530	-6.3	6.6	1.5	-6.4	6.6	1.5
531	-2.3	3.2	0.9	-2.1	2.8	0.8
610	-5.8	6.3	1.4	-5.9	6.2	1.4
611	-2.1	3.2	0.9	-1.9	2.7	0.8
630	-6.1	6.5	1.5	-6.3	6.4	1.5
631	-2.2	3.2	0.9	-2.0	2.7	0.8

Source: Author's elaboration from Monte Carlo simulations of 3,301 households of the Casen 2003 (segments selected for the EFH).

Table C2

Distribution of the absolute deviations of the income distribution of each factor type
(percentiles from 2,500 Monte Carlo simulations)

	Income distribution measured by percentiles						
Factor	P-1%	P-10%	P-25%	P-50%	P-75%	P-90%	P-99%
CMD	8.0	9.5	10.5	11.5	12.7	13.9	16.1
100	3.4	5.1	6.2	7.6	8.9	10.1	12.2
200	18.2	19.7	20.6	21.7	22.7	23.5	25.2
201	1.6	2.2	2.7	3.4	4.1	4.8	6.1
1200	19.4	20.6	21.4	22.3	23.2	24.1	25.3
1201	19.5	20.7	21.5	22.4	23.3	24.1	25.4
210	4.8	6.5	7.7	9.1	10.4	11.8	14.0
211	2.9	4.1	4.9	5.7	6.7	7.6	9.2
310	3.7	5.1	6.0	7.2	8.3	9.4	11.5
311	1.9	2.7	3.3	4.1	4.8	5.6	7.0
410	3.2	4.6	5.4	6.5	7.5	8.5	10.2
411	1.4	2.2	2.7	3.3	4.0	4.6	5.8
430	3.3	4.7	5.6	6.7	7.7	8.7	10.4
431	1.4	2.2	2.7	3.3	4.0	4.6	5.8
510	3.4	4.6	5.4	6.4	7.4	8.3	9.9
511	1.4	2.1	2.6	3.2	3.8	4.4	5.4
520	21.8	22.9	23.7	24.5	25.4	26.2	27.4
521	2.3	2.9	3.3	3.9	4.4	5.0	5.9
530	3.5	4.8	5.6	6.6	7.6	8.6	10.1
531	1.4	2.1	2.6	3.2	3.8	4.4	5.4
610	3.2	4.5	5.3	6.3	7.2	8.2	9.8
611	1.4	2.1	2.6	3.1	3.8	4.3	5.3
630	3.3	4.6	5.5	6.5	7.4	8.4	10.0
631	1.4	2.1	2.6	3.1	3.7	4.3	5.3

Source: Author's elaboration from Monte Carlo simulations of 3,301 households of the Casen 2003 (segments selected for the EFH).

Table C3

Distribution of the absolute deviations of the income distribution of each factor type
(percentiles from 2,500 Monte Carlo simulations)

	Income distribution measured by deciles						
Factor	**P-1%**	**P-10%**	**P-25%**	**P-50%**	**P-75%**	**P-90%**	**P-99%**
CMD	7.7	9.1	10.2	11.3	12.5	13.8	16.2
100	2.9	4.8	5.9	7.2	8.5	9.7	11.8
200	18.8	20.3	21.2	22.3	23.2	24.1	25.7
201	1.1	1.7	2.2	2.8	3.5	4.2	5.3
1200	20.0	21.3	22.2	23.0	23.9	24.7	25.9
1201	20.1	21.4	22.2	23.1	23.9	24.7	25.9
210	4.6	6.5	7.7	9.1	10.4	11.6	13.9
211	2.8	3.9	4.7	5.5	6.4	7.3	8.7
310	3.5	5.0	6.0	7.1	8.2	9.2	11.0
311	1.6	2.5	3.1	3.8	4.5	5.2	6.4
410	3.1	4.5	5.4	6.4	7.4	8.3	9.9
411	1.0	1.8	2.3	2.9	3.5	4.0	5.1
430	3.2	4.7	5.6	6.7	7.7	8.6	10.2
431	1.0	1.9	2.3	2.9	3.5	4.1	5.1
510	3.2	4.5	5.3	6.3	7.4	8.2	9.7
511	1.0	1.8	2.2	2.8	3.3	3.9	4.8
520	22.3	23.5	24.3	25.0	25.9	26.6	27.9
521	1.7	2.2	2.6	3.1	3.7	4.2	5.1
530	3.4	4.7	5.6	6.6	7.6	8.5	10.1
531	1.0	1.7	2.2	2.8	3.3	3.8	4.8
610	3.1	4.4	5.2	6.2	7.2	8.0	9.6
611	1.0	1.7	2.2	2.7	3.3	3.8	4.8
630	3.2	4.6	5.5	6.4	7.5	8.3	9.9
631	1.0	1.7	2.2	2.7	3.3	3.8	4.8

Source: Author's elaboration from Monte Carlo simulations of 3,301 households of the Casen 2003 (segments selected for the EFH).

APPENDIX D

SUMMARY OF THE SELECTED EFH EXPANSION FACTORS OVER TIME AND COMPARISON WITH STATISTICS FROM OTHER HOUSEHOLD SURVEYS

Table D1

Distribution of persons by age and education in Chile

Education	Casen09	EFH-CMD	EFH-0531	Age	Casen09	EFH-CMD	EFH-0531
Below high school	37.9%	32.3%	37.6%	0/17	27.0%	26.0%	26.5%
Some high school	40.7%	38.1%	39.2%	18/24	13.1%	13.0%	12.2%
Some college	14.3%	19.1%	15.5%	25/34	13.2%	14.0%	12.7%
Post-graduate	7.1%	10.5%	7.7%	35/44	13.5%	13.8%	13.2%
Total	100.0%	100.0%	100.0%	45/54	13.5%	13.7%	14.3%
Unemployment	ENE-Q4	EFH-CMD	EFH-0531	55/64	9.1%	9.3%	9.8%
Chile	7.2%	6.5%	7.3%	65/+	10.5%	10.1%	11.2%
Metropolitan Region	7.1%	5.8%	6.7%	Total	100.0%	100.0%	100.0%

Source: Author's elaboration from the EFH and Casen.

Table D2

Dispersion of the selected expansion factors across EFH survey waves

Factors	EFH 2007	EFH 2008	EFH 2009	EFH 2010	EFH 2011
Minimum	26	181	339	25	146
(Q10)	105	625	633	219	201
(Q25)	276	1021	850	448	334
(median)	412	1382	1333	828	687
(Q75)	1268	1937	2096	1250	1113
(Q90)	1519	2667	2799	1801	1601
Maximum	6663	3784	9844	3895	3691

Source: Author's elaboration from the EFH.

Table D3

Statistics of EFH versus Casen (urban households) - Metropolitan Area

	Casen 06	2007	2008	2009	Casen 09	2010	2011
Income (thousands of pesos)							
(Q25)	280	283	310	302	322	342	350
(median)	480	500	529	521	541	570	603
(Q75)	827	800	858	916	955	965	1,090
(mean)	726	944	814	809	852	838	1,025
Mortgage (% of households)	19.0%	14.5%	16.7%	16.5%	16.5%	17.3%	16.9%
Financial (% of households)		16.3%	10.9%	12.5%			11.3%
Unemployed (% of labor force)		6.7%	12.5%	9.2%		6.7%	6.6%

Source: Author's elaboration from the EFH and Casen.