SMALL AREA ESTIMATION IN TURKEY

Ömer L. Gebizlioðlu, Ankara University and State Institute of Statistics and
Hasibe Dedeþ and A. Ömer Toprak, State Institute of Statistics, Turkey

ABSTRACT

The need for small area estimates is growing fast in Turkish official statistics. This is true especially for some agricultural and household statistics required at sub-regional and provincial levels. The small area estimation experience of the country has developed through the recent crop acreage prediction studies and household surveys. Several methods have been in use to do small area estimation for the surveys mentioned here. The concerned methods tackle with multistage stratified sampling designs applied overall the Country. These methods have been based on the utilization of frames that were updated at the latest population census. It has been necessary to do cross-checking with the census and to incorporate the results of extended sample surveys into the sampling design and estimation efforts. Along these lines, this paper presents the surveys and assessment of the methods adopted for small area estimation.

KEYWORDS

Small Area Estimation, Household Surveys, Area Frame Sampling, Survey Based Estimates, Synthetic and Model Based Estimates

1. INTRODUCTION

The State Institute of Statistics of Turkey (SIS) produces about eighty-five percent of the official statistical information in Turkey. The internally and externally growing need for statistical information have enforced the Institute to enlarge the scope and coverage of its activities remarkably in the last decade. It is not anymore only macro level matters that attract the attention of decision makers on statistical grounds but also, with an increasing importance, the social and economic dynamics of small parts of the Country at sub-regional and even sub-provincial areas. Satisfaction of such local information requirements with a desirable accuracy and precision would have been possible only with very large normal survey samples at unbearable costs. Attempts of finding feasible ways for this purpose have led the statisticians to use the small area statistics methodology along with the planned conventional census and survey activities.

The main distinction of small area statistics from the usual population subset (domain) statistics is that regular sample survey designs can not provide sufficient amount of data for small areas or small subpopulations of interest for a valid and viable statistical inference. Nature of the variables of interest, complexity of the structure of population of small area , spatial and other relationships between small area units are some major elements of difficulty in creating dependable and standard methods of statistical approaches. Therefore sample size determination, estimation of characteristics of interest, and measurement of quality of statistics for small areas require special approaches in regard of general design based methodologies.

Among the several statistical activities of SIS , the critical ones that need small area estimates have been: Household Income and Consumption Expenditure Survey (HICS), Household Labor Force

Survey (HLFS) and Crop Area Estimation by Remote Sensing (CAERS). The important statistical data sources for small area estimates in addition to the surveys themselves are censuses of past years, administrative records and register systems. Censuses are limited in scope of variables and frequency of applications. Records and registers are usually insufficient to support the aim of estimates for small area. On the other hand, sample surveys can not always provide and support sample sizes for small area estimates of adequate precision for all and every small area overall the country. So, finding out some other sorts and sources of information has been necessary. New information and methods that make use of this information have been sought for.

Auxiliary information has been used to improve the sample designs of the surveys and studies discussed here. All knowledge about the population under study have been considered to be auxiliary information even if they have been unquantified. Such information sources have been singled out so far to provide data for sample units. This information has been used in different ways as (i) explanatory variables in regression, ratio and similar other estimators to improve statistical precision, (ii) explanatory variables for conditional inference in ratio or linear, non-linear model relations to provide efficiently stratified sample designs, and (iii) variables that indicate possible sub-areas in the analysis to determine poststratification and to ensure the desired precision by sub-areas. This kind of information has been used occasionally for estimation purposes at the aposteriori stage of statistical inference. So, auxiliary information has been combined with the small area estimation attempts in the studies of SIS to warrant validity in designs and accuracy in small area estimates.

Small area estimation here refers to small domains which are geographically defined. An extensive literature exist for the small domain estimation which is extended to geographically defined small area estimation problems. A good information on recent developments in this field can be found in the books of Platek, Singh, Rao and Sarndal (1987) and Central Statistical Office of Poland (1993 a and b). In the light of such reported international experience, methodological and practical progress have been developed in Turkish small area estimation case.

2.STATISTICAL ACTIVITIES WITH SMALL AREA ESTIMATION ATTEMPTS

The surveys considered in this work involve survey designs and integrated sample designs with the following essential components: Survey variables, sample design and data collection, analysis by models whenever necessary, choice of domains of analysis, identification and computation of sampling and nonsampling errors, and efforts to improve data quality. The common feature of these surveys is that they need to take the advantage of a set of knowledge on population to improve either the sampling design or the estimates for producing information on small areas. Depending on the survey goals, small area estimation in these surveys were for planned areas and unplanned areas both. Brief description and some methodological discussion for the three surveys follow below. Household Labor Force Survey is discussed in greater detail in Section 3.

2.1 Crop Area Estimation by Remote Sensing

The project of crop area (acreage) estimation by remote sensing (CAERS) was started in 1992 with the aim of establishing a satellite imagery based statistical prediction system that predicts the crop areas of the considered year timely overall the country. At the early stages of developing such a system, it was necessary to pursue image and ground survey based studies of all the designated surface areas. The initial goal was to get to learn the best possible image interpretation and image data based predictive modeling . So ground surveys were conducted actually to cross check the success of image processing and image data based prediction. In the progressive stages of the project this practice changed to be so that not all the surface areas considered in image analysis were subject to control by ground surveys.

The challenging aspect of CAERS studies was that the activities at ground survey and image analysis stages must be performed with minimal errors separately and then they must be merged to estimate the same unknowns: To do this, the image data had to be received from the selected earth surface areas first and then ground surveys were conducted on selected segments of the chosen earth surface areas. Through the minimization of the errors of the image based inference by cross-checking with the ground surveys, prediction of the crop areas were provided for the whole country. Some provinces have always been focal sub-areas in policy making; so small area estimation for those provinces was a subject of the study.

The secondary sampling units in this study that were to be selected for observation by ground surveys are the area segments of size 700 m. X 700 m. These segments took place in the 42 km.X42 km. area blocks that do not overlap and cover the whole country. Primary sampling units of the study were these blocks which need to be identified on the ground with maps in a digital environment. The satellite images of these 42 kmX42 km. blocks were received for selected areas two times in a year. Selection of the said area blocks and segments had to be made with an optimal and time dynamic sampling strategy.

Sampling of blocks and segments in the studies so far followed a two stage stratified sampling design. Variables and design of the sampling were determined mainly on the basis of the last available census information. A vegetation image of past years of the country were once used to classify the whole land into agricultural and non-agricultural classes. On the agricultural class, five categories were determined according to the crop area size and crop yield intensities per hectare using the last agricultural census data of the country. Agricultural census information was in terms of some geographical regions with known administrative boundaries. Since the collection of 42kmX42km. area blocks did not fit within the geographical regions of utilized agricultural census, the design of the CAERS samples had to consider this adverse effect by increasing the number of secondary sampling units in the sample.

The selection of 700mX700m area segments within each area block was performed by systematic sampling with a distance threshold to disallow two segments in the sample to fall too close to each other. That is to say that some relevant layers of information on the sampling frame elements were used to do the clustering scheme for stratification. This scheme consisted a combination of dissimilarity indicator between segments and geographic contiguity. The result was that the more the data available for small geographical units, the more the number of small and scattered pieces in a stratum.

The highlights of sampling design and sub-regional estimation for CAERS is as follows: On a continuous surface Y(z) is the variable of interest which is the crop area of the Country within a region A. z indicates area blocks. Let be the area of A. The mean value of Y(z) is needed. We use

(2.1)

to estimate the theoretical mean crop area

(2.2)

by n observations. Total crop area is obtained from . Suppose A is divided into H strata, and for stratum A_h, h=1,2....,H, each with an area of , we can obtain the crop area averages with properties;

(2.3)

where , and z and z' denote different blocks of areas, n_h is the size of sample of block areas from a stratum, and . By taking the expectations over the space we get the error variance;

(2.4)

where and z_h, z_h' are randomly selected block areas in stratum A_h . The aim is to minimize this error variance by designing an optimal sampling plan (Cressie(1993), Ripley(1991)).

Note that, if the sample size n_h for each stratum is let to be random, the design should handle also the post stratification. However, in this case, the probability law for n_h happens to be not binomial because variance estimates for n_h then depend on the shape, size and other geographic elements of the surface areas.

The efficiency of stratification is Eff=V_{nstr}/V_{str} where V_{nstr} is the estimate of variance without stratification while V_{str} is its counterpart with stratification. Sample size becomes nxEff without stratification at the same precision. In our case gain in stratification becomes large when strata with high crop area and density have a uniform distribution of crop acreage overall the region .

The strata sample size values n_h can be calculated by minimizing the following quadratic loss function

(2.5)

subject to the constraint n= . Here is the crop area estimate for strata h, , U_h is an auxiliary information indicator reflecting the importance of stratum

A_h with respect to the agricultural area and crop yield intensity levels, and . The optimal n_h has been obtained by Gebizlioðlu, Aral and Teksoy (1995) as

(2.6)

where S_h is the sample variance of Y_h for stratum $A_h$and . Note that if q=1 and Y_h=$U_h$the optimal n_h turns out to be Neyman allocation (Cochran, 1977). A value of q between 0 and 1 allows a strategy between Neyman and equal coefficient of variation allocation between the strata.

The total crop area is estimated using :

(2.7)

where is the probability of inclusion of area block *i* into the sample size *n*, X is the auxiliary information integrated into the design, and N is the total number of area blocks covering the whole Country. Expression (2.7) is a special case of a regression estimation model :

(2.8)

with , , and .

Using a specified precision value for sample estimates and the coefficient of variation quantity

, we can estimate the sample sizes from the general expression :

(2.9)

where is the standard normal deviate value corresponding to significance level ,

(2.10)

and

. (2.11)

The precision requirement can be used to take care of precision desired for various domain and small areas that are of interest.

After the completion of ground surveys and image analysis, estimation of crop areas for small areas becomes necessary whenever the ground survey information has been insufficient. Small area definition here applies to provinces .

The adopted small area estimation model is the so called Battese-Fuller regression estimation model (Battese and Fuller (1981), Walker and Sigman (1984)) which assumes a stratified sampling design . It is actually a linear random effects model. This model can be applied within the regions for all strata where classification and regression have been completed. The general form of the model is:

(2.12)

, $j=1,.....,n_{hp}$

where Y_{hpj} = crop area in stratum h, province p, and segment j ; and are regression parameters for stratum h, X_{hpj} = number of pixels on the imagery classified into the class of concerned crop area in stratum h, province p, and segment j, and e_{hpj} = total error, f_{hp} is the small area (province) effect, and is the random error.

Synthetic estimations can be made in those strata where regression model is not a reliable fit due to very few segments observed. In such cases, average crop area per segment for the stratum under consideration is used for synthetic estimation.

(2.13)

where N_{hp}=Number of population units in stratum h, province p and is the average observed crop area per segment in stratum h.

The estimation procedures and other relevant issues mentioned here can be followed from Sarndal, Swenson and Wretman (1992) and Skinner, Holt and Smith (1989).

The assessment of small area estimates are usually made by calculating , which is root mean square error relative to estimate . This assessment for CAERS shows that as correlation between ground survey and image analysis results increases, model fit gets better to produce more reliable crop area predictions. High spatial correlation between the observations contributes to the realization of efficient estimates.

2.2 Household Surveys

Among the household surveys that have been the subject of study for small area estimation, the Household Income and Consumption Expenditure Surveys(HICS) were restarted in 1994, after a long interval since 1987. As different from the previous surveys, the 1994 HICS used a new methodology with separate phases of applications for consumption expenditures and income distribution estimation purposes.

For the whole year of 1994, a completely rotating 2188 households were sampled per month for the consumption expenditure measurements. In the second phase of the survey, household income questionnaire were applied on the same 26256 households early in 1995. Consumption expenditures survey produced the following information for the whole country: residence type, residence ownership, residence facilities, basic household social and economic features, consumption expenditures by the type and magnitude of products and services bought. The income values in the same survey were obtained by recording monthly income and annual income with separate questions. The inconsistency between income and expenditures led SIS to conduct income survey in the second phase of studies . Similar procedures were applied for the 1995 HICS whose complete results are yet to come out.

The sampling design for HICS studies was multistage stratified cluster sampling designs with the objective of producing household income and consumption statistics at the levels of the whole country, urban areas, rural areas, seven geographical regions and 19 preselected provinces. The most important part of the sampling design was the determination of the best factors for stratification. The efforts spent on this purpose entailed the art of creating auxiliary information from existing sources, old surveys and censuses; and using it for the design of efficiently stratified samples including the additional stratification to achieve the required precision by small areas.

The outline of the approach in the sampling design for HICS studies is as follows: Letting Y to be the variable(s) of interest in the surveys, the auxiliary variable(s) X are included in the sampling design. A random sample S is drawn from the population with N units. is the first order selection probability of design p(.) for the primary sampling units. is the second order probability of inclusion. Each unit assumes values of variables Y and X. T(y) is a function of $(y_1,y_2,.......,y_N)$ of interest for estimation. It is usually the case that T(y) is a function of y_i's with w_i weights which depend on sample S and a known auxiliary information.

Realization of small variance for T(y) is accomplished by using an auxiliary variable X in relation to Y, and by conducting a sampling with inclusion probabilities proportional to X_i. In the case of getting sample based -estimators for fixed sample size, this approach leads to ^{2} values close to zero; so the variance of estimate is as small as possible. The selection of auxiliary variables can also be performed using the second order probabilities , if necessary, with the aim of building a design that yields the effect that is close to zero for -estimators of T(y).

In HICS studies, the Y variables of interest are various that can be grouped under three classes:(i) Conditions of residence and related social and economic features, (ii) Consumption expenditures by type, amount value and source, and (iii) Household composition, employment and income by type, value and place. Both the sampling design and estimation of parameters of interest were to be accomplished by using auxiliary variables with due attention to small area estimation for planned small areas.

The auxiliary information consists of about sixty variables under the following variable groups: population by age groups, education and professional status, status at work, professional skills and position at work, labour force and income in economic activities by sectors of the economy.

Variables at the intersection of economic sectors, income by economic activity, and age groups within geographical regions are the most important auxiliary information variables for sampling design and estimation purposes. Age groups happens to be the most distinctive piece of auxiliary information for efficient stratification in the sampling design. The profession and job of principal income earner of the household, conditions of residence, and number of working members of household are the factors for stratification at secondary stage, in order of importance.

As to the estimation of parameters of interest through the statistic T(y), like totals or means of variables of interest Y for the population, sample based expansion type estimations is the major estimation method in the studies. The elements of estimation are as follows: is the population size in stratum h for designated small area . is the counterpart of in sample which is the subset of overall sample S that fall in area . The expansion estimator T(y) is

(2.14)

where and w_i is the survey design weight for unit i. Under the -estimate strategy the weights become proportional to inclusion probabilities. This estimation method is the most practical one with respect to convenience in computing and inclusion of attribute type stratification variables into the estimation. Note that the strata considered in T(y) expression can be post-strata as well as design strata. Whenever measure type important auxiliary variables are found going well along with small area values of interest for estimation, ratio estimation becomes the preferred one for its smaller variance. Letting that auxiliary variable be X with known strata values of interest in estimation, the estimator turns out to be

. (2.15)

The ratio type estimation can be extended to regression prediction if covariates are available. However, the design inconsistency of this estimator is usually a problem. This disadvantage can be alleviated by regression estimators. This requires efforts to ensure the linear and additive relation of covariates X and variable of interest Y in the regression equation for small areas.. The general regression model built on this purpose is

(2.16)

where is expansion of direct sample based estimator and so is T(x).

HICS studies regards the seven geographical regions and the provinces in the design as levels of estimation. If there are quite similar areas within the levels, estimates for the levels are found to be reliable for the small areas. Given that the best auxiliary information has been found and the correct X variables are designated to represent this information in the ratio or regression equation, synthetic estimators yield quite precise estimates for this case. The auxiliary data in direct synthetic estimation are to be the population counts . In order to get the best result from synthetic estimators , the bias of the estimator must be reduced or removed. After the estimation of bias, correction of synthetic estimates yield adjusted synthetic estimators (Gonzales (1973), Ghangurde and Singh (1978)). The vulnerable assumption here is that the relationships between variables of interest and auxiliary variables at small areas stay stable over time and space.

The Household Labour Force Surveys (HLFS) in Turkey were started in 1966 by SIS. To improve the quality and international comparability of data, the questionnaire, the sampling scheme and definitions were completely revised in 1988, with biannual applications since then. The aim of HLFS studies was to produce data on labour force participation rate, unemployment ratio and number of persons employed, underemployed, and unemployed. The number of marginal workers, and some information about the informal sector were also sought after. Information about persons not in labour force was among the objectives. This last item consisted of eight subgroups. Information about income of households was a byproduct of the studies which was to be used also for the analysis of consistency with HICS results. The definitions, classification standards, and the main methodological setup of the studies were in compliance with the international recommendations of United Nations(1989).

The way of weighting the row data in the surveys changed in 1990 due to the recalculation of national results by extrapolating the sample results using the factors based on 1990 Population Census. The series of HLFS results since 1988 were revised accordingly. An important change in the design of questionnaire was realized in the 1994 HLFS application. The use of Eurostat guidelines was started then for the questionnaires and the data definitions. Two forms composed the questionnaire: A form for household demographic characteristic and labour force status, and another form for the labour force status of the household members who are aged 12 and over.

The small area estimation matters and relevant matters for HLFS studies of SIS since 1994 are discussed in more detail in the following section.

3. HOUSEHOLD LABOUR FORCE SURVEYS

The questionnaires and sampling design of HLFS were revised in 1994 to resolve a number of issues. Modifications in the questionnaires were made with the aim of covering new concepts and international conformity, and to reflect the possible chances in the labour force status in Turkey. The revision of sampling design was to respond to the need of taking into account the results of the last population census and the requirements of metropolitan and regional estimates. Meeting the special needs for the survey on child labour and for more precise results for specified individual regions were among the objectives.

A multistage stratified sampling design has been in existence for HLFS samplings. Random selection of ultimate sampling units, which are households, have been done from the clusters of these units at the final stage of the sampling. The design for the 1994 and 1995 surveys consisted of two sampling sets with four subsamples in each. The main sampling set was for the periodical applications while the other one was a complementary reserved sample set to serve to specified regional estimation purposes. The three subsamples of the main sample sets were used for the periodical applications by rotation, the other was again for the regional estimations. The subregional estimates required the use of small area estimation methods. It was indispensable to find the most effective auxiliary information variables to do the stratification and estimation.

The stratification in the surveys were done according to the following main factors:

- Geographical regions (8 regions)
- Rural and Urban locality by population size
- Urban settlements by population size
- Rural settlements by population size
- District types by economic and social features

x

The primary strata were designed in terms of locality size groups on the basis of the most recent population census of 1990. This stratification allowed rural, urban division. The data collection and sampling rate differed according to the primary stratum, in particular by city, town, and village stratification. The second level of stratification was by eight geographical regions. Further stratification were achieved by some social and demographic variables like female literacy, currently working female population, internal migration and labour mobility by localities. Female literacy was found to be the most effective stratification variable because it had a large variation and coefficient of variation by locality, and the effectiveness of stratification was increased by controlling this variation. For a given primary stratum, the sampling procedure and rate was identical accross all eight geographical regions. It was therefore not necessary to form geographical regions as explicit strata for the selection of sampling units. The cities with population over 200 000 made a separate stratum altogether for sample selection.

A controlled selection of sample units was considered in the application of the surveys. This was of particular importance for village level of rural areas due to the high variation of employment status at this level. Four village size groups were determined. Each replication within each region had a sample coming from each of the four village size groups. The sample apportionment within each size group was by the level of female literacy. This selection procedure retained the probability nature of the sample which was based on the stratified sampling in two stages for the rural area level. The sample rotation system was as follows: while fourth subsample designated at random was kept as reserved for special needs like sub-domain or small area estimation, the three replications were used for normal application of HLFS. Of the three, two replications were included in any one application for six monthly round. A replication remained in the sample for two successive applications and left the sample during one application to be reintroduced for the next two replications. At each replication, households were relisted and reselected.

The stratification of large rural and urban areas were done as four strata and sub-strata considering the urban and rural distinction also. Selection of samples were made in three stages; namely location, block and household selections.

The estimated number of households in each stratum were multiplied by the sampling rate “k.f_0” to obtain the target sample size. f_0 is the overall sampling rate for the selection of households in villages. For the other strata the rate of selection was found by “k_t.f_0” for towns, and by “k_c.f_0” for cities. k_t/k and k_c/k ratios were to be determined to achieve the overall sampling rate that was to meet the requirements of overall sample size of n households in each application (round) of the HLFS. This required that “n=f_0.k.Hhs”, Hhs being the number of households. In this context, k_t and k_c were oversampling factors for towns and cities, respectively, compared to villages. Determination of k,

k_t, and k_c was an optimal sample size determination problem for the said objectives of the HLFS studies under the constraint of the budget limitations.

3.1 Sample Size and Sampling Weights

The size of the 1994 and 1995 HLFS samples was determined by choosing strata sample sizes n_h to minimize the desired variances under the given constraints. The computational procedures were according to the known basic methodology (Cochran (1977, pp.89-149)).

The sampling intervals were so determined that the required overall sampling rates were to be achieved in the selection of households. So, the sample weights were to be computed systematically by making direct use of the information on design probabilities and response rates. Up to this aim, an overall blow-up factor F had to be computed at the national level. This required the comparison of population projections with the population size covered in the sample. The reliable population projections at several regional, demographic and administrative levels by the State Institute of Statistics (1995) were used for this purpose. All other sampling weights were in relative terms to the national level with average weight of 1.0 at each subsequent sampling stage.

The design weights were applied inversely proportional to the overall selection probabilities. In 1994 and 1995 HLFS applications, the multiplier k in the sampling rate computation formula "$k.f_0$" was 1.0 for the villages, and 1.5 and 2.0 for towns and cities respectively. So the design weights were proportional to 1, 1/1.5 and 1/2.0 in villages, towns and cities, respectively. In order to ensure that the average value per unit in the enumerated sample is 1.0 , the weights were normalized as follows:

so that (3.1)

where n_h is the sample size for strata h and w_h are the above mentioned weights.

Further adjustments had to be made to the design weights in account of non-responses and sample distribution of population. Homogenous sets of clusters were created within each stratum, and response rates information R_{hj} for each of those were utilized. The resulting normalized weights were:

(3.2)

where is the overall average response rate in the survey. On the other hand; the correction of sampling distribution was necessary to match it with the population control totals which were based on the population projections. The control totals were the distribution of national population by gender and age at the cross section of total population by regions.

The normal design of HLFS studies considered the requirement of independent estimator for 21 major cities. The total sample size was near the upper limit manageable from cost and practical matters point of view. The expanded sample for the 21 major cities was allocated to each city in proportion to square root of the city population size. Further constraints had to be placed on the maximum and minimum limits of the sample allocated to any city. The size of expansion, and maximum and minimum limits were computed according to the required precision in estimates subject to the given constraints.

The conduct of HLFS studies was planned in order to produce detailed results at the small geographical regions. The required minimum sample size for this purpose was considered to be of the order of 6000-8000 households per eight geographical regions of the sample design. The use of the mentioned reserve subsample was a part of the plan for sub-regional estimations. The problem of designing samples for sub-regional estimation arised mainly with regard to the unknown quantities of the stratification and weighting variables. Therefore, the use of auxiliary information from the most recent census data become necessary. Letting X_{gj} stand for j'th household which fall in the g'th category with respect to all categories of households by several ranges of values of X; the linear regression

(3.3)

was proposed for the determination of sampling weights. On the basis of the information generated by this regression model, the j'th sampling unit had to assume a g-weight, in addition to the sampling weight proportional to 1/ , being the selection probability reflecting the known auxiliary total X_g within the category which j'th unit belongs to. That is, the calibration of g-weights was to be used so that the g-weighted sample sum of auxiliary variable equals the known area totals of these values

The overall sampling weights are then proportional to "(1/)(g-weights)". The technical details of regression fitting and calibration, mentioned here, is a general knowledge (see, for instance, Sarndal, Swenson and Wretman (1992)). The auxiliary variable total for the known categories of units can be estimated by Hurwitz-Thompson estimator

(3.4)

where the summation is over the unit falling in the known category. The proposed g-weight for the j'th unit is

(3.5)

where U_j weight is found from the calibration and regression model fitting efforts.

Addition of the reserve subsample of HLFS sampling design into the overall sample serves to purpose of small area estimation with the sampling weights as explained above. Inclusion of the reserve sample to the normal one increases the size of the overall sample without disturbing the normal sample and its rotation pattern.

3.2 The Small Area Estimation of Labour Force

The small area estimation in HLFS studies entails the estimation of employment, unemployment and underemployment for designated small areas which are usually the centers of cities,cities other than the twentyone major cities, and large towns. There has been no official publication on small area estimates by SIS, whereas the results of the normal surveys are already available either as books (State Institute of Statistics (1994)) or as provisional results in the form of statistical news bulletins. Small area estimation of employment, unemployment and underemployment is a challenging job under the conditions of constantly changing social and economic features of the population. Dynamic structure of sectoral job markets (Bulutay, (1995)), educational status of women, and age group dependent demographic aspects of the employable household members have been the major elements of changing conditions.

Therefore, in addition to the size of the concerned population by small areas there has been another set of factors to be considered in small area estimation which are the population groups by the major elements of changing conditions of employable people. The size of those population groups within the small areas has never been known beforehand. So indirect estimation along with the direct estimation, whenever possible, has had to be performed.

Similar situations have been considered in some other countries. The reports by Falorsi, Falorsi and Russo (1993, 1995) and Elliot (1993) explain the estimation methodology aspects of small area estimation in some country specific labour force surveys. They report that the following estimators have been applied in their labour force surveys: Direct, ratio, post-stratified, synthetic, composite, regression, sample size dependent, and time series estimators. Empirical comparison of the estimators have been provided in their works with the remarks that there is no uniformly best estimator by the efficiency of estimator criteria. In each area there seems to be a different best estimator. The inclusion of auxiliary variables in the estimation procedures produce good results, as also reported. As to the general presentation and evaluation of small area estimation methods and models; the works of Singh, Mantel, Thomas (1994), Ghosh and Rao (1994) and Rao and Gosh (1994) are excellent surveys that can be cited among others. The growing number of proposed estimators and estimation procedures in this field shows that the subject of study is interesting from both theory and application points of view. The characteristics and methodology of small area estimation for labour force surveys in Turkey is outlined below.

The studies of HLFS small area estimation utilize several data sources such as the HLFS survey data, administrative registers, past population census and extended population surveys data, HICS survey data, and the auxiliary small area data on which synthetic estimators can be based. The estimation efforts are mainly for the planned small areas that are the provinces other than the 21 major provinces of the normal sampling design or the centers of the selected cities for which the labour market information is critical. Unplanned small areas emerge due to the fact expressed before that the literacy level of women and migration dynamics, among others factors, happen to be important factors effecting the variation between the small areas. The planned small areas are within the eight geographical regions of the sampling design for which separate samples are designed and selected as mentioned. Note that the estimation for 21 major cities can actually be considered as small area estimation which have been taken into account at the design stage separately. Design based estimation procedures are already being used for those provinces.

The HLFS small area estimation problem contains the intervention of the three groups of variables: (i) the interest variables Y for which small area estimates about labour force are required, (ii) the auxiliary variables Z used a priori which are included in the sample design and values of which are generally known for all the population units, (iii) the auxiliary variables X used a posteriori in the estimation procedure for which current small area statistics are available from several data sources. Both Y and X variable values are recorded in the sample surveys while Z values are not. It is the X variables which are used to do adjustments in the estimation stage.

The reference to the variables in the estimation for small areas is made by the following notation:

: Total of the interest variable Y for (hijga)

: Mean of the interest variable Y for (hijga)

where h is the stratum (h=1,2,.....,H), i is the primary sampling unit (i=1,2,.....,N_h), j is the secondary sampling unit (j=1,2,.....,M_{hi}), g is the group (g=1,2,.....,G) index, and is the area index (=1,2,.....,) such that the population U is supposed to be divided into nonoverlapping small areas. The groups in HLFS studies are built up as states composed by age, gender, education level, and migration status of the households. N_h indicate the number of primary sampling units in stratum h, and M_{hi} stands for the secondary sampling unit number in h'th stratum, i'th primary sampling unit. The counterpart of N_h and M_{hi} are n_h and m_{hi} which indicate the number of selected units in respective strata. The similar counts for small area units are indicated by the addition of subscript to the others.

The intersections of some small areas with the design strata may be null so target population in small areas is only an improper subset of U. Consequently the sample S contains an improper subset which is that part of S belonging to small area . Since HLFS samples are drawn by multistage stratified sampling design, most small areas cut accross all or some of the strata which is referred to as crossclasses. This is caused by the inclusion of group factors (effects) into the estimation. Having these groups in the estimation does not only improve the efficiency of estimates but also contributes to the production of statistics with more meaningful components for decision makers. If one or more strata corresponds entirely to the small areas in the HLFS samples, that is to say that small area may coincide with identified strata, the standard results for simple random sampling designs would be adequate to use in the estimation of totals or means of characteristic of interests. This is not the case in HLFS studies. So the complication brought by the crossclasses into the estimation of population totals or means has to be accommodated. The reason for the arising crossclasses is mainly the existence of factors of women literacy and internal migration. Subclasses like gender and age groups are almost uniformly distributed over the population. Whereas, the classes by the literacy level of women and duration of residency in an area are much less well distributed.

The estimation methods utilized in HLFS studies so far are as follows:
Direct estimation of totals are obtained by

(i) Expansion estimator

(3.6)

where the summation is three-fold over h, i, and j, v_{hij} is a random variable which assumes value 1 if unit (ij) in stratum (h) belongs to small area ( ), or zero, otherwise. The survey weight is

(3.7)

with SU_h as the number of units belonging to strata h, and SU_{hi} as the number of units belonging to the population subset of primary sampling units in strata h. Note that the further normalizations on the weights are necessary to cope with the selected village, town and city coefficients k, k_t and k_c respectively, and to account for nonresponses which are mentioned in Section 3.1.

(ii) Ratio (r) estimator

(3.8)

where is the is the number of units belonging to area ,

(3.9)

where the summation is threefold over h = 1,2,.... , is the total number of design strata which contain small areas, i = 1,2,..., , and j = 1,2,...M_{hi}. is its expansion estimator.

(iii) Post stratified ratio (p) estimator

(3.10)

where is the number of units belonging to group g of area and and are expansion estimators of and , respectively. The subscript g added in the subscripts of Y and SU is to indicate the restriction to the group classes.

The estimates by these direct estimation methods are easy to calculate and not complicated for interpretation. Expansion estimator does not use any auxiliary information. Since the sample size is a random variable in this case, the stratum variance of the estimator is found to be large unless the group factor variable is fairly uniformly distributed over the stratum. The variance of the expansion estimator for the whole country is large. The ratio estimation has a bias which effects the post-stratified estimator. The ratio and post-stratified ratio estimators are preferred for making estimations at those strata where sample sizes happen to be large enough. Ratio and post-stratified type estimators are applied within design strata as well as post-strata.

The average values of variable of interest is obtained from the estimates of totals in the usual way of dividing small area totals by the size of small area samples.

Model Based Estimation methods are used for HFSL results whenever implicit or explicit models and their assumptions are used. Among many known estimators, regression and synthetic estimators are the ones tried, so far.

(iv). Regression estimators are considered for application in order to get design unbiased estimators by using all data collected in the HLFS's. Hence design consistency of the estimators can be achieved for large sample sizes whenever the right auxiliary variables are found as covariates of the variables of interest. In this sense, regression estimators are direct sample based estimators within design strata or post-strata. However, they are actually model based since either the weights in regression or covariates carry assumptions. In very simple notation as in Section 2.1, the regression estimator is

(3.11)

with l Î , where W_l is the survey weight for unit l, x and y are observed values, X is the auxiliary variable and v_l is the regression weight.

Another regression model that takes into account the small area subset of population and subsample value of auxiliary variable X is

. (3.12)

Here, and can be either post-stratified or expansion estimates. is obtained by the usual regression parameter estimation methods.

(v) Synthetic Estimation assumes that the larger area which contain a small area has very similar features in regard of the estimation problem in general, and sub-groups in particular. The synthetic estimation methodology and its advantages and disadvantages for use in unemployment surveys are discussed in detail by Gonzales and Hoza (1978). Synthetic estimators are biased and the size of the bias in HLFS studies increases as the said assumption of the methodology fails to be realized. However, they are useful estimators when the small area sample size is very small. They are design inconsistent even if the sample size is large when assumptions do not hold true.

The mostly used synthetic estimator in HLFS studies is the ratio synthetic estimator based on stratification,

(3.13)

where and are the expansion estimates of totals of Y and X in strata h, h=1,2,....H.

Regression synthetic estimator within post-strata proves to be reliable for some sub-strata whenever the linear association measure correlation between Y and its covariate X is high enough. In some stuations bivariate ratio synthetic estimators are seen to produce good results if a second variate is found in post-stratification stage which is highly explanatory for variation in Y.

Sample dependent estimators and composite estimators are in the application stages. The unstable behaviour of linear combination coefficient between several small areas for crossclass groups is a problem in composite estimators. However, as more sample outcomes are received through the repetition of surveys, sample dependent estimator, as a particular case of the composite estimator, is expected to yield good results. This is in the sense of less bias of expansion estimator component and smaller variance of the post-stratified estimator component of the estimator.

4. CONCLUSION

The small area estimation in Turkey is being continued for several surveys with better and persistent application periods, now. So the identification of domains and small areas is done better in advance for the design and redesign of the surveys. Pooling of estimates over systematically repeated surveys is expected to increase the quality of data and reliability of estimates especially for the time dependent labor force characteristics. The use of suitable time series methods is introduced into the studies up to the satisfaction of this need.

Careful selection of auxiliary variables and stratification in sampling design are always important matters for small area estimation irrespective of how well the advance planning is. A keen attention needs to be paid for the identification of the interacting interest , a priori design and a posteriori estimation variables. The determination of the correct design weight is critical for both the design and estimation stages. The adverse effects of nonresponse problem in the surveys must be minimized. Normalization of design weights in account of nonresponse rates may be a requirement in most repetitions of surveys.

Generating several estimates of parameters and approximation of the variance in proposed estimates is done using the data from complex surveys. The complexity of surveys make the estimation of variance quite complicated. Suitable procedures like pseudosampling (half replication), bootstrapping and jackknifing are recommended for the approximations (Skinner Holt, and Smith (1989)) at each repetition using all the accumulated information of relevance.

The household statistics are always influenced by the population dynamics and demographic phenomena including the migration movements. The inter-regional impulse response behaviour of the household labor force, and similar other household characteristics varies with the short, medium and long term movements associated with changes in the spatial distribution of population growth, population mobility and migration, status of women in every aspect, production, wages and so on. Depending on how the spatial system of interest is defined, external linkages and interaction of characteristics of interest with the multivariate temporal and location characteristics can be modeled (Martin, Thrift and Bennett (1978)). Such spatial interaction and distribution models cast light on the identification of variables and defining samples with better understanding of the phenomena and the parameters that underlie the surveys.

REFERENCES

BATTESE, G.E., FULLER, W.A., (1981). “Prediction of County Crop Areas Using Survey and Satellite Data”, Survey Section Proceedings, 1981 American Statistical Association Annual Meeting, Detroit, Michigan, 500-505.

BULUTAY, T., (1995). Employment, Unemployment and Wages in Turkey, Ankara: ILO and SIS, SIS Printing Division.

CENTRAL STATISTICAL OFFICE OF POLAND, (1993a). Small Area Statistics and Survey Designs, Volume I, Warsaw:Central Statistical Office of Poland.

CENTRAL STATISTICAL OFFICE OF POLAND, (1993b). Small Area Statistics and Survey Designs, Volume II, Warsaw:Central Statistical Office of Poland.

COCHRAN, W.G., (1977). Sampling Techniques, NewYork;Wiley.

CRESSIE,N.A.C., (1993). Statistics for Spatial Data, NewYork: Wiley.

ELLIOT, D., (1993). “A Project to Examine the Potential For Small Area Estimates of Employment and Health Data in the UK”, Small Area Statistics and Survey Designs, Vol. II, 143-151, Warsaw:Central Statistical Office of Poland.

FALORSI, P.D., FALORSI, S., RUSSO, A., (1993). “Empirical Comparison of Small Area Estimation Methods for Italian Labour Force Survey”, Small Area Statistics and Survey Designs, Vol. I., 17-33, Warsaw: Central Statistical Office of Poland.

FALORSI, P.D., FALORSI, S., RUSSO, A., (1995). “Small Area Estimation at Provincial Level in Italian Labour Force Survey”, Proceedings of 1995 Annual Research Conference, 617-633, Arlington, Virginia: US Department of Commerce, Bureau of the Census.

GEBÝZLÝOÐLU, Ö.L., ARAL, H.M., TEKSOY, N., (1995). “A Statistical Approach to Crop Yield Estimation Via Remote Sensing”, Turkish Journal of Physics, 19, 1041-1048.

GHANGURDE, P.D., SINGH, M.P., (1978). “Evaluation of Efficiency of Synthetic Estimates”, Proceedings of the Social Statistics Section of the American Statistical Association, 53-61.

GHOSH, M., RAO, J.N.K., (1994). “Small Area Estimation: An Appraisal”, Statistical Science, 20, 55-93.

GONZALES, M.E., (1973). “Use and Evaluation of Synthetic Estimators”, Proceedings of the Social Statistics Section of the American Statistical Association, 33-36.

GONZALES, M.E., HOZA, C., (1978). “Small Area Estimation With Application to Unemployment and Housing Estimates”, Journal of the American Statistical Association, 73,7-15.

MARTIN, R.L., THRIFT, N.R., BENNETT, C.,eds, (1978). Towards the Dynamic Analysis of Spatial Systems, London: Pion Limited.

PLATEK, R., SINGH, M.P., RAO, J.N.K., SARNDAL, C.E., eds, (1987). Small Area Statistics: An International Symposium, NewYork: Wiley.

RAO, J.N.K., GHOSH, M., (1994). “Some Current Developments in Small Area Estimation”, Proceedings of the 1994 Annual Research Conference, 306-332, Arlington, Virginia: U.S. Department of Commerce, Bureau of the Census.

RIPLEY, B.D., (1991). Statistical Inference for Spatial Processes, London: Cambridge University Press.

SARNDAL, C.E., SWENSON, B., WRETMAN, J., (1992). Model Assisted Survey Sampling, Berlin: Springer-Verlag.

SINGH, A.C., MANTEL, H.J., THOMAS, B.W., (1994). “Time Series EBLUPS for Small Areas Using Survey Data”, Survey Methodology, 20, Statistics Canada.

SKINNER, C.J., HOLT, D., SMITH, T.M.F., eds, (1989). Analysis of Complex Surveys, New York: Wiley.

STATE INSTITUTE OF STATISTICS, (1994). 1994 Household Labour Force Survey Results, Ankara: SIS Printing Division.

STATE INSTITUTE OF STATISTICS, (1995). The Population of Turkey, 1923-1994: Demographic Structure and Development with Projections to the mid-21st Century, Ankara: SIS Printing Division.

UNITED NATIONS, (1989). National Household Survey Capability Programme Household Income and Expenditure Surveys, NewYork: United Nations Division of Statistics

WALKER, G., SIGMAN, R., (1984). “The Use of Landsat for County Estimates of Crop Areas: Evaluation of the Huddleston-Ray and the Battese-Fuller Estimators for the Case of Stratified Sampling”, Communications in Statistics-Theory and Methods, 13, 2975-2996.