Gad Nathan, 40 Years of Contributions to Official Statistics

Luisa Burck*
Central Bureau of Statistics, Jerusalem, Israel louiza@cbs.gov.il

Abstract

One of Gad Nathan's most outstanding accomplishments was combining his academic career with work for the Israeli Central Bureau of Statistics (CBS) where he started as the Director of Statistical Methods Division and ended as the Bureau's Chief Scientist. As a leading survey statistician with a well-established international recognition, Gad Nathan initiated and carried out scientific research at the Bureau in all areas concerning official statistics; development of new sampling methods, building questionnaires, conducting duplicate surveys, application of analytical inference from complex surveys, analysis of categorical data, analysis of non-sampling errors, record matching processes and their evaluation, longitudinal data analysis, etc. As of January 2012, CBS made a transition to a monthly Labour Force Survey from a quarterly system of measuring labour force characteristics. Up to the transition, many statistical issues were addressed anew at the CBS and some major changes were introduced to sampling design and estimation as well as to relating logistics and operations of the survey. This paper, dedicated to Gad's memory, summarizes the statistical research conducted during and following the transition period, while emphasizing the principal statistical theories and applications concerning this particular survey.

Key Words: Panel Survey, composite estimation, calibration.

1. Introduction

The Israeli Labour Force Survey (LFS) is the major household survey conducted by the Israeli Central Bureau of Statistics (CBS) since 1954. The survey follows the development of the labour force in Israel, its size and characteristics, as well as the extent of unemployment and other trends. It is also a data source on living conditions, education and schooling, immigrants of 1990 and after, etc. This survey attempts to cover all persons aged 15 and older in the permanent population of Israel. For urban areas, the sample is generally drawn from addresses provided in the municipal tax records and supplemented by addresses for new dwellings. In smaller localities, the sample is drawn from list of households or in case of kibbutzim, persons. Most of the final sampling units are selected using two-stage cluster sampling within strata, with clustering of persons within households. In the first stage, localities are sampled with probability proportional to size. Typically, large urban areas are included with certainty. In the second stage, dwellings are sampled from the selected localities in a manner that provides every household of the population the same inclusion probability. Until 2012, the LFS was conducted quarterly, with households asked to participate in four interviews, waves. The waves are spread over a year and a half, with households included in the sample for two consecutive quarters, excluded for two quarters, and then included for two final quarters. This rotation structure was designed to increase the precision of estimates of change across adjacent quarters and

years, while not placing too great of a burden on any household. The national sampling fraction for the quarterly LFS in the past years was 0.5% of the population of households (according to the annual report, the sampling fraction was 1% till 1982). For the entire country, the CV of the quarterly labour force participation rate was 0.70 percent and of the unemployment rate was 3.10 percent (this translates to standard error of 0.30 for the unemployment rate obtained as a direct estimate).

As of January 2012, CBS made a transition to a monthly LFS from the quarterly system in order to produce more reliable monthly estimates of levels and changes as requested by the decision makers, researchers and others. Although, in the statistical system of a small country, as Israel, relatively large samples in proportion to the population have to be used to attain significant results, the national sampling fraction for the monthly survey was set to 1% of the population due to budget issues.

In the following, the attempts to achieve sophisticated sample designs and estimation procedures to improve the precision of the monthly estimates, based on prior evaluation of the parameters involved, are presented. In Section 2, we briefly examine the monthly rotation schemes used in different countries and the associated response rates and discuss the implications for the monthly estimates of level and change. In section 3, different estimates obtained through composite estimation considered by the CBS are presented, while a special focus is set upon regression composite estimation, Fuller & Rao (2001). The results, based on simulated data that allow comparing between the efficiency of different estimates, are illustrated in Section 4.

2. Choice of Monthly Rotation Scheme

A quick survey of rotation schemes used in different countries suggests variation in both rotation patterns and implied priorities. Some countries design their rotation schemes to obtain significant overlap between dwellings in both adjacent months and quarters and years. For instance, the United States interviews households in four consecutive months, excludes them for eight months followed by additional interviews in the same months of the next year **(4-8-4)**. This leads to a 75% overlap between adjacent months and a 50% overlap between adjacent years. Other countries emphasize the precision of monthly and quarterly changes. Probably, the large annual samples sizes also reduce the need for a further increase in the precision of yearly change. For example, Australia keeps households in the sample for eight consecutive months **(8in)** leading to an 87.5% overlap between adjacent months, but has no overlap across years. In **2-2(4)** rotation scheme the households are interviewed for two consecutive months and excluded in the next two months; this pattern is repeated four times within a fourteen month period. This leads to a 50% overlap between adjacent months and a 25% overlap between adjacent years. Note that in all three rotation schemes, in any given month interviews are attempted in equal proportions for cases from eight panels. Thus, the sampling errors of changes over a specific time can be reduced significantly by the choice of the appropriate rotation scheme while the sampling error of the monthly level remains unaffected.

Let Y_t be the monthly level estimate at time t. Assuming that $\mathrm{Var}(Y_{t-s}) \approx \mathrm{Var}(Y_t)$, the variance of the change over s months can be written as:

$$\mathrm{Var}\left(Y_t - Y_{t-s}\right) = 2\mathrm{Var}(Y_t)\left[1 - K(s)\rho(s)\right]$$

where $K(s)$ denotes to the proportion of overlap of the sample s months apart and $\rho(s)$ is the true correlation between Y_t and Y_{t-s} (it is assumed that it does not depend on t |or that the process is stationary). On one hand the rotation scheme can reduce the sampling errors of

the changes over time, and on the other hand the sampling errors of level estimates may become greater, in particular, of quarterly and yearly level estimates. Based on, the quarterly LFS the true correlations of the labour force characteristics over time were estimated. Given these correlations, the sampling errors with no rotation scheme can be compared to those obtained with a rotation scheme. In Table 1, for a given rotation pattern the ratio of the sampling errors obtained with rotation to those with no rotation.

Table 1: The ratio of the sampling errors of estimates of equal size samples with and without rotation

Estimate	Type of Estimate	No rotation	With rotation		
		1in	2-2 (4)	4-8-4	8in
Monthly Total (level)	Level estimate	1	1	1	1
Total Employed Persons	Monthly change	1	0.73	0.55	0.42
	Quarterly change	1	0.82	0.88	0.67
	Yearly change	1	0.89	0.77	1
Total Unemployed Persons	Monthly change	1	0.82	0.71	0.64
	Quarterly change	1	0.91	0.94	0.85
	Yearly change	1	0.97	0.93	1
Percent Unemployed Persons	Monthly change	1	0.81	0.69	0.62
	Quarterly change	1	0.91	0.94	0.84
	Yearly change	1	0.96	0.93	1

3. Choice of Estimation Method

3.1 Generalized Regression Estimator

Calibration estimation is a flexible and integrating approach for incorporating auxiliary information at the estimation stage of a survey. It is flexible because it allows for both binary and continuous auxiliary variables, as well as for combinations of those to be used. It applies equally well for simple and complex, stratified multi-stage sample designs. It is integrating because it encompasses several estimators which are widely used by practitioners, such as poststratified, ratio and regression estimators. They possess the nice "calibration property", by which weighted estimates for the auxiliary variables considered will match the known population totals of these auxiliary variables. They are relatively simple to implement, because the estimators are linear given a set of calibrated weights, which need to be computed only once for each survey data set. These estimates as proposed by Deville and Särndal (1992) are widely in use in the CBS.

Let us denote $U = \{1,...,k,...,N\}$ population of households of size N, s $(s \subseteq U)$ sample of size n and $\pi_k = \Pr(k \in s)$ the sampling probability of household k. Let $\mathbf{x}_k = (x_{k,1},...,x_{k,q},...,x_{k,Q})$ be vector of Q auxiliary variables also observed for n sample units where population totals $t_x = \sum_{k \in U} \mathbf{x}_k$ of these variables are known. We are interested in estimating $Y = \sum_{k \in U} y_k$, target variable such as number of employed persons. $\hat{Y}^{HT} = \sum_{k \in s} d_k y_k = \sum_{k=1}^{n} d_k \sum_{i=1}^{C_k} y_{k,i}$,where $d_k = \pi_k^{-1}$, is the Horvitz-Thompson (HT) estimator of the target variable and $\hat{\mathbf{t}}_x^{HT} = \sum_{k \in s} d_k \mathbf{x}_k$ is the sum obtained from the

sample for the auxiliary variable. Usually, $\hat{\mathbf{t}}_x \neq \mathbf{t}_x$ and calibration is needed to achieve the equality. The Generalized Regression estimator (GREG) is given by

$$\hat{Y}_{GREG} = \sum_{k\in s} w_k y_k = \hat{Y}^{HT} + (\mathbf{t}_x - \hat{\mathbf{t}}_x^{HT})\hat{\boldsymbol{\beta}}_s$$

where $w_k = d_k\left\{1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_x{}^{HT}\right)\left(\sum_{k\in s} d_k \mathbf{x}_k{}'\mathbf{x}_k\right)^{-1}\mathbf{x}_k^{'}\right\}$ and the regression coefficient can be written as $\hat{\boldsymbol{\beta}}_s = (\sum_{k\in s} d_k \mathbf{x}_k{}'\mathbf{x}_k)^{-1}\sum_{k\in s} d_k \mathbf{x}_k{}' y_k$. In the quarterly survey, in order to eliminate negative and very small weights the following distance function is used $G_k(w_k, d_k) = (x-L)\log\frac{x-L}{1-L} + (u-x)\log\frac{U-x}{U-1}$ where U and L are two limits set with $L<1<U$ and $Ld_k < w_k < Ud_k$ and x is *the ratio* w_k/d_k. Estimates obtained using these weights are asymptotically equal to those estimated by GREG.

3.2 The Regression Composite Estimator

In this paper, composite estimates, in the framework of sampling and calibration for LFS, refer to exploiting the past data, mainly those of recent months for the estimation at time *t*. For example, composite estimates will be based on data for months *t* and $(t-1)$ as opposed to direct estimates that are based on data for time *t* only. For this example, M_t (Matched) will be those panels interviewed both at time *t* and $(t-1)$ and B_t (Birth) will be the new panels interviewed at time *t*. Modified Regression Estimate of type 1 (MR1) as appears in Singh (1996) reduces mainly the standard errors of the level estimates. The MR1 can be implemented within the ordinary regression scheme by introducing a new variable for each person *i* in household *k* into the calibration:

$$x_{k,i}^{(L)} = \begin{cases} \bar{y}_{t-1} & \text{if } k \in B_t \\ y_{k,i,t-1} & \text{if } k \in M_t \end{cases}$$

where $\bar{y}_{t-1}$ is the mean (proportion) of the labour force status at $(t-1)$. MR2 estimate reduces significantly the sampling error of the monthly change and is implemented by introducing the following variable for each person *i* in household *k* into calibration:

$$x_{k,i}^{(C)} = \begin{cases} y_{k,i,t} & \text{if } k \in B_t \\ y_{k,i,t} + R(y_{k,i,t-1} - y_{k,i,t}) & \text{if } k \in M_t \end{cases}$$

where $R = K^{-1} = \sum_{k\in(M_t\cup B_t)} w_{k,t} \Big/ \sum_{k\in M_t} w_{k,t}$. Fuller and Rao (2001) suggested the weighted average of the two variables above: $z_{k,i,t} = (1-\alpha)x_{k,i}^{(L)} + \alpha x_{k,i}^{(C)}$ with $0 \le \alpha \le 1$. Obviously, the choice of the parameter α depends on the rotation pattern. In their paper, Gambino et al. (2001) report a big gain in efficiency, in the Canadian LFS (6in) resulting from the use of these estimates instead of direct estimates.

3.3 Estimation

The use of the regression composite estimator above may lead to multicollinearity of the variables used in calibration and as a result the weights may differ greatly from the sampling weights. As a solution, we consider first using Ridge regression with $\lambda \neq 0$. The weights obtained in this case are defined by $w_k = d_k + \left(\mathbf{t}_x - \hat{\mathbf{t}}_x{}^{HT}\right)\left(\sum_{k\in s} d_k \mathbf{x}_k{}'\mathbf{x}_k + \lambda \mathbf{I}_Q\right)^{-1} d_k \mathbf{x}_k^{'}$ and the estimate is given by $\hat{Y}_{Ridge} = \sum_{k\in s} w_k y_k = \hat{Y}^{HT} + (\mathbf{t}_x - \hat{\mathbf{t}}_x^{HT})\hat{\boldsymbol{\beta}}_R$, $\hat{\boldsymbol{\beta}}_R = (\sum_{k\in s} d_k \mathbf{x}_k{}'\mathbf{x}_k + \lambda \mathbf{I}_Q)^{-1}\sum_{k\in s} d_k \mathbf{x}_k{}' y_k$.

Note that Ridge regression introduces bias but at the same time reduces the variance significantly so that the MSE will be smaller than that of $\hat{Y}_{GREG}$. Another solution is Raking where the coefficients are obtained iteratively.

4. Results from Simulated Data

In order to investigate the rotation scheme and the parameter α relevant to Israel, we created a data set based on the quarterly data for 2004-2006 (12 quarters). Taking into consideration the rotation pattern of the quarterly survey and based on two consecutive quarters, data was imputed for the two missing months using logistic regression model. The population created included 22,053 households with 52,115 individuals. A simple random sample was drawn within each stratum and this sample was divided into panels accordingly for each rotation scheme.
For each rotation scheme a grid search was applied for different values of α and MSE(GREG)/MSE(COMPOSITE) was calculated. Table 2 shows the gain in MSE for some selected variables for different α for each rotation scheme.

Table 2: The gain in MSE of the composite regression estimator (versus direct estimator) - percent

Rotation pattern	α	Unemployed persons - Males	Unemployed persons - Females	Employed persons - Males	Employed persons - Females	Unemployed persons - Total	Employed persons - Total	Percent of employed persons	Monthly change in percent unemployed persons	Number of households	Persons with more than 12 years of education	Employed persons in construction
8in	0.00	12.82	9.67	13.81	12.11	10.82	10.70	11.85	14.64	-38.83	4.14	2.23
	0.25	15.72	10.96	16.62	17.41	13.52	15.11	14.45	24.05	-39.33	6.39	2.47
	0.40	16.42	10.98	16.48	19.22	13.83	16.26	14.65	29.46	-39.91	7.29	2.53
	0.50	16.22	10.37	15.43	19.51	13.34	16.06	14.08	32.17	-40.42	7.50	2.53
	0.60	15.32	9.21	13.44	18.80	12.21	14.79	12.76	34.17	-41.00	7.40	2.52
4-8-4	0.00	15.58	12.90	17.26	21.62	13.76	16.73	14.68	22.06	-46.52	8.43	1.34
	0.15	16.23	15.07	22.35	26.76	14.32	21.75	15.43	30.87	-49.06	9.95	0.90
	0.30	16.46	14.96	25.46	30.30	14.31	25.04	15.19	42.46	-50.43	10.76	0.57
	0.45	13.45	12.09	24.59	30.34	10.91	24.64	11.46	51.27	-52.29	9.70	0.22
	0.60	8.35	7.53	19.70	25.90	5.57	19.70	5.69	57.23	-54.18	6.83	-0.16
2-2 (4)	0.00	15.06	11.63	11.23	12.22	9.86	9.68	10.17	20.61	-44.40	3.67	0.48
	0.08	15.29	7.60	14.57	15.21	7.29	12.69	7.74	20.23	-46.31	5.27	0.47
	0.10	13.89	5.86	15.55	15.95	5.02	13.42	5.54	20.68	-46.86	5.63	0.39
	0.15	11.50	3.39	17.00	17.40	0.91	14.76	1.51	23.73	-47.60	6.13	0.19
	0.20	9.69	2.10	18.30	18.62	-2.13	16.08	-1.54	27.73	-48.75	6.61	-0.32

In 8in and 4-8-4 rotation schemes $\alpha = 0.4$ and $\alpha = 0.3$ seemed to be optimal, respectively. In 2-2(4) rotation scheme no such value was found (but for further work it was set to 0.1). These results were further checked for different sample sizes and under different sampling designs (different stratification).
Furthermore, three different estimation methods were compared, GREG, Ridge regression and Raking. For Ridge regression the best results were obtained for $\hat{\lambda} = 25$. For 4-8-4 rotation pattern with $\alpha = 0.3$, the gain in MSE was highest through Raking in all level estimates and for a large number of change estimates as can be seen in Table 3.

Table 3: Average gain in MSE for different estimation methods - percent

	Level estimate			Monthly change		
		$\lambda = 25$			$\lambda = 25$	
Variable	GREG	Ridge	Raking	GREG	Ridge	Raking
Not in labour force - Males	17.20	19.60	19.20	29.90	30.20	29.40
Employed persons - Males	19.20	19.70	17.40	30.50	33.90	32.20
Unemployed persons - Males	8.10	19.00	15.00	19.50	22.90	19.50
Not in labour force - Females	19.30	14.80	22.40	29.70	28.00	31.40
Employed persons - Females	20.10	18.00	24.50	32.20	27.40	33.30
Unemployed persons - Females	11.70	12.90	14.50	18.90	16.80	21.70
Employed persons - Total	20.50	18.50	22.40	31.80	30.60	31.90
Unemployed persons - Total	8.50	20.20	18.80	18.60	20.70	20.40
In labour Force - Total	16.60	13.30	17.50	28.60	26.30	27.00

References

Bell, P. (2001). Comparisons of Alternative Labour Force Survey Estimation, *Survey Methodology*, **27**, 53-63.

Cochran, W.G. (1977). *Sampling Techniques*. 3rd Ed. New York: John Wiley & Sons, Inc, *Chapter 12.*

Deming, W.E. and Stephan, F.F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginals are Known. *The Annals of Mathematical Statistics***, 11**, 427-444.

Deville, J.C. and Särndal, C.E. (1992): Calibration Estimators in Survey Sampling, *JASA*, Vol. **87**, No. 418. 376-382**.**

Fuller W.A. and Rao J.N.K. (2001) A Regression Composite Estimator with Application to Canadian Labor Force Survey, *Survey Methodology* , **27**, 45-51.

Gambino, J.G., Kennedy, B. and Singh, M.P. (2001) Regression Composite Estimation for Canadian Labour Force Survey: Evaluation and Implementation, *Survey Methodology*, **27**, 65-74**.**

Särndal Carl-Erik, Swensson Bengt, Wretman Jan. (1992). *Model Assisted Survey Sampling*. Springer Series in Statistics, *Chapter 9***.**

Singh, A.C. (1996) Combining Information in Survey Sampling by Modified Regression, Proceedings of the Section on Survey Research Methods, *American Statistical Association*, 120-129.