# مجلة التجارة والتمويل

https://caf.journals.ekb.eg/

## كلية التجارة – جامعة طنطا


العدد : الرابع

ديسمبر 2023

(الجزء الثانى)

AL-AZHAR UNIVERSITY
GIRLS' BRANCH CAIRO
FACULTY OF COMMERCE
DEPARTMENT OF
STATISTICS

# Efficient Goal Programming Approach in Statistical Matching

Prepared by

**Abeer Mohammed Mokhtar Esmail Hussein Elrefaey**

Assistant Lecturer of Statistics, Faculty of commerce,
Girls' Branch Cairo, Al-Azhar University

Email: abeerelrefaey@ azhar.edu.eg

Supervised by

**Prof. Ramadan Hamed Mohamed**
Professor of Statistics
Faculty of Economics and Political Science
Cairo University
Email: ramadanh@aucegypt.edu

**Prof. Elham Abd Elrazik Ismail**
Professor of Statistics and Dean of the Faculty
Faculty of Commerce, Girls' Branch Cairo
Al-Azhar University

Email: elhamismail@azhar.edu.eg

**Dr. Safia Mahmoud Ezzat Mohallal**
Assistant Professor of Statistics
Faculty of Commerce, Girls' Branch Cairo
Al-Azhar University
Email: safiahamed@azhar.edu.eg

# Efficient Goal Programming Approach in Statistical Matching

**Abstract**

Statistical matching methods goal is to combine several data sources to build datasets. The main goal of statistical matching is to make helpful and informative synthetic data without collecting more data or making new surveys. The study aims to use the goal programming approach in statistical matching to complete the data in two files, where the first file contains variables different from the second file, with one or more of the common variables. To reach this goal, a linear regression model is designed for each of the variables in each file in terms of the variables in the two files. The goal programming approach was used to estimate the parameters of the two regression models, and from it the estimated value of the variables presents in the first file and not present in the second file, and so on, hence we get a file with all the variables. The goal programming approach has the advantage of minimizing the effect of outliers with estimates because it uses minimization of the sum of absolute deviations. Moreover, the proposed approach has a constraint that guarantees significant estimations of the parameters. In addition to formulating the model, A simulation study evaluates the proposed approach's performance by generating and imputing data for dependent variables from different distributions. Results show the efficacy of the approach in accurately estimating missing values while maintaining data quality and minimizing errors.

**Keywords**

Statistical matching, goal programming, linear regression, L1 (least Absolute), efficiency

## 1. Introduction

The origins of statistical matching can be traced back to the mid-1960s, when a comprehensive data set with information on socio-demographic variables, income and tax returns by family was created by matching the 1966 Tax File and the 1967 Survey of Economic Opportunities [Okner (1972)]. Then, in the early 1970s different matching techniques were applied to social surveys in the US [(Ruggles and Ruggles 1974)], but these techniques were severely criticized on the grounds that they rely on assumptions neither justified nor testable [Kadane (1978)], [Rodgers (1984)]. Interest in procedures for producing information from distinct sample surveys rose in the following years.

Statistical matching is a process of combining two or more datasets that have some overlapping variables or observations. The goal of statistical matching is to create a larger, more comprehensive dataset that includes all the relevant information from each of the individual datasets. In the usual statistical matching framework, the variables X and Y are observed the survey A, while X and Z are observed in B; while the X variables are common to both the surveys, the variables Y and Z are not jointly observed. Data file A contains $n_A$ observations of (X, Y), and data file B contains $n_B$ observations of (X, Z). The aim of statistical matching, namely the gain of joint information about variables not jointly observed [ D'Orazio *et al*. (2006)].

Statistical matching techniques play a valuable role in several situations, including the matching of two non-overlapping surveys that share common variables, the integration of Big Data into survey or administrative datasets, and the determination of imputation values for cases where specific groups intentionally lack data on certain variables [de Waal (2015)].

Kim and Park (2019) explored statistical matching as a technique to combine different data sources gathered from the same population, particularly in cases where the variables of interest are not simultaneously observed. This method is cost-effective and can create synthetic data using existing sources. The paper presents various statistical micro matching methods, specifically designed for categorical or categorized variables, a common occurrence in sample surveys. It introduces methods that account for conditional independence assumption (CIA) and proposes approaches for both scenarios: without auxiliary information and with auxiliary information to mitigate bias. A statistical matching method with auxiliary information is proposed for situations where CIA assumptions don't hold. This method shows improved performance compared to the random hot deck when there's a higher association between Y and Z. The recommended approach here uses auxiliary information and performs better when moderate association between Y and Z is suspected. Overall, the simulation study indicates that the size of file C, containing auxiliary information, doesn't need to be large, suggesting that overcoming the limitations of CIA isn't a major practical concern in terms of cost.

Moretti and Shlomo (2023) explored the application of statistical matching methods to integrate different datasets containing information about various social aspects. The challenge often arises from the assumption of conditional independence among variables observed in distinct data sources. To address this, the article introduces an auxiliary dataset (file C) that encompasses all variables, aiming to enhance statistical matching by incorporating the correlation structure of variables across datasets. The proposed methodology involves modifying prediction models using calibration from file C, which is particularly helpful in cases where the auxiliary file may have uncertainties. The research addressed the limitations of conditional independence assumption between variables in different files. It proposes utilizing an auxiliary dataset to relax this assumption and improve matching quality. Calibration of prediction models within file C is recommended to enhance robustness against model misspecification and errors. Simulation results demonstrate that the proposed approach provides more accurate estimates of relationships between variables, particularly when correlation between certain variables is significant. The method also addresses model failures and non-normality.

There are several methods of statistical matching, but they all generally involve identifying common variables or observations in each dataset and then using statistical methods to match or merge the datasets based on these commonalities. Common methods include record

linkage, regression analysis, and imputation. Also, statistical matching can be performed using mathematical programming techniques, such as linear programming or integer programming. In this approach, the statistical matching problem is formulated as an optimization problem, with the objective of minimizing the difference between the combined dataset and a target dataset with complete information [Harris-Kojetin and Groves (2017)].

Donatiello *et al.* (2022) discussed the application of statistical matching methods to combine data from the EU Statistics on Income and Living Conditions and the Household Budget Survey (HBS). The aim is to create a synthetic dataset that enables a comprehensive multidimensional analysis of economic poverty among households in Italy. The study builds upon previous experiences at the Italian National Institute of Statistics and introduced a modified approach to matching complex sample survey data. The designed method allows for the creation of a synthetic dataset that maintains the marginal distribution of target variables. Unlike simpler donor-imputation techniques, this method considers the final survey weights, making it more intricate and necessitating additional validation steps. Preliminary results are encouraging due to meticulous pre-harmonization of reference surveys and the collection of pertinent data for statistical matching.

These works provide a starting point for understanding the mathematical programming approach to statistical matching, and there is a wealth of additional literature on specific techniques and

applications. These recent works demonstrate the ongoing development and innovation in the field of statistical matching using mathematical programming.

The remainder of this article is organized as follows: in section 2 proposed the suggested goal programming approach for statistical matching in two files. Section 3 introduced a simulation study to assess the performance of the suggested approach. Then the results were discussed in section 4 and provided conclusions in Section 5.

## 2. Suggested Goal Programming Approach

The suggested goal programming approach aims to minimize the sum of absolute deviation between the observed observations $y_i$ and the estimated observations $\hat{y}_i$. Then the objective function of the problem is to minimize:

$$\sum_{i=1}^{n}|y_i - \hat{y}_i| = \sum_{i=1}^{n} y_{diff} \qquad (1)$$

The equivalent goal programming model can be formulated as follows:

$$min \sum_{i=1}^{n}(d_i^+ + d_i^-) \qquad (2)$$

Subject to:

$$\sum_{j=0}^{k} x_{ij}\beta_j + d_i^- - d_i^+ = y_i, i = 1,2,\dots\dots,n \qquad (3)$$

$$\left|\frac{\bar{y}_{diff}}{s_{\bar{y}_{diff}}}\right| \leq t_{\left(\frac{\alpha}{2},n-1\right)} \qquad (4)$$

$$d_i^+ \geq 0, i = 1,2,\dots\dots,n$$

$$d_i^- \geq 0, i = 1,2,\dots\dots,n$$

$$\beta_j \; unrestricted \; in \; sign, j = 0,1,\dots\dots,k$$

where :

$X_{ij}$ are the observed independent variables.

$\beta_j$ are the parameters of the regression model.

$Y_i$ are the observations of dependent variable.

$d_i^+$ and $d_i^-$ are the positive and negative deviational variables, respectively.

The inequality (4) is added to ensure that significant estimates of the parameters were obtained using t distribution.

It can be proved that the objective function (2) is equivalent to minimizing the sum of absolute deviations between the observed and estimated values of the dependent variable $Y_i$ as follows:

$$d_i^- = \text{Max}\left(0, y_i - \sum \beta_{ij} x_j\right)$$

$$= 0.5\left(\left(y_i - \sum \beta_{ij} x_j\right) + \left|y_i - \sum \beta_{ij} x_j\right|\right)$$

$$d_i^+ = \text{Max}\left(0, \sum \beta_{ij} x_j - y_i\right)$$

$$= 0.5\left(\left(\sum \beta_{ij} x_j - y_i\right) + \left|\sum \beta_{ij} x_j - y_i\right|\right)$$

$$d_i^- + d_i^+ = \left|y_i - \sum \beta_{ij} x_j\right|$$

$$\sum_{i=1}^{n}(d_i^- + d_i^+) = \sum_{i=1}^{n}\left|y_i - \sum \beta_{ij} x_j\right| = \sum_{i=1}^{n}|y_i - \hat{y}_i|$$

## 3. Simulation study

Statistical matching aims to integrate two statistical sources. These sources can be two samples or a sample and data from population. If two samples have been selected from the same population and information has been collected on different variables of interest, then it is interesting to match the two surveys to analyze. Statistical matching uses variables common to both data sets to identify similar records that can be linked to generate a new synthetic data set that allows more flexible analysis than would be possible with the two discrete data sets. This section presents simulation study for the goal programming approach proposed method which allows the matching of two data files by using the imputation of one file on another file.

To evaluate the performance of the goal programming approach proposed method, the simulation method was used as follows:

1- Generating random data for two data files, where the first file contains the data of an independent variable $x_1$ the second file contains generated data for independent variable $x_2$ where $x_1$, $x_2$ the common variables ($x_1$, $x_2$ generated from a standard normal distribution).

2- Generating random variables for the first data file contains a dependent variable $y_1$ (generated from a lognormal distribution (1.2, 1), and for the second data file contains a dependent variable $z_2$ (generated from Chi square distributed (2)) from two heavy tailed or peaked tailed distribution.

3- The study considers different samples sizes: 200 and 500.
4- The simulation was repeated 1000 times for each sample size.
5- In each replication, the suggested goal programming approach ((2) – (4)) was used to estimate the regression coefficient.
6- The estimated regression model of the first model was used to find the estimated values of $y_1$ in the second file (named $y_2$) and the estimated regression model of the second file was used to find the estimated values of $z_2$ of the first file (named $z_1$).
7- The first file, the donor data, contains the data of an independent variable $x_1$ and a dependent variable $y_1$ is used to estimate target values $y_2$ in the second file, the recipient data.
8- The second file, the donor data, contains the data of an independent variable $x_1$ and a dependent variable $z_2$ is used to estimate target values $z_1$ in the first file, the recipient data.

To evaluate the performance of the goal programming approach through the simulation method that has been applied, the significance of the differences between the means of the generated values and the estimated values for each of the dependent variables was tested. In addition, the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), were used for comparison where the lower values of them are considered better:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \qquad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where

$n : sample\ size.$ $y_i : original\ value.$ $\hat{y}_i : imputed\ value.$

## 4. Results

Table (1) summarizes the results of the simulation including the mean of the means and mean of standard deviations for y and z. The table also presents percentage of insignificant differences between the means generated and estimated values of y and z.

**Table (1): Mean of the means and mean of standard deviations (for y and z) and percentage of insignificant when sample sizes (200,500) and 1000 replication.**

| Sample Size | Variables | Mean of Means | Mean of standard Deviations | % of insignificant differences |
|---|---|---|---|---|
| 200 | $y_1$ | 1.652629832 | 4.793742259 | 0.998 |
| | $y_2$ | 1.408056783 | 0.017009149 | |
| | $z_1$ | 1.762563462 | 0.030249684 | 0.989 |
| | $z_2$ | 1.99352654 | 3.987990877 | |
| 500 | $y_1$ | 1.652381086 | 4.693370298 | 0.999 |
| | $y_2$ | 1.495415923 | 0.007485333 | |
| | $z_1$ | 1.849951236 | 0.012576331 | 0.998 |
| | $z_2$ | 1.997772167 | 4.015548981s | |

The results of Table (1) indicate that the differences between the means of generated and estimated y are insignificant in 0.998 of the replications in sample size 200 and in 0.999 in sample size 500. Almost the same results exist for the generated and estimated

values of z. In addition, the results indicate less variability in the estimated values compared to generated values for both y and z and in the two samples.

Table (2) presents the means of MAE and RMSE of both y and z and for the sample sizes, 200 and 500 with 1000 replication.

**Table (2) means of MAE and RMSE**

| Sample Size | Variables | Mean of MAE | Mean of RMSE |
|---|---|---|---|
| 200 | y | 1.190531893 | 2.122820679 |
| | z | 1.417394367 | 2.00076172 |
| 500 | y | 1.213744239 | 2.136710622 |
| | z | 1.437098015 | 2.006300125 |

Results of Table 2 indicate that the mean of absolute deviations is less than the mean of Root Mean Square Error which justifies the objective of minimizing the sum of absolute deviations. The results also indicate slight differences according to the sample size. The results confirmed that goal programming approach has the advantage of minimizing the effect of outliers with estimates because it uses minimization of the sum of absolute deviations.

The goal programming approach was used to estimate the parameters in statistical matching is efficient in integrating data files that have common units and helpful and informative synthetic data without collecting more data. That means the goal programming approach can be used to integrate data files that have common units. The goal programming approach also has the advantage of minimizing the effect of outliers with estimates so the two dependent variables y1 and z2 were generated from two heavy tailed or peaked tailed distribution.

## 5. Conclusions

This paper presents a goal programming approach for statistical matching using linear regression. The goal programming approach was used to estimate the parameters in statistical matching is efficient in integrating data files that have common units and helpful and informative synthetic data without collecting more data. That mean the goal programming approach can be used to integrate data files that have common units,

The goal programming approach effectively addresses the challenge of combining datasets with differing variables, offering accurate imputations while minimizing the impact of outliers. The simulation study highlights the approach's success in achieving meaningful estimations and maintaining data quality. Further research could explore the application of this approach to more complex datasets and real-world scenarios.

**References**

De Waal, T., 2015. *Statistical matching: Experimental results and future research questions*. Statistics Netherlands.

Donatiello, G., D'Orazio, M., Frattarola, D. and Spaziani, M., 2022. The joint distribution of income and consumption in Italy: an in-depth analysis on statistical matching. *REVIEW OF OFFICIAL STATISTICS*, p.77.

D'Orazio, M., Di Zio, M. and Scanu, M., 2006. *Statistical matching: Theory and practice*. John Wiley & Sons.

Harris-Kojetin, B.A. and Groves, R.M., 2017. Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps. Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/24893.

Kadane, J.B., 1978. Some Statistical Problems in Merging Data Files. In 1978 Compendium of Tax Research, Office of Tax Analysis, U.S. Department of the Treasury, Washington, D.C.: U.S. Government Printing Office, 159-171.

Kim, K. and Park, M., 2019. Statistical micro matching using a multinomial logistic regression model for categorical data. *Communications for Statistical Applications and Methods*, 26(5), pp.507-517.

Moretti, A. and Shlomo, N., 2023. Improving Statistical Matching when Auxiliary Information is Available. *Journal of Survey Statistics and Methodology*, 11(3), pp.619-642.

Okner, B. A., 1972. Constructing a New Microdata Base from Existing Microdata Sets: The 1966 Merge File. *Annals of Economic and Social Measurement* 1, 325–362.

Rodgers, W.L., 1984. An evaluation of statistical matching. *Journal of Business and Economic Statistics*, 2, 91-102.

Ruggles, N. and Ruggles, R., 1974. A strategy for merging and matching microdata sets, *Annals of Economic and Social Measurement* 1(3) 353-371

Vantaggi, B., 2008. Statistical matching of multiple sources: A look through coherence. *International Journal of Approximate Reasoning*, *49*(3), pp.701-711.