Revista Colombiana de Estadística - Applied Statistics
July 2025, volume 48, issue 2, pp. 163–182
http://doi.org/10.15446/rce.v48n2.117060

Estimation of Variances using the Generalized Variance Function: Labor and Population Indicators in the Colombian Household Survey 2022

Estimación de varianzas utilizando la función generalizada de varianza: indicadores laborales y de población en la encuesta de hogares de Colombia 2022

Julian Diaz[1,a], Cristian Tellez[1,b], Felipe Ortiz[2,c]

[1]Facultad de Estadística, Universidad Santo Tomás, Bogotá, Colombia
[2]Facultad de Ciencias Matemáticas, Universidad Complutense de Madrid, Madrid, España

Abstract

This study addresses the challenge of estimating variances in household surveys, particularly when sampling design variables are absent in publicly available microdata. By implementing the Generalized Variance Function (GVF), Colombia's Household Survey for 2022 serves as a case study. GVF models were developed and validated using the standard errors published by the National Administrative Department of Statistics (DANE) of Colombia. These models demonstrated high accuracy and robustness for estimates across various levels of disaggregation and periodicities. Additionally, their validation with 2023 data confirmed their predictive capacity and applicability in similar contexts, underscoring their effectiveness as tools for evaluating the quality of estimates in complex surveys.

Key words: Generalized variance function; Complex surveys; Variance estimation; Sampling error household survey.

Resumen

Este estudio aborda el desafío de estimar la varianza en encuestas de hogares, causado por la ausencia de variables del diseño muestral en los microdatos públicos. A través de la implementación de la Función Generalizada de Varianza (FGV), se utiliza como caso de estudio la Encuesta de

[a]M.Sc. Student. E-mail: juliandiaza@usantotomas.edu.co
[b]Ph.D. E-mail: cristiantellez@usta.edu.co
[c]Ph.D. Student. E-mail: andror04@ucm.es

Hogares de Colombia para 2022. Los modelos de FGV se desarrollaron y validaron con base en los errores estándar publicados por el Departamento Administrativo Nacional de Estadística (DANE) de Colombia. Estos modelos demostraron alta precisión y robustez en estimaciones a diferentes niveles de desagregación y periodicidades. Asimismo, su validación con datos de 2023 confirmó su capacidad predictiva y aplicabilidad en contextos similares, destacando su eficacia como herramienta para evaluar la calidad de las estimaciones en encuestas complejas.

Palabras clave: Función generalizada de varianza; Encuestas complejas; Estimación de la varianza; Error de muestreo; Encuesta de hogares.

1. Introduction

Surveys are often designed to produce estimates for various demographic subgroups and geographic areas. Due to the large number of required estimates, statistical offices frequently opt to report a subset of direct estimates[1] along with their respective variances. This choice arises because, even with modern computational resources, directly calculating the variance for a large number of estimates remains both costly and time-consuming (Wolter, 2007).

Furthermore, databases available to external users of national statistical offices in Latin American and Caribbean countries often lack the necessary information to calculate the variances of estimators in household surveys, such as the "Encuesta Permanente de Hogares" (EPH) in Argentina, the "Gran Encuesta Integrada de Hogares" (GEIH) in Colombia, and the "Encuesta Nacional de Empleo, Desempleo y Subempleo" (ENEMDU) in Ecuador, among others. This omission stems from privacy concerns, as these databases exclude critical variables such as primary sampling units (PSUs), design strata, and clusters (Gutiérrez et al., 2020).

These limitations have prompted the adoption of models to approximate the standard error for estimates when direct variance is unavailable. Wolter (2007) proposes a model that expresses variance as a function of the expected value of the survey estimate, the sample size, and other variables that allow for model adjustments. This methodology is widely known as the Generalized Variance Function (GVF).

In large-scale surveys, where statistics are published across a wide range of topics, estimates for all elements of the survey can be calculated by simply evaluating a model on the survey's estimates, as noted by Handayani & Aunuddin (2005). This approach enables the use of GVF models when public-use databases lack sampling design variables.

According to Lohr (2021), it is crucial to consider how the model relates the variance of the estimate to the estimate itself. Particular caution should be exercised when applying GVFs to estimates that were not included in the parameter calculations, as this may result in unreliable outcomes.

[1] According to Lohr (2021), a direct estimator for a group is defined as one calculated exclusively using data from the sampled observations within that specific group. Examples of these estimators include those proposed by Horvitz-Thompson and Hájek.

Likewise, Wolter (2007) emphasizes that GVFs are particularly valuable in cases where the variance of direct estimates is not consistent across different groups. A GVF allows for the identification and mitigation of heteroscedasticity issues and outliers, which can have a significant impact on results. This is especially important because direct variance estimators often become unreliable in domains with small sample sizes.

According to Morales et al. (2021), the objective should not be to develop a highly predictive model but rather to smooth variance estimates based on the design of the direct estimators, prioritizing a model that is reasonably interpretable. As highlighted by Handayani & Aunuddin (2005), efforts have been made to develop GVF techniques applicable to quantitative characteristics. However, these attempts have encountered challenges due to the complexity of ensuring that all statistics within a group conform to the same mathematical model.

The literature on GVF implementation includes several notable studies. For instance, Alegria & Scott (1991) evaluated various GVF models to estimate the sampling error of the Forest Inventory and Analysis (FIA) survey conducted by the United States Department of Agriculture (USDA) for Kentucky in 1989. Their findings indicated that, for continuous variables, a GVF model incorporating the sampling errors of row and column totals achieved the best performance.

Similarly, Salvucci et al. (1995) analyzed the Schools and Staffing Survey (SASS), conducted by the U.S. Department of Education during 1990-91. Their study revealed that disregarding the sampling design used to collect the data resulted in an underestimation of sampling variances, which ultimately led to the adoption of GVF methods.

In the same line, Handayani & Aunuddin (2005) conducted an empirical study to evaluate the application of the GVF model for binomial variables in the 1998 Indonesia Labor Force Survey, focusing on the provinces of Java Island. Their study compared the relative variance estimates obtained using the GVF model with those calculated directly, highlighting the effectiveness of the GVF approach in this context.

Additionally, Kubacki & Jędrzejczak (2011) tackled the challenge of producing reliable estimates for small areas and assessed their accuracy within the Polish Household Budget Survey at the county level. Their findings demonstrated that the GVF model could generate reliable estimates not only for sampling errors but also for counties that were not part of the sample, showcasing its versatility.

Moreover, McIllece (2018) developed GVF models to estimate the mean and median unemployment duration statistics published by the Current Population Survey (CPS) in 2018. Their analysis evaluated the precision of standard errors during the modeling period and over a two-year projection, demonstrating that GVF models were sufficiently reliable throughout the reference period.

Furthermore, Zhang et al. (2019) extended GVF models to the Longitudinal Generalized Variance Function (LGVF), which simplifies to the GVF when applied to cross-sectional data. Their model incorporated the effect of time to capture dynamic changes over the years, using CPS data from 2008 to 2010. Their findings revealed that the ratios of relative variances to predicted relative

variances from the proposed LGVF converge in probability to 1 under certain regularity conditions.

On the other hand, Fúquene-Patiño et al. (2021) applied the GVF model in conjunction with the Fay-Herriot model to estimate the variance of direct estimators for calculating the Proportion of Households with at least one usual Member Living Abroad (PHMLA) at the municipal level in Colombia. They leveraged auxiliary information from the 2005 census and the 2015 Demographic and Health Survey (DHS), enabling them to adjust the model and reduce standard errors, particularly in municipalities with small sample sizes. Their study demonstrated that combining the GVF and Fay-Herriot models enhances the precision of estimates compared to direct methods.

A recent study by Gutiérrez & Babativa-Márquez (2023) for the "Comisión Económica para América Latina y el Caribe" (CEPAL) analyzed household survey data from 18 Latin American countries, employing three GVF models to estimate variances and sampling errors for indicators derived from these surveys, including the GEIH of Colombia. This research focused on a limited set of social indicators, such as per capita income, monetary poverty, extreme poverty, "Necesidades Básicas Insatisfechas" (NBI) dimensions, and labor force participation rates. The primary objective was to develop a robust methodology for calculating standard errors, particularly in contexts where information on sampling design is limited.

Some studies have not employed the GVF model but instead proposed alternative approaches. For instance, Carter & Rolph (1974) introduced an approximation for the variance of a proportion using a variance-stabilizing transformation based on the *arcsin* of the square root of the estimator. In their approach, the variance is expressed as $var\left(\frac{y_d}{n_d}\right) = \frac{1}{4n_d}$.

Additionally, Fay & Herriot (1979) proposed approximations for the coefficient of variation as a measure of estimate quality. They suggested calculating the coefficient of variation as $cve(\hat{\bar{y}}) = \frac{3}{\sqrt{N_d}}$. Furthermore, they observed that applying a logarithmic transformation to the variance stabilized it, resulting in an approximate value of $\frac{9}{N}$.

This study aims to provide a comprehensive framework for constructing GVF models, incorporating various variables to adjust the model to the available data and grouping indicators to enhance interpretability. This approach differs from that proposed by Gutiérrez & Babativa-Márquez (2023). In addition, this work offers a practical guide with examples to facilitate replication of the analysis by other researchers. The ultimate objective is to propose an alternative solution to address the absence of sampling design variables in publicly accessible datasets, a prevalent issue in official surveys.

As a case study, the 2022 GEIH was used to construct the GVF models. The results obtained from these models were compared to those published by the "Departamento Administrativo Nacional de Estadística" (DANE), demonstrating that GVF can serve as an efficient alternative for estimating variance.

Despite the complexity of the model and the limited familiarity with its implementation, GVF is expected to gain relevance in future studies. However, it is

important to note that, according to Valliant (1987), specific characteristics of certain surveys may render the GVF model less effective or even unsuitable. Finally, the results from the 2023 GEIH were used to validate the GVF models proposed in this study.

2. Generalized Variance Function (GVF)

Traditional techniques for estimating variances in simple random samples are inadequate for complex surveys due to limitations such as correlations between observations within strata, differing probabilities of inclusion for sampling units, and the effects of stratification and clustering. For these reasons, alternative techniques must be employed to estimate variances in complex surveys. According to Lohr (2021), in a complex survey with multiple levels of stratification and clustering, the variances of means and estimated totals are calculated at each level and then combined as one progresses through the survey design. Additionally, post-stratification adjustments and nonresponse significantly impact variance. Therefore, more advanced methods than traditional approaches are required to achieve precise variance estimates in complex surveys.

Furthermore, Lohr (2021) further notes that the GVF is a statistical model that relates the variance of a statistic to its expected value. The model coefficients are estimated using a set of statistics, such as totals, proportions, means, and their variances, derived through linearization or replication.

For practical purposes and ease of use, several surveys and statistical agencies provide GVF models to calculate standard errors. Examples include the CPS and the National Crime Victimization Survey (NCVS) in the United States, as well as Statistics Canada (STATCAN). The steps for calculating variances using GVF are outlined below:

1. Use a method to directly estimate the variance. In this case, the estimated population totals for the D domains of interest are $\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_D$. Let v_d denote the relative variance (*relvar*)[2] of $\hat{t}_d$, defined as $v_d = \frac{\hat{V}(\hat{t}_d)}{\hat{t}_d^2}$, for $d = 1, 2, \ldots, D$.

2. Postulate a model that relates v_d to $\hat{t}_d$. In most surveys using this approach, a linear regression model is adopted, with v_d as the response variable and $\frac{1}{\hat{t}_d}$ as the explanatory variable:

$$v_d = \beta_0 + \frac{\beta_1}{\hat{t}_d} + \varepsilon_d, \quad \text{where} \quad \beta_1 > 0, \tag{1}$$

where ε_d represents random errors assumed to follow a normal distribution with mean zero and variance σ_A^2, that is, $\varepsilon_d \sim N(0, \sigma_A^2)$. According to

[2]Most GVF models are formulated with relative variance as the response variable because it is based on the premise that relative variance v_d is a decreasing function of the magnitude of the expectation of $\hat{t}_d$ (Wolter, 2007).

Wolter (2007), special attention should be given to the value of β_0, as it could be negative. If $\hat{t}_d$ is sufficiently large for a particular domain, the variance estimate could become negative. In such cases, a constraint must be introduced.

3. Use regression techniques to estimate β_0 and β_1. At this stage, the ordinary least squares (OLS) method is applied to estimate the parameters. The regression coefficients are computed using the following expression:

$$\hat{\beta} = \left(\sum_{d=1}^{D} z_d z_d' \right)^{-1} \sum_{d=1}^{D} z_d \left(\hat{V}(\hat{t}_d) \right), \tag{2}$$

where z_d are the explanatory variables of the model, and $\hat{V}(\hat{t}_d)$ represents the estimated variance associated with the population total estimate $\hat{t}_d$.

4. Use the estimated regression equation to predict the relative variance of a new estimated total $\hat{t}_{\text{new}}$. The predicted relative variance is given by:

$$\hat{v}_{\text{new}} = \hat{\beta}_0 + \frac{\hat{\beta}_1}{\hat{t}_{\text{new}}}, \tag{3}$$

where $\hat{v}_{\text{new}}$ represents the predicted relative variance $\hat{V}(\hat{t}_{\text{new}})/\hat{t}^2_{\text{new}}$. Using GVF, the variance estimate $V(\hat{t}_{\text{new}})$ is calculated as:

$$\hat{V}(\hat{t}_{\text{new}}) = \hat{\beta}_0 \hat{t}^2_{\text{new}} + \hat{\beta}_1 \hat{t}_{\text{new}}. \tag{4}$$

According to Morales et al. (2021), the GVF model smooths variance estimates based on the survey design. It is recommended that those constructing GVF models avoid excessive smoothing. The GVF model does not need to provide precise predictions of v_d for different domains and parameters; instead, it focuses on generating smoothed estimates that are more robust to measurement errors and small sample sizes. However, a model with low predictive capacity could result in excessive smoothing of variance estimates. To mitigate this issue, auxiliary variables can be incorporated as explanatory factors in the GVF model. This approach reduces excessive smoothing and is particularly valuable when the auxiliary variables are interpretable.

Based on the insights of Lohr (2021), the GVF offers significant advantages in contexts involving complex sampling designs, such as the GEIH conducted in Colombia. Unlike methods such as Taylor series linearization, which requires differentiable expressions and can be labor-intensive for complex functions, the GVF smooths variance estimates using regression models, significantly reducing the time and effort needed to produce annual reports. Compared to methods like Bootstrap and Jackknife, which involve computationally intensive calculations and potential inconsistencies in small subsamples, the GVF provides a more efficient alternative, particularly when limited information is available in public data files.

Additionally, in contrast to replication methods such as Balanced Repeated Replication (BRR) and Random Groups, the GVF does not face strict limitations

in sampling design, as it models the relationships between estimators and variances more flexibly. Although its primary drawback lies in the potential inadequacy of the model for subpopulations with high clustering or deficiencies in regression parameters, its ability to optimize resources and adjust models makes it an ideal choice for large-scale surveys with high periodicity.

For this study, 12 anonymized microdata files from the GEIH of 2022[3], each containing an average of 80 000 records, were used as inputs to construct the GVF models. Similarly, 2023 data files[4], with similar monthly volumes, were used to validate the models. The GEIH, conducted monthly by DANE, is a survey based on a complex sampling design with different periodicities depending on the geographical level: monthly for the national total and the group of the 13 main cities, quarterly for municipal capitals and rural areas, and annually for intermediate cities and departmental capitals. The survey measures and analyzes the structure of the labor market, living conditions, and sociodemographic characteristics of households, serving as an essential tool for formulating public policies, planning social programs, and conducting technical and academic studies.

3. Variance Estimation using GVF

This section presents the GVF models used to estimate the variance of the estimators of interest. Additionally, the indicators analyzed in the study are defined.

3.1. GVF Models

According to Gutiérrez & Babativa-Márquez (2023), a GVF can be fitted to any estimator $\widehat{\theta}$, provided that a constraint is imposed on the relative variance, defined as $\widehat{v}_d = \frac{\widehat{V}(\widehat{\theta}_d)}{\widehat{\theta}_d^2}$. Without this constraint, undesirable negative variance estimates may arise. To address this issue, it is recommended to transform the dependent variable using a logarithmic function, as also suggested by Gutiérrez & Babativa-Márquez (2023). This transformation enables the modeling of $\log(\widehat{v}_d)$ instead of $\widehat{v}_d$, stabilizing the estimates during the adjustment process and preventing undefined values. By applying an exponential transformation to the predictions, this approach ensures that the variance estimates remain positive, thereby guaranteeing the model's stability and consistency.

As mentioned earlier, various GVF models have been developed for complex surveys to adjust the variances of direct estimators. In studies using survey data from Colombia, Fúquene-Patiño et al. (2021) proposed a log-linear model based on the GVF approach, specifically designed to estimate the proportion of households with at least one emigrant in Colombian municipalities. However, this model was excluded from the present study due to multicollinearity issues arising from the

[3] `https://microdatos.dane.gov.co/index.php/catalog/771/get-microdata`. Last accessed: 25 Jun 2025.

[4] `https://microdatos.dane.gov.co/index.php/catalog/782/get-microdata`. Last accessed: 25 Jun 2025.

simultaneous inclusion of related variables, such as the direct estimate and its square root.

Conversely, Gutiérrez & Babativa-Márquez (2023) proposed three GVF models in their analysis of social indicators for household surveys in Latin America. The first, the Krenzke log-linear model (LK), is based on the approach developed by Krenzke (1995), who introduced a model that flexibly analyzes variance in scenarios where design effects among estimates cannot be assumed constant or similar. This model was developed to address limitations identified by Valliant (1987), who noted that direct variance estimates may be unreliable when design effects vary. The model is defined as:

$$\log\left(\widehat{v}_d^{LK}\right) = \mathrm{z}_d'\beta = \beta_0 + \frac{\beta_1}{\hat{\theta}_d} + \frac{\beta_2}{\hat{\theta}_d^{1/2}} + \varepsilon_d, \tag{5}$$

where ε_d represents random errors assumed to follow a normal distribution with mean zero and variance equal to σ_A^2. The other two models proposed in that study, the common log-linear model (LC) and the empirical log-linear model (LE), were not considered in this work, as initial goodness-of-fit tests indicated better performance for the LK model (referred to in the rest of this document as the theoretical model).

It is important to note that, although the study by Gutiérrez & Babativa-Márquez (2023) included predictions of social indicators for Latin American countries, Colombia was not considered in the formulation of the models because the GEIH microdata did not include sampling design variables. This limitation was also observed in seven other countries in the region, making it impossible to adequately estimate direct variances and, consequently, to construct the basis for the GVF model. However, the methods developed in that study are noteworthy for their applicability to surveys with similar characteristics, offering a robust solution for scenarios where sampling design information is limited.

In this study, additional models were considered, building on the idea proposed by Wolter (2007), who emphasized the importance of developing alternatives to improve the fit of GVF models. Furthermore, studies such as that of Johnson & King (1987) demonstrated that incorporating categorical variables into GVF models captures the specific variability of population subgroups. The integration of auxiliary variables, including categorical ones, is essential to address heterogeneity in variance estimates and to achieve more accurate and robust models in the context of complex surveys. The proposed models are defined as follows:

$$\log\left(\widehat{v}_d^{Prop1}\right) = \mathrm{z}_d'\beta = \beta_0 + \beta_1 \mathrm{n}_d + \beta_2\hat{\theta}_d + \beta_3\ \text{Outlier} + \varepsilon_d, \tag{6}$$

where the variable "Outlier" indicates whether the record is classified as an outlier (the justification and definition of this variable are detailed in Section 4). Another proposed model is defined as:

$$\log\left(\widehat{v}_d^{Prop2}\right) = \mathrm{z}_d'\beta = \beta_0 + \beta_1 \mathrm{n}_d + \beta_2\hat{\theta}_d + \beta_3\ \text{Periodicity} + \beta_4\ \text{Outlier} + \varepsilon_d, \tag{7}$$

where the variable "Periodicity" indicates whether the estimate is classified as "monthly" or "quarterly." Finally, another model is defined as:

$$\log\left(\widehat{v}_d^{Prop3}\right) = \mathrm{z}_d'\beta = \beta_0 + \beta_1 \mathrm{n}_d + \beta_2\hat{\theta}_d + \beta_3 \text{ Indicator} + \beta_4 \text{ Outlier} + \varepsilon_d, \quad (8)$$

where the variable "Indicator" is categorical and represents the specific indicator being estimated.

Additionally, the variance predictions of the estimators using GVF, under any of the studied models, are calculated as:

$$\widehat{V}(\widehat{\theta}_d) = \exp\left(\mathrm{z}_d'\beta\right) \cdot \widehat{\Delta} \cdot \widehat{\theta}_d^2, \quad (9)$$

where $\mathrm{z}_d'\beta = \log\left(\widehat{v}_d\right)$ represents the relative variance estimates, with z_d' as the vector of explanatory covariates and $\hat{\beta}$ as the regression coefficient estimates. Meanwhile, $\widehat{\Delta} = \frac{\sum_{d=1}^{D} \widehat{v}_d(\widehat{\theta})}{\sum_{d=1}^{D} \exp(\mathrm{z}_d'\widehat{\beta})}$ is the bias correction factor in a log-linear regression, where D represents the total number of domains considered in the model. Ignoring this correction factor leads to an underestimation of the true variances when applying GVF (Morales et al., 2021).

3.2. Indicators of Interest

In this study, key indicators related to labor conditions and population distribution were analyzed. These indicators are defined as follows:

- **Unemployment Rate**: This is the percentage ratio between the number of people actively seeking employment (DS) and the labor force or "población económicamente activa" (PEA) (MinSalud, 2024). The PEA includes individuals of working age who are either employed or actively seeking work. This indicator is expressed as:

$$\text{Unemployment Rate} = \frac{\text{DS}}{\text{PEA}} \times 100 \quad (10)$$

- **Employment Rate**: This represents the percentage ratio between the number of employed individuals (OC) and the working-age population or "Población en Edad de Trabajar" (PET) (MinSalud, 2024). The PET includes all individuals aged 12 or older in urban areas and those aged 10 or older in rural areas. The formula for this indicator is:

$$\text{Employment Rate} = \frac{\text{OC}}{\text{PET}} \times 100 \quad (11)$$

- **Global Participation Rate**: This is the percentage ratio between the PEA and the PET (MinSalud, 2024). This indicator is defined as:

$$\text{Global Participation Rate} = \frac{\text{PEA}}{\text{PET}} \times 100 \quad (12)$$

- **Total Number of Employed People**: According to MinSalud (2024), this refers to individuals who, during the reference period, met one of the following conditions: worked at least one hour for pay (in money or kind) during the reference week, had a job but were temporarily absent from work, or were unpaid family workers who worked at least one hour during the reference week.
- **Total Number of Unemployed People**: As per MinSalud (2024), this includes individuals of working age who are unemployed, available to work, and actively seeking employment during the reference period.
- **Population Outside the Labor Force**: This refers to individuals of working age who did not participate in the labor market during the reference period, meaning they were neither employed nor actively seeking employment. This group includes students, homemakers, retirees, and others who chose not to engage in economic activities (DANE, 2024*b*).

Based on these definitions, the indicators are classified into two groups: labor market indicators, which include the unemployment rate, employment rate, and global participation rate; and population indicators, which encompass the total number of employed people, total number of unemployed people, and the population outside the labor force.

4. Results and Discussion

The sampling design of the GEIH is considered complex due to its combination of advanced techniques: it is probabilistic, stratified, clustered and multistage. This means that sampling units are selected in stages (municipalities, census sectors, households), strata are used to ensure geographic and socioeconomic representativeness, and households are grouped into clusters to reduce operational costs. Additionally, expansion factors are applied to adequately represent the population. DANE (2024*a*) publishes periodic bulletins containing information on labor market and population indicators, including estimates at different geographic levels, accompanied by standard errors as measures of statistical precision.

For the analysis of 2022, two GVF models were constructed, each tailored to one of the two groups of indicators studied: labor market indicators and population indicators. Following the recommendations of Wolter (2007), which emphasize the importance of grouping statistics into GVF models when they share fundamental characteristics–such as belonging to the same economic or demographic dimension, targeting the same population group, or referring to a common geographic level– this separation was deemed appropriate.

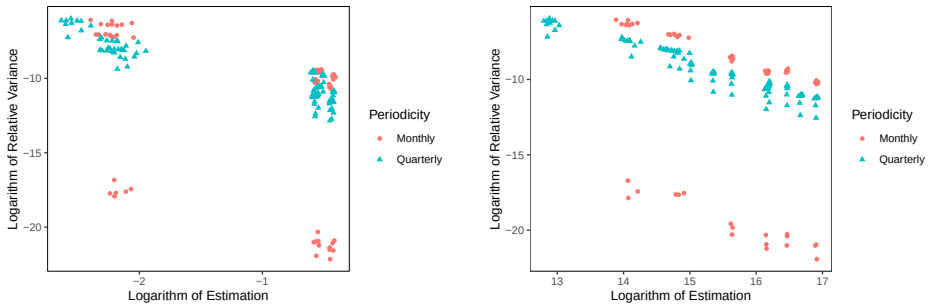
This decision was further justified by the fundamental differences between the two types of indicators. Labor market indicators, as they represent proportions within the active population, exhibit distinct patterns of relative variance and statistical properties compared to population indicators, which correspond to absolute counts of the general population. Separating the indicators into two models

allows for a more precise capture of the statistical characteristics of each group, optimizing variance adjustments and result quality. Previous studies, such as that of Salvucci et al. (1995), have demonstrated the effectiveness of this methodology in appropriately addressing methodological differences between groups of statistics.

For the construction of both models, anonymized GEIH microdata were used, which did not include variables related to the sampling design. Therefore, a design with unequal selection probabilities based on expansion factors was assumed to calculate monthly and quarterly estimates, verifying that they matched those reported in the DANE (2024*a*) bulletins. For the standard errors, those found in the DANE annexes were used. It is important to note that, due to the absence of annexes for January 2022, the monthly analysis considered only the remaining 11 months of the year.

For each group of indicators, estimates were evaluated at different levels of disaggregation and periodicity. Monthly estimates were analyzed at the national level and for the group of the 13 main cities, considering 11 available months. Quarterly estimates, covering 10 quarters[5], were assessed at four levels: national, urban, rural, and the group of the 13 main cities. This process yielded a total of $k = 3 \times (2 \times 11 + 4 \times 10) = 186$ estimates per group, enabling the analysis to capture the specific characteristics of each type of indicator.

To assess the statistical relationships between the variables, the logarithm of the relative variance $\widehat{v}_d$ was plotted against the logarithm of the estimates, differentiating by periodicity:



(a) Labor market indicators (b) Population indicators

FIGURE 1: Scatter plot of the logarithm of relative variance versus the logarithm of the estimate.

From Figure 1, it is evident that, for both labor market indicators and population indicators, there are outliers with monthly periodicity that do not align with the rest of the data. Upon analyzing these records, it was identified that they correspond to the standard errors for March, April, and May 2022. According to

[5]For the calculation of quarterly estimates, microdata files corresponding to three consecutive months (e.g., January-February-March) were integrated, directly applying the expansion factors. Estimates were obtained by summing the values of the three files and, in the case of population indicators, dividing the total by three to calculate the quarterly average. Finally, the results were compared with the data reported by DANE (2024*a*) to validate the accuracy of the calculations.

the annexes of the bulletins reported by DANE (2024*a*), these values are close to 0. Consequently, 18 outliers were identified for labor market indicators and 18 for population indicators. To account for these anomalies when constructing the GVF models, a variable was included to consider the presence of these outliers, as their origin was determined to be due to a specific phenomenon.

The statistics evaluating the fit of proposed Models 1, 2 and 3, and the theoretical model are presented in Tables A1, A2, A3 and A4, respectively, available in Appendix A. For labor market indicators, proposed Models 1, 2 and 3 exhibited similar performance in terms of fit metrics, whereas the theoretical model failed to adequately capture the data's characteristics. In contrast, for population indicators, proposed Model 3 demonstrated the best performance, achieving the lowest Root Mean Square Error ($RMSE$) and showing a good fit compared to the other models. Despite the good fit of the models and the limited number of variables used in their construction, Nagelkerke's R^2 measure, due to its uniformity, did not prove to be a differentiating factor in selecting the most suitable model.

Proposed Model 3 was selected as the most suitable for both groups of indicators due to its excellent fit and simplicity in the number of included variables. Tables 1 and B1, presented in Appendix B, detail the estimated parameters for this model. The coefficients associated with sample size ($\hat{\beta}_1$) are close to 0 in both groups, while the coefficient corresponding to the estimate ($\hat{\beta}_2$) reflects significant differences attributable to the scale of the indicators. Although some additional terms were not statistically significant, they were retained to capture residual variability and prevent potential biases, ensuring consistency with the other explanatory variables. This approach improved the model's global metrics, such as deviance and Akaike's information criterion (AIC), strengthening the quality of the fit. Furthermore, the correction factor applied to the transformation ($\widehat{\Delta}$) is approximately 1.08 for both groups, ensuring precise and consistent interpretation of the results. Collectively, these characteristics demonstrate the effectiveness of proposed Model 3 in adequately representing the variance properties of the analyzed indicators.

Table 1: Regression coefficients for numerical variables in the proposed GVF Model 3.

Indicator group	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\widehat{\Delta}$
Labor market	-16.31^{***}	-9×10^{-6} ***	-7.113^{***}	1.088
Population	-19.88^{***}	-1×10^{-5} ***	1×10^{-8}	1.085

*** Variable significant at the 0.1 % level.
** Variable significant at the 1 % level.
* Variable significant at the 5 % level.
· Variable significant at the 10 % level.
(Empty) Insignificant variable.

To obtain the estimated standard error for a particular disaggregation and indicator, the coefficients of the proposed model for that specific indicator must be used, and the square root of Equation 9 must be computed. Below are the results for each group of indicators, including an example to illustrate how the standard error prediction is performed using a GVF model.

Labor Market Indicators

Figure 2 compares the standard errors reported by DANE (2024*a*) with those obtained from the GVF models for the different models studied. The results indicate that the GVF models provide a good fit, as supported by the various evaluated measures.

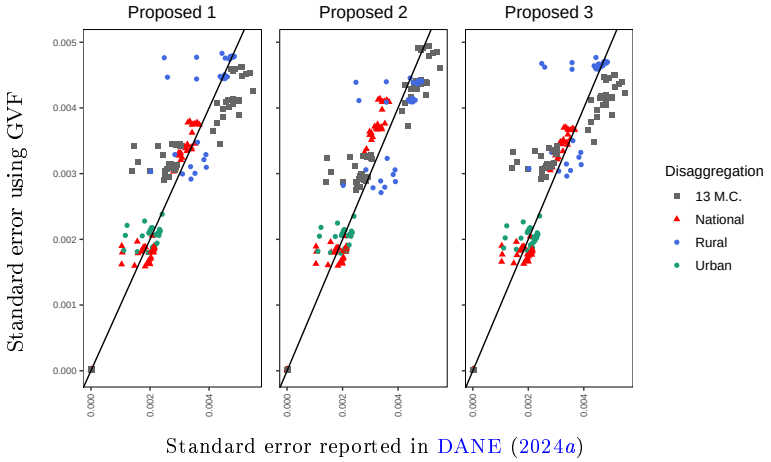


FIGURE 2: Comparison of standard errors: GVF vs. reported values in labor market indicators.

The parameter estimation for proposed Model 3, used for standard error estimation based on the GVF, is expressed as:

$$\widehat{SE}_{Prop3}\left(\widehat{\theta}_d\right) = \sqrt{\exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 n_d + \widehat{\beta}_2 \hat{\theta}_d + \widehat{\beta}_3 \text{ Indicator} + \widehat{\beta}_4 \text{ Outlier}\right) \cdot \widehat{\Delta} \cdot \widehat{\theta}_d^2} \tag{13}$$

where $\widehat{\Delta}$ equals 1.088, $\widehat{\beta}_0 = -16.31$, $\widehat{\beta}_1 = -9 \times 10^{-6}$, and $\widehat{\beta}_2 = -7.113$. The coefficient $\widehat{\beta}_3$ depends on the indicator being calculated, with specific values listed in Table B1; for non-outlier data, $\widehat{\beta}_4 = 10.91$. Additionally, n_d represents the sample size used to calculate the estimate $\widehat{\theta}$ for domain d.

Example 1. We aim to calculate the standard error of the employment rate estimate at the national level for November 2022, where the estimate is 0.57362942 and the sample size is 73 614. Using the proposed GVF Model 3 and the provided parameters, the standard error is calculated as follows:

$$\widehat{SE}\left(\widehat{\theta}\right) = \sqrt{\exp\left((-16.31) + (-9 \times 10^{-6}) \cdot 73614 + (-7.113) \cdot 0.573 + (0) + (10.91)\right)} \cdot \sqrt{(1.088) \cdot (0.573^2)}$$

$$\widehat{SE}\left(\widehat{\theta}\right) = 0.0038$$

If we compare this result to the value reported for November 2022 by DANE (2024*a*), which is 0.0035, we observe a difference of just 7.8 %. This demonstrates the model's effectiveness, even when only considering sample size, the estimate, and the indicator to be calculated.

Population Indicators

Figure 3 compares the standard errors of the estimates reported by DANE (2024*a*) with those obtained from the GVF models for a set of population indicators. It can be observed that proposed Model 3 aligns well with the reported values, consistent with the goodness-of-fit metrics for this group of indicators. In contrast, proposed Models 1 and 2 fail to align with the identity line, indicating poorer performance.

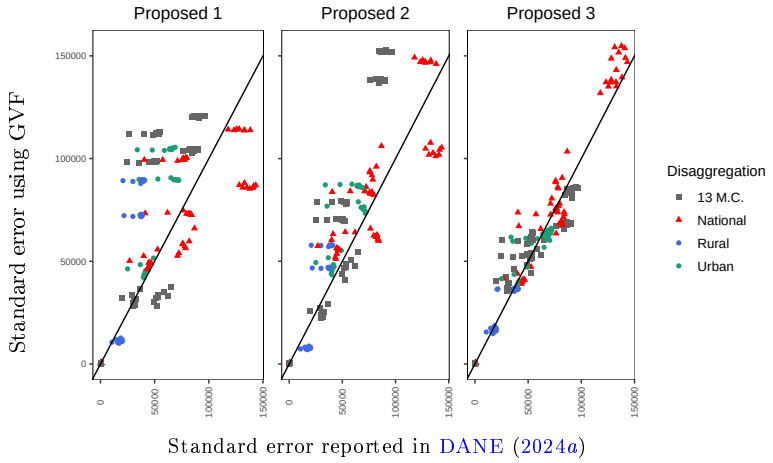


FIGURE 3: Comparison of standard errors: GVF vs. reported values in population indicators.

Substituting into Equation 13, $\widehat{\Delta}$ equals 1.085, $\widehat{\beta_0} = -19.88$, $\widehat{\beta_1} = -1 \times 10^{-5}$, and $\widehat{\beta_2} = 1 \times 10^{-8}$. The coefficient $\widehat{\beta_3}$ is calculated based on the indicator being considered, with values listed in Table B1. For non-outlier data, $\widehat{\beta_4}$ equals 10.96.

Example 2. We aim to calculate the standard error for the total number of employed people at the urban level for the July-August-September 2022 quarter, where the estimate is 17 557 565 and the sample size is 198 787. Using the proposed GVF Model 3 and the provided parameters, the standard error is calculated as follows:

$$\widehat{SE}\left(\widehat{\theta}\right) =$$

$$\sqrt{\exp\left((-19.88) + (-1 \times 10^{-5}) \cdot 198\,787 + (1 \times 10^{-8}) \cdot 17\,557\,565 + (-0.621) + (10.96)\right)} \cdot \sqrt{(1.085) \cdot (17\,557\,565^2)}$$

$$\widehat{SE}\left(\widehat{\theta}\right) = 62\,638.68$$

Comparing this result to the standard error reported by DANE (2024*a*), which is 69 516.57, we observe a difference of approximately 9 %. This demonstrates the model's effectiveness in capturing the variance properties of the analyzed indicator.

To validate the proposed GVF Model 3, developed using the anonymized GEIH 2022 microdata, data from 2023 were utilized. The model coefficients, detailed in Tables 1 and B1, enabled the generation of a total of 384 estimates: 192 for labor market indicators and 192 for population indicators. These estimates included monthly indicators, covering the 12 available months, analyzed at the national level and for the group of the 13 main cities, as well as quarterly indicators evaluated over 10 quarters at four levels of disaggregation: national, urban, rural, and the group of the 13 main cities. During this process, no estimates were identified as outliers. The calculated results were subsequently compared with the values reported by DANE (2024*a*) for 2023, yielding the following metrics:

Table 2: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Efron's pseudo R^2 for the validation of the proposed Model 3.

Indicator group	RMSE	MAE	MAPE	R^2 Efron
Labor market	0.0003	0.0003	9.809	86.5 %
Population	10 039.030	7 050.592	10.330	89.6 %

According to Table 2, the MAPE for both models (the GVF model for labor market indicators and the GVF model for population indicators) is approximately 10 %. This value is considered acceptable, indicating reasonable accuracy in the predictions of both models. Regarding Efron's pseudo R^2, both models exhibit results exceeding 85 %, suggesting that they are capable of explaining a significant proportion of the variability in the data.

To enable a comparison, the study conducted by Gutiérrez & Babativa-Márquez (2023) for CEPAL proposed a GVF model to estimate the standard errors of social and labor indicators in Colombia. However, as previously mentioned, this model did not include GEIH microdata in its initial formulation due to the absence of information on sampling design variables. This limitation prevented the proper estimation of the direct variances required to construct the model's foundations. Although this issue was also present in the current work, a methodological alternative was proposed to utilize anonymized GEIH microdata by validating and adjusting the standard errors through comparisons with the values reported by DANE, ensuring the estimates matched the official figures exactly.

Both studies overlapped exclusively in the evaluation of the unemployment rate and employment rate indicators, as these were the only shared values between them, despite each study considering additional indicators. However, significant differences were identified in the periodicities and domains analyzed. In the CEPAL study, estimates were produced annually and disaggregated by domains such as sex, education level, area, age and ethnicity. In contrast, the present study focused on broader domains, including the national level, main cities, urban areas, and rural areas, adopting monthly and quarterly periodicities. These methodological differences make a direct comparison of the results from both studies challenging.

Moreover, in this study, when comparing the estimates of the shared indicators, differences in the observed minimum sample sizes were noted. For the unemployment rate, the smallest estimate was derived from a sample of 21 206 observations, while for the employment rate, the smallest estimate was calculated using a sample of 28 585 observations. Conversely, in the CEPAL study, a minimum threshold of 400 records was established, above which the smoothed variance was equalized to the direct variance, based on the assumption that larger sample sizes are sufficiently representative. However, this assumption is questionable, as the absence of sampling design variables can result in unreliable estimates, even with representative sample sizes.

This limitation became evident when applying the GVF model developed by Gutiérrez & Babativa-Márquez (2023) to the mentioned indicators, where the coefficients associated with sample size were negative. Due to the logarithmic transformation employed in the models, the standard errors decreased drastically in domains with large sample sizes, even resulting in zero variance estimates. This finding contrasts with the results reported by DANE and highlights a limitation of the CEPAL model in adequately handling domains with large sample sizes.

In conclusion, although both the study conducted by Gutiérrez & Babativa-Márquez (2023) and the present work share the objective of estimating the standard errors of indicators based on the GEIH, differences in methodology and data usage make a direct comparison impractical. The CEPAL model, by not incorporating the specific microdata from the GEIH of 2022 and adopting a generalized approach, demonstrates limitations when adapting to the Colombian context, particularly in domains with large sample sizes. In contrast, the present study proposed a methodology that integrates GEIH microdata with the validation of standard errors using values reported by DANE. This approach enabled the generation of more representative estimates for the national context, albeit applied to broader disaggregations compared to CEPAL's more granular model.

5. Conclusions

This study focused on evaluating Generalized Variance Function (GVF) models to calculate the standard error of key indicators in the "Gran Encuesta Integrada de Hogares" (GEIH). Four GVF models were tested for two groups of indicators: labor market indicators and population indicators. Among these, proposed Model 3 proved to be the most effective based on the evaluated fit metrics, demonstrating consistency when compared to the official errors published by DANE (2024*a*). This model simplifies implementation by calculating the standard error using only the indicator estimate, sample size, and type of indicator being evaluated. It is important to note that the model assumes that the outlier variable will not present unusual values during its application.

Therefore, the main contribution of this study lies in the incorporation of categorical variables into the GVF models, enabling more precise adjustments tailored to the characteristics of the available data. Additionally, organizing the indicators into specific groups facilitated the development of robust and adaptable GVF models suitable for various types of statistics.

Given the lack of access to specific sampling design variables in the anonymized microdata provided by the different National Statistics Institutes, this study proposed an alternative methodology. This approach involved replacing the standard errors calculated under a simplified design (without the original design variables) with the official errors published by DANE, which consider a complex sampling design. This methodology addressed the absence of key information in the microdata, enabling the construction of GVF models better aligned with the survey's sampling design.

The application of this methodology to the GEIH 2022 data and its validation with 2023 results demonstrated appropriate fits, supported by metrics such as RMSE and MAE. Furthermore, practical examples were included to illustrate its implementation, making this study a valuable reference for similar contexts and surveys where anonymized microdata lack sampling design variables.

In summary, the GVF models developed in this study are easy to apply and effective when calculating standard errors for the evaluated indicators, maintaining consistency with the values published by DANE. One of their main advantages is their ability to estimate standard errors with monthly periodicity, even for disaggregations such as rural and urban areas, whose official values are only available quarterly in the annexes of DANE reports. This feature broadens the scope of estimates, providing more detailed and frequent information, which is particularly valuable for the continuous monitoring of indicators. However, given that the initial construction of these models depends on the availability of official standard errors, additional validations are recommended, particularly in domains with small sample sizes or unique characteristics, before a generalized implementation. This approach would ensure that the estimates remain robust and reliable in broader applications.

Finally, this study lays the foundation for developing more robust models, especially if complete sampling design variables for the GEIH become available in the future. Access to this information would significantly expand the model's applicability, enabling more detailed disaggregations, such as specific subnational levels, analyses of vulnerable population groups, or domains with smaller sample sizes than those considered in this study. Furthermore, the potential implementation of a Longitudinal Generalized Variance Function (LGVF) model is proposed, incorporating data from multiple years of the GEIH, following the approach of Zhang et al. (2019). Such an extension would strengthen the analytical capabilities for larger-scale studies, offering a comprehensive framework for variance estimation over time.

[Received: October 2024 — Accepted: January 2025]

References

Alegria, J. & Scott, T. C. (1991), *Generalized variance function applications in forestry*, Vol. 345, US Department of Agriculture, Forest Service, Northeastern Forest Experiment.

Carter, G. & Rolph, J. (1974), 'Empirical bayes methods applied to estimating fire alarm probabilities', *Journal of the American Statistical Association* **69**, 880–885.

DANE (2024*a*), 'Mercado laboral (empleo y desempleo) históricos', `https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-y-desempleo/geih-historicos`. Last accessed 25 Jun 2025.

DANE (2024*b*), 'Población fuera de la fuerza laboral', `https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/poblacion-fuera-de-la-fuerza-laboral`. Last accessed 25 Jun 2025.

Fay, R. E. & Herriot, R. A. (1979), 'Estimates of income for small places: an application of james-stein procedures to census data', *Journal of the American Statistical Association* **74**(366a), 269–277.

Fúquene-Patiño, J., Cristancho, C., Ospina, M. & Morales Gonzalez, D. (2021), 'Fay-herriot model-based prediction alternatives for estimating households with emigrated members', *Journal of Official Statistics* **37**(3), 771–789. https://doi.org/10.2478/jos-2021-0034

Gutiérrez, A. & Babativa-Márquez, G. (2023), Efectos de diseño para indicadores sociales en américa latina: función generalizada de varianza para estimadores directos provenientes de encuestas de hogares, Serie de la CEPAL 67980, Comisión Económica para América Latina y el Caribe (CEPAL).

Gutiérrez, A., Fuentes, A., Mancero, X., López, F. & Molina, F. (2020), Criterios de calidad en la estimación de indicadores a partir de encuestas de hogares: una aplicación a la migración internacional, Technical report 101, Comisión Económica para América Latina y el Caribe (CEPAL), Santiago, Chile. https://repositorio.cepal.org/handle/11362/45681

Handayani, A. & Aunuddin, I. (2005), Generalized variance functions for binomial variables in stratified two-stage sampling, *in* 'Forum Statistika dan Komputasi', Indonesia, pp. 1–8.

Johnson, E. G. & King, B. F. (1987), Generalized variance functions for a complex sample survey, Research Report RR-87-06, Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1987.tb00210.x

Krenzke, T. (1995), Reevaluating generalized variance model parameters for the national crime victimization survey, *in* 'Proceedings of the Section on Survey Research Methods, American Statistical Association', Alexandria, VA, USA, pp. 327–332. Published in the ASA 1995 conference proceedings.

Kubacki, J. & Jędrzejczak, A. (2011), The comparison of generalized variance function with other methods of precision estimation for polish household budget survey, *in* 'Proceedings of the 7th Conference "Survey Sampling in Economic and Social Research"', University of Economics in Katowice, Katowice, Poland, pp. 58–69.

Lohr, S. L. (2021), *Sampling*, Chapman and Hall/CRC.

McIllece, J. (2018), 'On generalized variance functions for sample means and medians. jsm 2018–survey research methods section, 584–594'.

MinSalud (2024), 'Definiciones del mercado laboral', `https://www.minsalud.gov.co/trabajoEmpleo/Paginas/definiciones.aspx`. Last accessed 25 Jun 2025.

Morales, D., Dolores Esteban, M., Pérez, A. & Hobza, T. (2021), *A Course on Small Area Estimation and Mixed Models: Methods, Theory and Applications in R*, Statistics for Social and Behavioral Sciences, Springer.

Salvucci, S., Weng, S. & Holt, A. (1995), *Design Effects and Generalized Variance Functions for the 1990–91 Schools and Staffing Survey (SASS): User's Manual*, Vol. 1 of *NCES Technical Report Series*, U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, Washington, DC. https://nces.ed.gov/pubs95/95342.pdf

Valliant, R. (1987), 'Generalized variance functions in stratified two-stage sampling', *Journal of the American Statistical Association* **82**(398), 499–508.

Wolter, K. (2007), *Introduction to Variance Estimation: Statistics for Social Science and Behavorial Sciences*, Springer.

Zhang, G., Cheng, Y. & Lu, Y. (2019), 'Generalised variance functions for longitudinal survey data', *Statistical Theory and Related Fields* **3**(2), 150–157.

Appendix A. Goodness-of-Fit Measures for Models

TABLE A1: Measurements of Akaike's information criterion (AIC), Bayesian information criterion (BIC), deviation, RMSE and Nagelkerke's pseudo R^2 for the proposed Model 1.

Indicator group	AIC	BIC	Deviance	RMSE	R^2 Nag
Labor market	203.257	219.386	30.78	0.007	99.99 %
Population	512.408	528.536	162.225	439 635.7	99.99 %

Table A2: Measurements of Akaike's information criterion (AIC), Bayesian information criterion (BIC), deviation, RMSE and Nagelkerke's pseudo R^2 for the proposed Model 2.

Indicator group	AIC	BIC	Deviance	RMSE	R^2 Nag
Labor market	190.457	209.811	28.426	0.006	99.99 %
Population	439.283	458.637	108.32	33 6375.9	99.99 %

Table A3: Measurements of Akaike's information criterion (AIC), Bayesian information criterion (BIC), deviation, RMSE and Nagelkerke's pseudo R^2 for the proposed Model 3.

Indicator group	AIC	BIC	Deviance	RMSE	R^2 Nag
Labor market	205.695	228.275	30.523	0.007	99.99 %
Population	200.562	223.142	29.692	133 200.4	99.99 %

Table A4: Measurements of Akaike's information criterion (AIC), Bayesian information criterion (BIC), deviation, RMSE and Nagelkerke's pseudo R^2 for the theoretical model.

Indicator group	AIC	BIC	Deviance	RMSE	R^2 Nag
Labor market	961.156	974.059	1830.542	0.258	94.13 %
Population	953.789	966.692	1759.451	4 672 747	91.88 %

Appendix B. Regression Coefficients for Categorical Variables

Table B1: Regression coefficients for categorical variables in proposed GVF Model 3.

Variable	**Categories**	**Indicator group**	
		Labor market	**Population**
Indicator $\widehat{\beta_3}$	Unemployment rate	-0.340	-
	Employment rate	0^{***}	-
	Global participation rate	0.143	-
	Total number of employed people	-	-0.621^{***}
	Total number of unemployed people	-	2.734^{***}
	Population outside the labor force	-	0^{***}
Outlier $\widehat{\beta_4}$	No	10.91^{***}	10.96^{***}
	Yes	0^{***}	0^{***}

*** Variable significant at the 0.1 % level.
** Variable significant at the 1 % level.
* Variable significant at the 5 % level.
· Variable significant at the 10 % level.
(Empty) Insignificant variable.