International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 13 (2016) pp 8031-8036

Simulation Study for Evaluation of Weighting Methods in National Socio-Economics Survey (Susenas)

Ary Santoso[1], Anang Kurnia[2*], Bagus Sartono[3] and Siti Muchlisoh[4]
Departemen of Statistics Bogor Agricultural University,
Jalan Pajajaran, Kampus IPB Baranangsiang, Bogor 1615, Indonesia.

**corresponding author: Anang Kurnia*
Corresponding author name: Anang Kurnia
Meranti wing 22 level 4 IPB Dramaga, Bogor, West Java

Abstract
Statistics Indonesia (a.k.a. BPS) utilizes the generalized least squares (GLS) approach to determine sample weights in their survey data, including National Socio-Economics Survey (Susenas). The weakness of this approach is the possibility of the appearance of weird weights such as negative values and extremely large values. This present paper discusses evaluating techniques of GLS to weight data. Two weigthing methods approacheswhich are examined are weighting by the sampling design and weighting by the use of generalized least squares.A simulation studywas conducted to explore the characteristics of weight design and GLS as well asto solve the problem of negative weights of GLS. The study revealed that the negative values of GLS can be solved by adjustment matrix of auxiliary variable (X).

Keywords: Sample Weights, Weighting by the sampling design, Weighting by Generalized Least Square (GLS)

INTRODUCTION

National Socio-Economics Survey (Susenas) is a survey with household approach conducted by BPS and aims to obtain data of socio-economic characteristics of the population especially those relate to the measurement of the level of prosperity. Since 1993 until now, Susenas enumeration period has changed, from an annual, semiannual and in 2011 implemented quarterly (March, June, September, and December). Although a change of the enumeration period, a sampling method of probability Proportional to Size (PPS) with household as the size and Systematic Random Sampling (SyRS) are still used, conducted in three-stages (multistage sampling).

Susenas is used to estimating population parameter.BPS utilizes weight for estimating population from the Susenas.Since, in a sample survey, we are estimating the value of a characteristic for the whole population on the basis of data on a part of it selected as sample, the sample observations are weighed with certain weights for obtaining an estimate of the population parameters (Murthy 1967). It shows that sample weight plays important role in a sample survey. Herlawati et al (2013) studied weighted sampling for stratified random sampling.

Before 2011, BPS used weights by design. After that, BPS utilizes weights by model (GLS). But, in the formula of weights GLS contains weights by design. Weights by design (a.k.a. Ω) was obtained by sampling design formed. The value of Ω reverses of multiplication of probability in each stage of sampling. GLS was developed by Zieschang (1990). GLS weighting methods is a calibration weighting method using distance function of chi-square by utilizing additional information. Zieschang (1990) showed weighting GLS is possibility of the appearance of weird weights such as negative values and extremely large values. Because of the problem, study of sample weights is important. The study would explore the characteristics of weight design and GLS as well assolve the problem of negative weights of GLS.

PRELIMINARY THEORY

Sample Weights

Murthy (1967) mentioned that the weights is the value used to assess the observation of the sample under consideration (selected sample). Weights without adjustment is weights which is obtained based on sampling design(Ω). The formula of basic sample weights (weight without adjustment) is:

$$\Omega_i = \frac{1}{\mathrm{fs}_i} \quad (1)$$

Ω_i = weights of the sample unit i-th

fs_i = sampling fraction of the sample unit i-th

Murthy (1967) also mentioned that ΣW_i is unbiased estimator for the total individual of the population (ΣW_i = N). Weighting with adjustment is weights obtained in accordance with the conditions and characteristics of the sample, as well as added by using ancillary information.

Sampling Method of National Socio-Economics Survey (Susenas)

Determination of province and district/city in Susenas was fixed. At the level of district/city were stratified into urban and rural. There are primary sampling unit (PSU), census block (BS), and households (HH) in each strata. PSU is enumeration area consisted of adjacent several census blocks (BS). Susenas is designed based on three stages sampling design. First, selecting primary sampling unit (PSU) by probability proportional to size (PPS) sampling. Second, selecting census block (BS) by PPS sampling from the each PSU that was selected at the first stage. Third, selecting household by linear systematic sampling (SyRS) from the each census block that was selected at the second stage.

Weighting Based on Susenas Design in 2011

Design weights (Ω) is determined based on sampling design conducted. It is obtained by calculating the probability and sampling fraction of each stage sampling first. Formula of probability of each stage shown in Tables 1 to 3.

Table 1: Sampling of PSU in district/city *d* strata *s*

Unit	The number of units in the district/city *d* strata *s*		Sampling method	Probability
	Population	Sample		
PSU	N_{ds}	n_{ds}	*Ppss*	$\frac{M_{dsi}}{\sum_{i=1}^{N_{ds}} M_{dsi}}$

Table 2: Sampling of census block (BS) in each PSU

Unit	The number of units in PSU *i* district/city *d* strata *s*		Sampling method	Probability
	Population	Sample		
Census block	B_{dsi}	1	*Ppss*	$\frac{M_{dsij}}{M_{dsi}}$

Table 3: Sampling of household (HH) in each BS

Unit	The number of unit in BS *j* PSU *i* district/city *d* strata *s*		Sampling method	Probability
	Population	Sample		
Household (without updating data)	M_{dsij}	10	*Systematic Random Sampling*	$\frac{1}{M_{dsij}}$
Household(with updating data)	M^{u}_{dsij}	10	*Systematic Random Sampling*	$\frac{1}{M^{u}_{dsij}}$

Based on Table 1 to 3 then the design weights (Ω) at of household level is

Total of probability of sampling

$$= \frac{M_{dsi}}{\sum_{i=1}^{N_{ds}} M_{dsi}} \text{ x } \frac{M_{dsij}}{M_{dsi}} \text{ x } \frac{1}{M_{dsij}} = \frac{1}{\sum_{i=1}^{N_{ds}} M_{dsi}} \quad (2)$$

Sampling fraction (fs)

$$= (10 \text{ x } n_{ds}) \text{ x } \frac{1}{\sum_{i=1}^{N_{ds}} M_{dsi}} \quad (3)$$

So that, weight of household on BS *j* PSU *i* strata *s* in the district/city *d* is

$$\Omega_{dsij} = \frac{1}{fs}$$

$$\Omega_{dsij} = \frac{\sum_{i=1}^{N_{ds}} M_{dsi}}{10\, n_{ds}} \quad (4)$$

Where

Ω_{dsij} = Weight of household in cencus block*j* PSU i district/city *d* strata *s*,

M_{dsi} = Total of household in PSU *i* (from sampling frame) district/city *d*strata *s*,

M_{dsij} = Total of household in census block*j* PSU *i* (from sampling frame) district/city *d*strata *s*,

M^{u}_{dsij} = Total of household from updating data in cencus block*j* PSU *i*district/city *d*strata *s*,

n_{ds} = Total of sample census block in district/city *d*strata *s*.

Above formula of Ω_{dsij} is weight design which is data of HH without updated data (ideal condition). But in fact, BPS often calculates the weights of design based on HH data with updated data and the formula as follows:

$$\Omega^{*}_{dsijk} = \frac{M^{u}_{dsij} \sum_{i=1}^{N_{ds}} M_{dsi}}{10\, n_{ds} M_{dsij}} \quad (5)$$

Weighting Generalized Least Square

Zieschang (1986) told that GLS in survey weight was usedfirst by Luery in 1980 in Current Population Survey (CPS). The GLS procedure adjusts the sample weights from prior stages of weighting by minimizing the weighted squared adjustments subject to a set of linear 'control' constraints the adjusted weights must satisfy.

The goal is to minimize

$f(w) = min_W (\Omega - W)^T \Lambda^{-1} (\Omega - W)$

subject to $X^T W = P_x$. The solution of this problem yields

$$\widehat{W} = \Omega + \left(\Lambda X (X^T \Lambda X)^{-1} (P_x - X^T \Omega)\right) \quad (6)$$

where

Ω = n x 1 vector of design sample weights for a sample of n sample units, representing the inverse of the design probability of selection, π, from a population of unit;

X = K x n matrix of control characteristics of each sample unit whose aggregate population values are known with certainty, such as number of person in K cells defined by age, race, and sex in each consumer unit;

P_x = K x1 vector of control counts of aggregate population values of characteristics X that are taken to be known with certainty, such as number of persons in the population on cells defined by age, race, and sex;

W = n x 1 vector of adjusted weights; and

Λ = n x n weighting matrix. Both Luery (1980) and Roman (1982) in Zieschang (1990) assumed either Λ=diag(Ω).

MATERIALS AND METHODS

Data

This study utilized secondary data, both the 2010 Indonesia population cencus (SP2010) and the 2011 national socio-economics survey (Susenas 2011) data. They were obtained by Statistics Indoensia (BPS). The data SP2010 was used as populaiton data on simulaiton process. To apply calculation of weights which referring to simulation study utilized Susenas data 2011.Variables (profiles) which would be used in the simulation were chosen subjectively by researchers. they were variable of gender, religion, and education.

Simulation

Steps used in this study were:

1. Exploration of calculating design weights (Ω) and GLS (W).
   a. Tabulate the data of SP2010 as population data
   b. Sampling of the population data with some variations sampling design, that are:

-sampling 1-stage: Simple Random Sampling (SRS)
-sampling 2-stage: SRS-SRS
-sampling 3-stage: SRS-SRS-SRS

c. Arrange Px from the population data (1a) using the characteristics of the age group and gender.
d. Arrange X from the sample data (1b) using the characteristics of the age group and gender.
e. Calculate design weight (Ω) of each sampling design in step b
f. CalculateΛ, where $\Lambda = diag\ (\Omega)$
g. Estimate GLS weight (W) of each sampling design in step b with formula

$$\widehat{W} = \Omega + \left(\Lambda X(X^T \Lambda X)^{-1}(P_x - X^T \Omega)\right)$$

h. Evaluate steps e and g

2. Estimating design weights (Ω) and GLS (W) use Susenas sampling design (PPS, PPS, and Systematic Random Sampling).
   a. Arrangethe data of SP2010 as population data and choose the variables gender, religion, and education in population data
   b. Sampling the population data with Susenas design sampling
   c. Arrange Px from the population data (1a) using the characteristics of the age group and gender.
   d. Arrange X from the sample data (1b) using the characteristics of the age group and gender
   e. Calculatedesign weight (Ω)
   f. CalculateΛ, where $\Lambda = diag\ (\Omega)$
   g. Calculate$\widehat{W} = \Omega + \left(\Lambda X(X^T \Lambda X)^{-1}(P_x - X^T \Omega)\right)$
   h. Repeat steps (b) until (g) with 1000 repeatation
3. Estimate variables profile population used sample weight from steps 1e, 1g, 2e, and 2g
4. Evaluate the result of estimation profil population step 3.

Data Analysis for Case Study of Susenas Data 2011

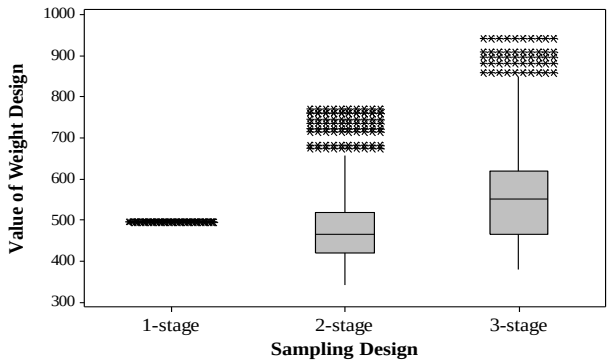
Steps used in this study were:

1. Tabulate Susenas data 2011
2. Arrange Px and X refer to result of simulation study
3. Calculate Ω
4. Calculate $\widehat{W}$
   - Weights calculation referred to BPS technique
   - Weightscalculation referred to modification technique which was obtained by simulation study

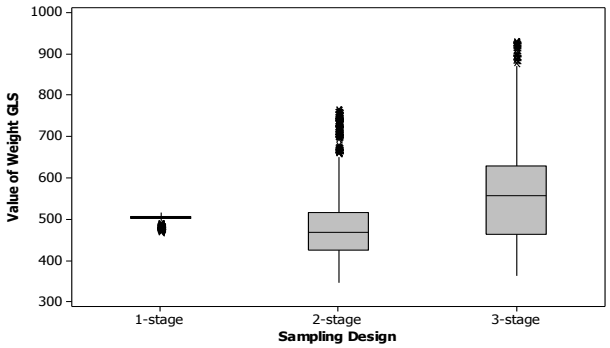
RESULTS AND DISCUSSION

Exploration of calculation sample weights in level of Households (HH)

Exploration results in the level of HH were presented in Figures 1a and 1b show that the number of stages of sampling affects the results of the sample weights both weights Ω and W. In Figure 1a, the 1-stage sampling design gave weight value relatively constant and Ω did not have a high level of diversity. Similarly in Figure 1b, the value W of the 1-stage sampling design less diverse and weight values of W between sample unit was not extremely large values. In addition, the weight values of GLS (W) obtained negative value (Figure 2). A negative value was only possessed by weight of W, while the weight Ω was always positive values.

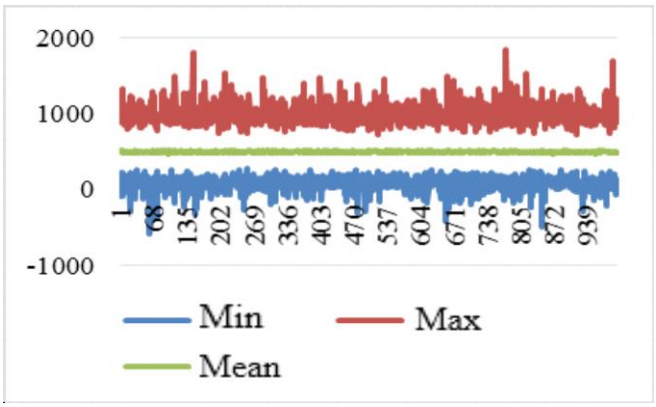


(a)

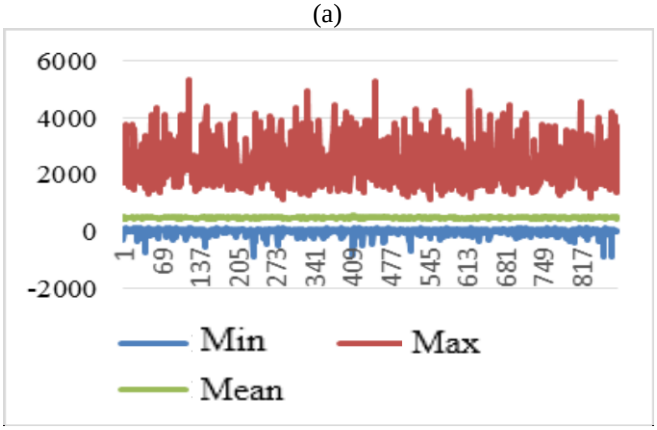


(b)

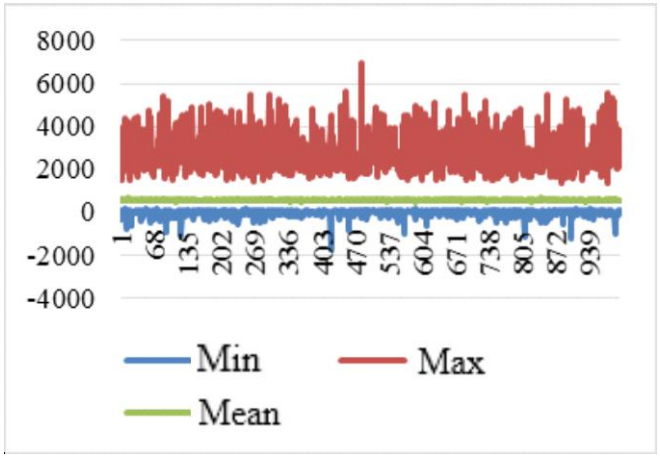
Figure 1: Sample weights based on sampling design 1-stage SRS, 2-stage SRS, and 3-stage SRS in level of HH: (a) Weights by design (Ω), (b) Weights by GLS (W)



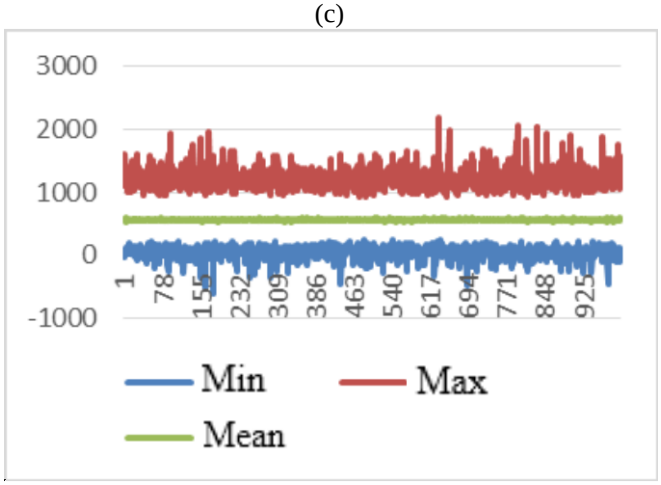
(a)



(b)



(c)



(d)

Figure 2: Weights GLS (W) in level of HH with ordinary matrix X: (a) 1-stage design of SRS, (b) 2-stage design of SRS, (c) 3-stage design of SRS, (d) Susenas Design

The exploration above described that the characteristics of weights value of both Ω and W were in line with research which had been done by zieschang in 1990. The characteristics of weights value of GLS could be negative and have extremely large values between sample unit. This condition still faced by Statistics Indonesia (BPS) when using GLS as the Susenas survey weighting method

Negative value of GLS weigths (W)

There were some components that caused value of W. Mathematically,the causal component of negative value of W was correction factor $[\Lambda X(X^T\Lambda X)^{-1}(P_x - X^T\Omega)]$ also known as C. Weights GLS (W)was negative when the score of C was higher negative than score of Ω. Parts of components of C which could be negative were $(X^T\Lambda X)^{-1}$also known as A and$(P_x - X^T\Omega)$also known as B. The component of C would be negative if both A and B were different signs each other.

The exploration above gives the concerning to the matrix of auxiliary variable X. When calculation weights in level of households (HH), arranging of matrix X.with BPS method would provide non binary elements of matrix X (Table 5). Matrix X in HH level was obtained by aggregation of matrix X which was arranged in individual level (Table 4).Elements of matrix X which were more than 1 (non binary) would be great multiply component in component of A. Whereas, element of A consisted of various value (positive of negative). It means thatif the element of X is high number,the component of A will be high value (positive or negative). Besides that,component of B can be also high negative value if $X^T\Omega > P_x$.The works of exploration drived in the justification that component of C could be negative more than value of Ω when matrix X was not arranged binery. One also justified that the component caused negative value of W was matrix X which was not binary.

Table 4: Illustration ofmatrix X for calculation weights in individual level

HH	Personal Number	Sex	Age	X											
				Male (Sex=1)						Female (Sex=2)					
				0-4	5-9	...	25-29	...	>75	0-4	...	20-24	...	>75	
1	1	1	25	0	0	0	1	0	0	0	0	0	0	0	
1	2	2	24	0	0	0	0	0	0	0	0	1	0	0	
1	3	1	7	0	1	0	0	0	0	0	0	0	0	0	
1	4	1	5	0	1	0	0	0	0	0	0	0	0	0	
2	1	1	25	0	0	0	1	0	0	0	0	0	0	0	
2	2	1	4	1	0	0	0	0	0	0	0	0	0	0	
2	3	1	1	1	0	0	0	0	0	0	0	0	0	0	

Table 5: Illustration ofmatrix X for calculation weights in households level

HH	Personal Number	Sex	Age	X											
				Male (Sex=1)						Female (Sex=2)					
				0-4	5-9	...	25-29	...	>75	0-4	...	20-24	...	>75	
1	1	1	25	0	2	0	1	0	0	0	0	1	0	0	
2	1	1	25	2	0	0	1	0	0	0	0	0	0	0	

Recomendation of matrix X arranging for calculation GLS weights (W) in households level

Justification of arranging matrix X was elements of matrix ought to be binary (0 and 1). Thus, one needs to be adjusted so that the matrix X which was arranged has binary value of the elements (0 and 1) when work in households level, then called X adjustment. Matrix X adjustment was formed into categories according to the subjectivity of the researcher. Categorizing ought to be accommodated all the conditions of sample which was determined the characteristics. This research, X adjustment consisted of several categories as follows:

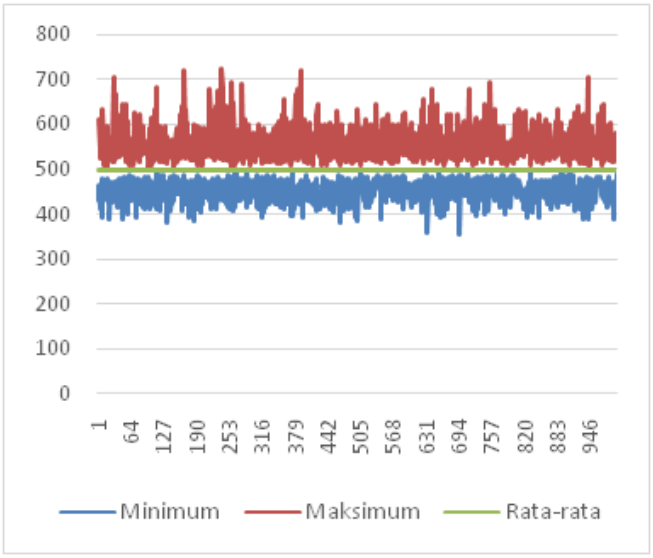
1. No member of the households (HH)<15 years old (category 1)
2. Member of households <15years oldand all males (category 2)
3. Member of households <15years old and all females (category 3)
4. Member of households <15years old consisted of males and females (category 4)

As an illustration, if the condition HH, members of HH, sex, and age of each member of the household such as in Table4, the composition of the matrix X adjustment based on the above categories are presented in Table6.

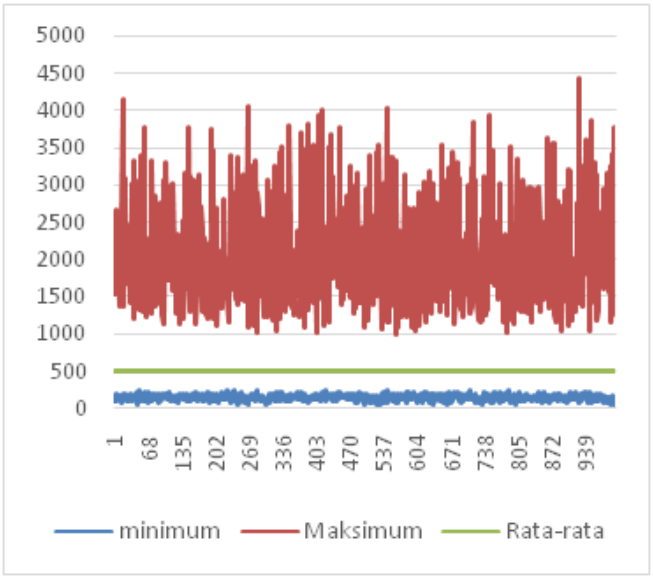
Tabel 6: Illustration of matrx X adjustment

RT	ART	JK	Usia	X			
				kategori 1	kategori 2	kategori 3	kategori 4
1	1	1	25	0	0	0	1
2	1	1	25	0	1	0	0

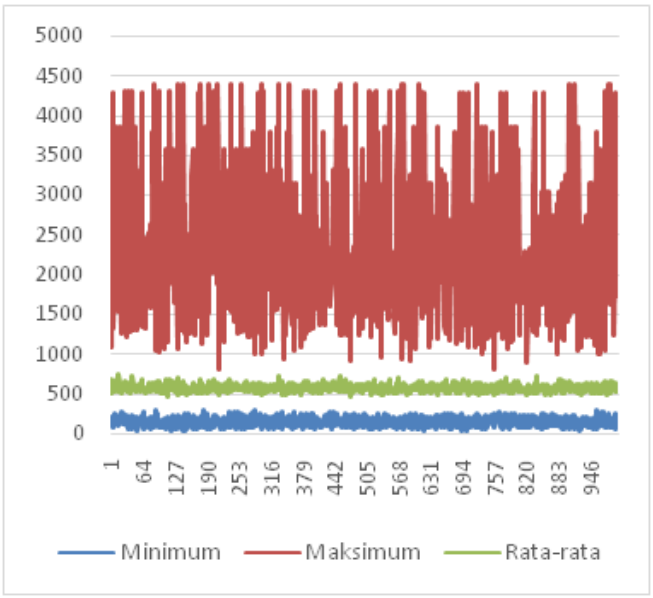
The results of the application of the matrix X adjustment (as illustrated above) on the design exploration (1-stage, 2-stage, and 3-stage of SRS) showed that the weights W which was worked at the households level did not produce a negative value (Figure 3). However, the diversity of the sample weights were still high and had extremely large distance between sample unit in sampling design of 2-stage and 3-stage of SRS.



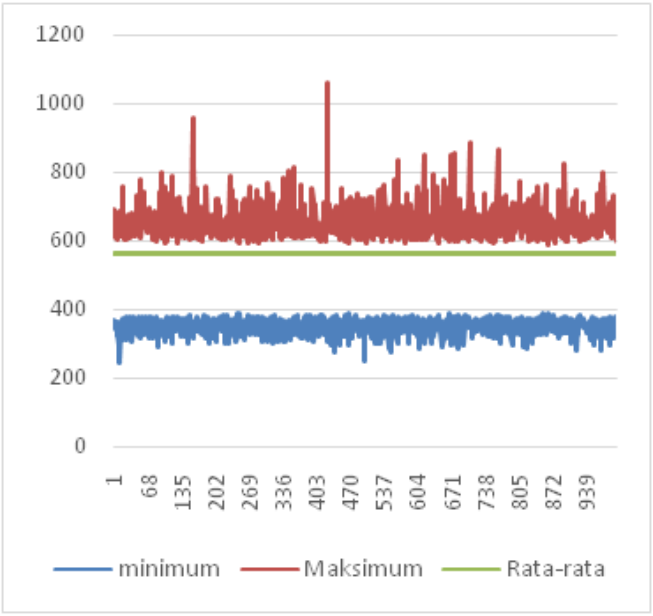
(a)



(b)



(c)



(d)

Figure 3: Weights GLS (W) in level of HH with matrix X adjustment: (a) 1-stage design of SRS, (b) 2-stage design of SRS, (c) 3-stage design of SRS, (d) Susenas Design

Simulation of calculation weights of Susenas with matrix X adjustment

Weights Susenas which was worked in households level with X unadjustment shoed in Figure 2d. Calculation weight Susenas in this discussion was the weight Susenas calculated in householsd level and utilized matrix X adjustment. all values of GLS weight (W) of Susenas obtained by using X adjustments showed positive values in 1000 replications (Figure 3d) and the diversity of the weights Ω and W is not so great, inspite of the presence of outlier (Figure4).

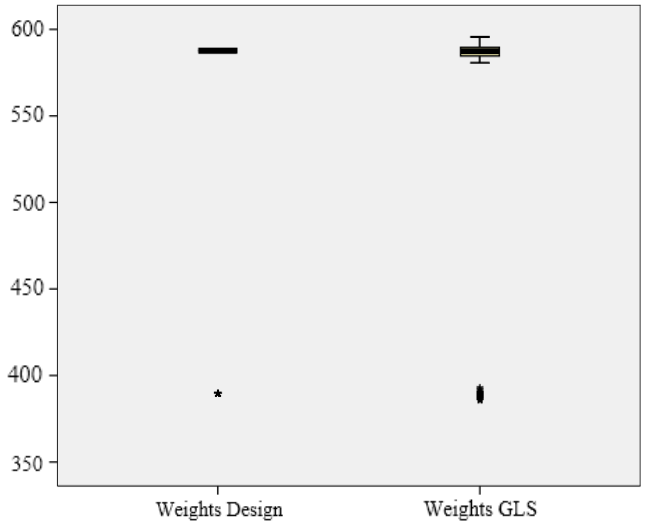


Figure 4: Weights of design Susenas in households level with matrix X adjustment

CONCLUSSION AND REMARKS

The values of GLS weight which was calculated in the household level could be negative and extremely large distance between unit sample. The negative values could be answered through justify adjustments to the arranging of the matrix auxiliary variable X into a matrix that contained the value 0 and 1 (binary).

ACKNOWLEDGMENT

I would like to thank you to Statistics Indonesia (BPS) for providing the 2010 Indonesia population cencus (SP2010) and the 2011 national socio-economics survey (Susenas 2011) data.

REFERENCES

[1] BPS. 2011. Buku Pedomen Susenas untuk Kepala BPS. Jakarta.

[2] Braun JW, Murdoch JD. 2007. A First Course in Statistical Programming With R. New York: Cambridge University Press

[3] Friedman EM, Jang D, Williams TV. 2002. Combined Estimates From Four Quarterly Survey Data Sets. *Proceedings of the Section on Survey Research Methods*. *American Statistical Association.* 1064-1069.

[4] Herlawati I, Kurnia A, Afendi FM. 2013. Penentuan Nilai Pembobotan dan Penduga Ragam untuk Penarikan Contoh Bertahap. *Xplore*. 1(1): e7(1-8).

[5] Murthy NM. 1967. Sampling Theory and Methods. Calcutta: Statistical Publishing Society.

[6] Zieschang KD.1990. Sample Weighting Methods and Estimation of Totals in the Consumer Expenditure Survey. *Journal of the American Statistical Association.* 85: 986-1001.