

OUTLIER ROBUST GENERALIZED REGRESSION ESTIMATOR

Hyunshik Lee¹ and Zdenek Patak²

ABSTRACT

The problem of outliers is common in sample surveys, particularly in business surveys. The Generalized Regression Estimator (GREG) is now well accepted and widely used in survey sampling when auxiliary data are available. However, the GREG estimator is vulnerable in the presence of outliers. Many robust procedures have a smaller variance than nonrobust procedures when outliers are present in the data set. However, the bias can be dominant in the mean square error when the sample size becomes large. These procedures are not consistent. In this paper we propose an outlier robust GREG estimator that is consistent, yet more efficient in terms of MSE than ordinary GREG estimator in outlier situations. Some simulation results that demonstrate this are also presented.

KEY WORDS: Consistency; Bias reduction; Adaptive method; Mean square error.

RÉSUMÉ

Le problème des valeurs aberrantes est commun dans les enquêtes par sondage, particulièrement pour les enquêtes entreprises. L'estimateur de régression généralisé (GREG) est maintenant bien accepté et largement utilisé dans les enquêtes par sondage lorsque des données auxiliaires sont disponibles. Toutefois, l'estimateur GREG est vulnérable en présence de valeurs aberrantes. Plusieurs procédures robustes ont une variance inférieure aux procédures non robustes en présence de valeurs aberrantes. Cependant, le biais peut être dominant dans l'erreur quadratique moyenne (EQM) lorsque la taille de l'échantillon devient grande. Ces procédures ne sont pas convergentes. Cette communication propose un estimateur GREG robuste par rapport aux données aberrantes, convergent et encore plus efficace en termes d'EQM que l'estimateur GREG ordinaire, en présence de situations avec valeurs aberrantes. Quelques simulations démontrant ces résultats seront aussi présentées.

MOTS-CLÉS: convergence; réduction du biais; méthode adaptative; erreur quadratique moyenne.

1. INTRODUCTION

The problem of outliers is common in sample surveys, particularly so in business surveys. Known outliers can be dealt with by using an efficient sample design. We often use stratification by size to enhance the efficiency of the sample design, and also to alleviate potential outlier problems using a size measure that is highly correlated with the survey variables. However, this alone is not enough to prevent the occurrence of outliers. When they occur, they exert undue influence on the estimates so as to make them very unreliable or sometime useless.

The Generalized Regression Estimator (GREG) is now well accepted and widely used in survey sampling when auxiliary data are available.

Statistics Canada, based on the GREG theory, has developed the Generalized Estimation System (GES). The availability of such a system allows for a routine use of the GREG estimator. However, the GREG estimator is vulnerable in the presence of outliers since it is based on the weighted least squares estimator of the regression parameters assuming a linear regression model.

Many robust procedures proposed for sample surveys are very efficient in terms of the mean square error (MSE) in the presence of outliers. However, this efficiency is obtained with an introduction of some bias. Hence, outlier robust procedures are usually looked at as a result of bias-variance trade off (Lee, 1995). This is quite acceptable when the sample size is small and the variance is the dominating factor in

¹ Hyunshik Lee, Westat, 1650 Research Blvd., MD 20850, U.S.A.

² Zdenek Patak, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6

the MSE. In this case a small bias is gratefully accepted to reduce a large variance. However, in some cases, the bias can be substantial even though it is not dominating to such a degree that statistical inferences based on the procedure are grossly distorted. It is so, particularly when the sample size increases because the bias stays the same but the variance decreases so that the bias term becomes dominant. This is characteristic of inconsistent robust procedures. This has been of great concern to the user of such procedures. Recognizing the bias problem of a robust procedure, Chambers (1986) proposed a bias-corrected (partially) robust ratio estimator, in which the bias is estimated by a robust technique and the estimated bias is added back to the robust estimator. Welsh and Ronchetti (1994) provided a justification for an additive bias correction from a Bayesian point of view. However, the estimator is still inconsistent.

As a solution to this inconsistency problem, Hulliger (1996) proposed a robust consistent procedure called the Minimum Estimated Risk (MER) estimator. He uses an M-estimator to robustify the Horvitz-Thompson estimator in such a way that the different tuning factors of the M-estimator are tried and selects one that gives the minimum estimated MSE. Since it includes the Horvitz-Thompson estimator as a possible choice when it reaches minimum estimated MSE, the estimator is consistent. The procedure is adaptive since the final form of the estimator is adapted to the sample information.

In this paper we also try to derive a consistent robust procedure that is applied to the GREG estimator. Our approach is also adaptive but we use a different framework. That is, we follow the Chambers' approach in which the bias is partially corrected. The main difference between Chambers' and ours is the way that the magnitude of the additive bias correction term is determined

2. ROBUST GREG ESTIMATOR

The GREG estimator assumes a linear model given by

$$\xi: y_i = \beta x_i + \varepsilon_i \quad (1.1)$$

where β is p -dimensional column vector of regression coefficients, x_i is the p -dimensional auxiliary vector, and $E_\xi(\varepsilon_i \varepsilon_j) = \sigma_i^2$ for $i = j$, $= 0$, otherwise. Then the GREG estimator of the population total Y has a general form given as

$$\hat{Y}_G = \sum_U \hat{\beta}'_L x_i + \sum_s \frac{1}{\pi_s} (y_s - \hat{\beta}'_R x_s) \quad (1.2)$$

where $\hat{\beta}'_L$ is the weighted least square estimator for β , and π_s is the inclusion probability of unit s in the sample s of a fixed size n selected from a finite population U of size N (see Särndal et al., 1992., p.225).

Lee (1995) discussed the robustification of the GREG estimator that has the following form:

$$\hat{Y}_{RG1} = \sum_U \hat{\beta}'_R x_i + \theta \sum_s \frac{1}{\pi_s} (y_s - \hat{\beta}'_R x_s) \quad (1.3)$$

where $\hat{\beta}'_R$ is a robust estimate of β , and θ is a number between 0 and 1 to be predetermined by the user. The θ should be chosen in such a way that the estimator becomes robust and consistent.

The second term in (1.3) is an additive bias-correction term. Another formulation based on the same idea can also be obtained as follows:

$$\hat{Y}_{RG2} = \sum_U \hat{\beta}'_R x_i + \theta \left(\hat{Y}_G - \sum_U \hat{\beta}'_R x_i \right) \quad (1.4)$$

where \hat{Y}_G is the GREG estimator given in (1.2). Alternatively, the estimator in (1.4) can be expressed as

$$\hat{Y}_{RG2} = \theta \hat{Y}_G + (1 - \theta) \sum_U \hat{\beta}'_R x_i \quad (1.5)$$

which is a shrinkage estimator. In the following discussion, the usual regularity conditions are assumed for the estimator to be consistent. Also, an appropriate asymptotic set-up is assumed in this discussion without specifying it.

Now, an important question is how to choose θ . In choosing θ , two factors should be considered: efficiency and consistency. The consistency requirement demands that $\theta \rightarrow 1$ as $n \rightarrow \infty$. On the other hand, the efficiency requires that θ should be close to zero when the variance of \hat{Y}_G is large because of outliers. A good choice of θ would be then

$$\theta = \frac{B^2}{V(\hat{Y}_G) + B^2} \quad (1.6)$$

where $B = E(\sum_U \hat{\beta}'_R x_i) - Y$, which is the bias of $\sum_U \hat{\beta}'_R x_i$ as an estimator of Y . Since the variance of \hat{Y}_G approaches zero with B^2 remaining positive as

$n \rightarrow \infty$, the θ defined by (1.6) will ensure consistency. On the other hand, when the variance of \hat{Y}_G is large compared to the squared bias, the θ should be small so that the estimator is made to be more robust rather than more bias-corrected. That is the property the θ in (1.6) has.

However, $V(\hat{Y}_G)$ and B are not known. We consider two ways of getting around this problem. The first is using historical information, which may enable us to determine a reasonable value of θ . This is the recommended approach as long as it is feasible. The other way is estimating $V(\hat{Y}_G)$ and B from the current sample. This approach can be regarded as adaptive in nature since the sample information is used to determine the final form of the estimator. We define two different θ 's in this way as follows:

$$\hat{\theta}_1 = \frac{\hat{B}_1^2}{\hat{V}(\hat{Y}_G) + \hat{B}_1^2} \quad (1.7)$$

where $\hat{B}_1 = \sum_s \frac{1}{\pi_i} (y_i - \hat{\beta}'_R x_i)$, and $\hat{V}(\hat{Y}_G)$ is some consistent estimator of $V(\hat{Y}_G)$, and

$$\hat{\theta}_2 = \frac{\hat{B}_2^2}{\hat{V}(\hat{Y}_G) + \hat{B}_2^2} \quad (1.8)$$

where $\hat{B}_2 = \sum_U \hat{\beta}'_R x_i - \hat{Y}_G$.

Obviously, \hat{B}_1 is more variable than \hat{B}_2 while \hat{B}_2 can be a biased estimator of bias when the GREG estimator is biased for Y . Normally, however, the bias is small. Note also that both bias estimators are consistent for estimating the bias. (B itself is sample size dependent and thus we assume that B approaches a certain constant as $n \rightarrow \infty$, which can be proved. As an example, see Hulliger, 1995)

3. SIMULATION RESULTS

In this section, we evaluate the performance of the proposed estimator using several artificial populations. It is assumed that the ratio model is appropriate and the ratio estimator (a GREG estimator under the model) is intended to use. We compare the performance of the robust GREG (ratio) estimator with the traditional ratio estimator. Five artificial populations were generated as follows:

- 1) First 128 points of x -values were generated from a mixed distribution, $0.95 \exp(1) + 0.05 \exp(3)$, where $\exp(\mu)$ is an exponential distribution with mean μ .
- 2) For each of the x -values, 5 different y -values were generated using the following 5 population models:

Population model 1: $N(3x_i, x_i^2)$;

Population model 2: $N(2x_i, x_i/4)$;

Population model 3: $N(2x_i, x_i/4)$ for 120 points and $\exp(2.5)$ for 8 points;

Population model 4:

$$\Gamma(\alpha, \beta), \alpha = \frac{0.4 + 0.25x_i}{0.0625x_i^{3/2}}, \beta = \frac{(0.4 + 0.25x_i)^2}{0.0625x_i^{3/2}};$$

Population model 5: $\Gamma(\alpha, \beta)$ as above for 120 points and $\exp(2)$ for 8 points.

These populations were used in Hulliger (1995). Populations 1 and 2 do not have outliers. Population 2 is the most ideal population for the ratio estimator. Population 3 represents a case with bad leverage points, which is very difficult to handle. In Population 4, the conditional mean of y given x is $0.4 + 0.25x$ and the variance is proportional to $x^{3/2}$. Thus, it represents a population with misspecified mean (intercept term) and variance. The last population represents a badly contaminated population. Plots of all these 5 populations are shown in Figures 7.1-7.5.

The sampling method is simple random sampling without replacement (SRSWOR). The sample sizes are 8, 16 and 32 for the artificial populations, and 5 and 30 for Populations 6 and 7, respectively.

We studied four estimators as given below:

$$\begin{aligned} \hat{Y}_G &= \hat{\beta}_L X \\ \hat{Y}_{RG0} &= \hat{\beta}_R X \\ \hat{Y}_{RG1} &= \hat{\beta}_R X + \hat{\theta}_1 \hat{B}_1 \\ \hat{Y}_{RG2} &= \hat{\beta}_R X + \hat{\theta}_2 \hat{B}_2 \end{aligned} \quad (1.9)$$

where $\hat{\beta}_L = \sum_s y_i / \sum_s x_i$ and thus \hat{Y}_G is the usual ratio estimator, $\hat{\beta}_R$ is the robust M-estimator of Huber for β ,

Fig.1 Rel. MSE Eff'cy (n = 8)

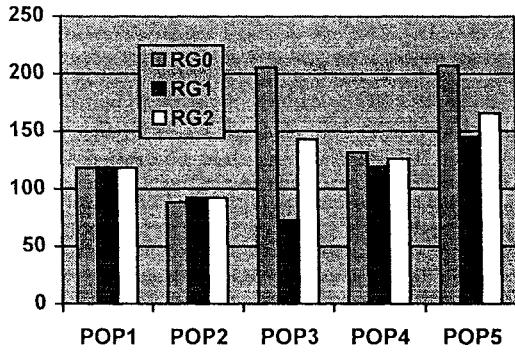


Fig.2 Rel. MSE Eff'cy (n = 16)

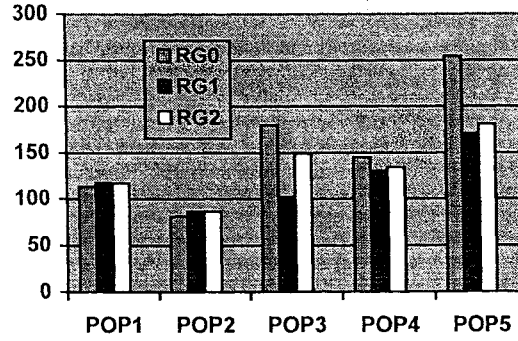


Fig.3 Rel. MSE Eff'cy (n = 32)

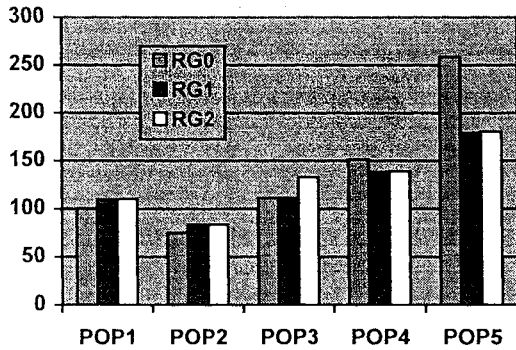
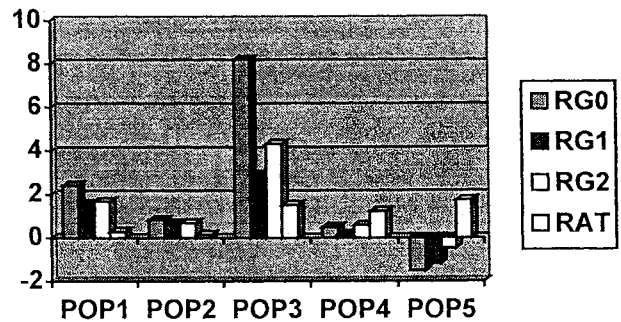


Fig.4 Relative Bias (n = 16)



and $\hat{\theta}_i$ and \hat{B}_i are defined as in (1.7) and (1.8). \hat{Y}_{RG0} is the robust prediction estimator without a bias correction. The results of the sampling experiment with 5,000 replicate simple random samples from each population are summarized in Figures 1-4.

Figures 1-3 show relative MSE efficiencies of RG0, RG1 and RG2 with respect to the ratio estimator. As expected under POP2, which is the ideal population for the ratio estimator, the robust estimators are less efficient. Under POP1 the robust estimators are slightly more efficient and they are similar among themselves. Under POP4 where there are quite a few outliers but residuals are fairly symmetric, the robust estimators are substantially more efficient and RG0 is slightly better than the other two. POP5 is somewhat similar to POP4 but the outliers are more extreme and thus the robust estimators perform much better. In this case, RG0 is the clear winner. This is another confirming evidence that a simple robust alternative based on M-estimation works very well when the model error is symmetric (see Lee, 1991, for example).

Note that for this situation the simple expansion estimator is almost as efficient as or even more than the ratio estimator (Rel, MSE efficiencies of the expansion estimator are 92, 105, and 102, for $n = 8, 16,$ and $32,$ respectively).

The results for POP3 are very interesting. This is the population that has asymmetric model error and the outliers are extreme. Therefore, RG0 is very MSE efficient but has a large bias. On the other hand, RG1 and RG2 are still much more efficient than the ratio estimator but have much smaller bias than RG0 (see Figure 4, which shows relative percent biases of the 4 estimators including the ratio estimator with respect to the population total). It is very interesting to observe that the advantage of the RG0 in terms of the MSE efficiency over the other robust estimators quickly diminishes as the sample size increases because the bias overtakes the dominance in the MSE as the sample size increases. Among RG1 and RG2, RG2 is superior. It would be better to use a fixed θ

Fig.5 Rel. MSE Effcy of RG2 for Different Fixed Theta (POP3)

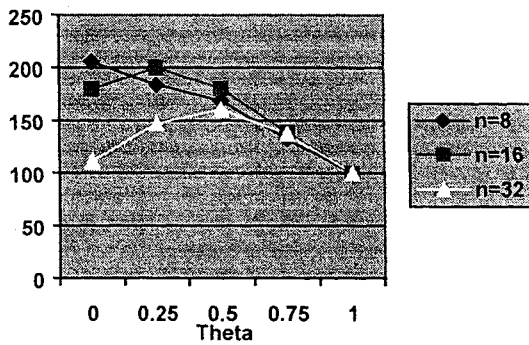
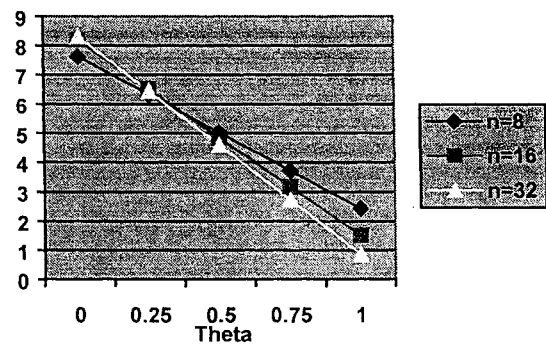


Fig.6 Rel. Bias of RG2 for Different Fixed Theta (POP3)



rather than adaptively determined θ if we can choose an appropriate value. Figure 5 shows the relative MSE efficiencies of RG2 when different fixed θ 's are used. The optimal θ in terms of the MSE changes as the sample size changes: 0, 0.25 and 0.5 for sample size 8, 16 and 32, respectively. This clearly indicates that, as the sample size increases, more bias should be corrected. Figure 6 gives the relative biases (w.r.t. the population total) and shows how the bias changes with θ . It is clear that the bias reduces linearly as θ increases. Note that when $\theta=1$, the RG2 estimator is the same as the ratio estimator and it still has a visible bias when the sample size is small and which decreases as the sample size increases.

The RG2 with adaptively determined (random) θ performed less efficiently than with the optimal fixed θ paying a dear price for using the sample to determine θ (Table 1). However, the bias is smaller than that of the estimator with the optimal fixed θ so that the confidence interval will be less distorted. One may attempt to use θ that gives minimum estimated MSE (a formula is not given here). However, the bias could be substantial enough to distort the confidence interval. Therefore, it may be desirable to have a smaller efficiency with a smaller bias.

Samp Size	Rel. Eff.	Rel. Bias	Ave $\hat{\theta}$	Std. of $\hat{\theta}$
8	142.7	4.56	0.144	0.209
16	148.9	4.31	0.235	0.236
32	132.8	3.77	0.360	0.257

4. CONCLUDING REMARKS

In this paper, we studied robust and design-consistent alternatives to the GREG estimator in the case where there is a danger of committing too big a bias by a simpler robust estimator such as RG0. If one has a good idea of what θ should be used, then such θ should be used. Otherwise, θ can be determined adaptively using the sample with some sacrifice in terms of lower MSE efficiency. However, such estimators can have smaller bias. It is again noted that the simpler robust estimator (RG0) is very efficient in most cases and should be seriously considered when there is some danger of outliers and model error is reasonably symmetric. Furthermore, if the sample size is small, such an estimator would be even more preferable since the variance component will be dominating in the MSE. However, if the sample size is moderate and there is a serious danger of a big bias, then the alternative estimators proposed in this paper may be preferable.

More research is needed, however. A variance (or MSE) estimator should be derived and the coverage property of a confidence interval procedure has to be studied. It would be worthwhile to study the estimators under more complex sampling situations and a more complex linear model as well.

REFERENCES

Chambers, R.L. (1986). Outlier Robust Finite Population Estimation. *Journal of American Statistical Association*, 81, 1063-1069.

Hulliger, B. (1995). Outlier Robust Horvitz-Thompson Estimator. *Survey Methodology*, 21, 79-88.

Colledge, and P.S. Kott, New York: Wiley and Sons, pp.503-526.

Lee, H. (1991). Model-Based Estimators That Are Robust to Outliers. In *Proceedings of the 1991 Annual Research Conference*, pp. 178-202. Washington, D.C.: USBC.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model-Assisted Survey Sampling*. New York: Springer-Verlag

Lee, H. (1995). Chapter 26: Outliers in Business Surveys, in *Business Survey Methods*, eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J.

Welsh, A.H., and Ronchetti, E.V. (1994). Bias-Calibrated Estimation of Totals and Quantiles from Sample Surveys Containing Outliers. Technical Report.

Fig. 7.1 Population 1

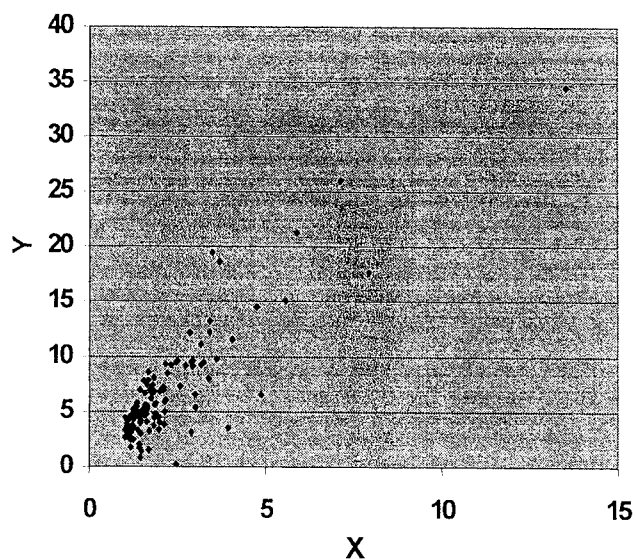


Fig. 7.2 Population 2

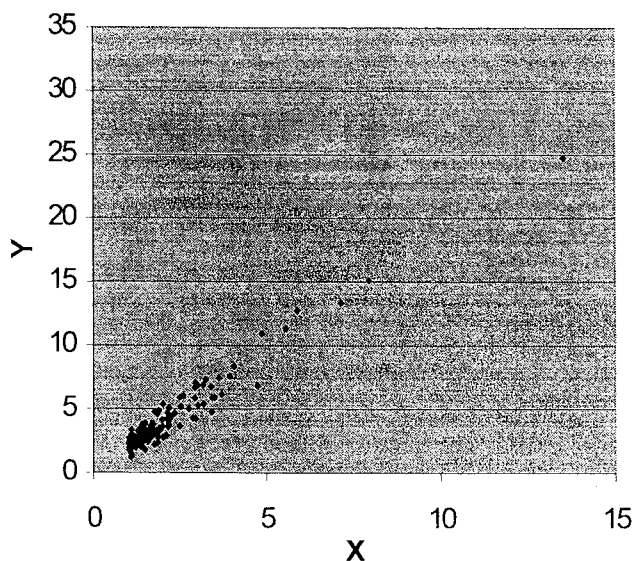


Fig. 7.3 Population 3

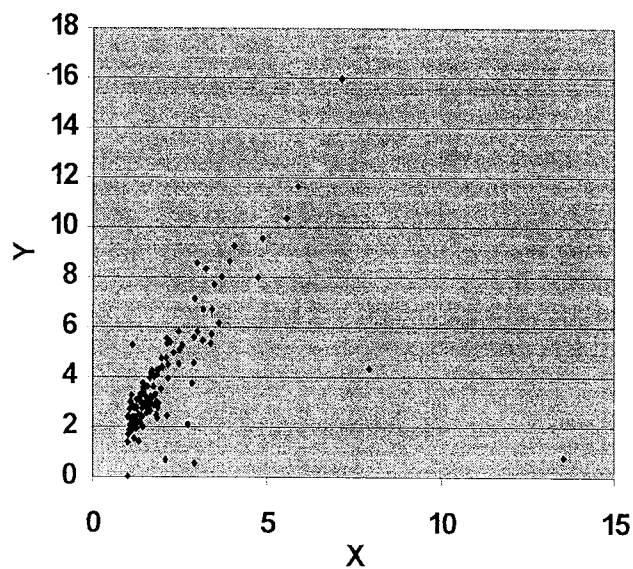


Fig. 7.4 Population 4

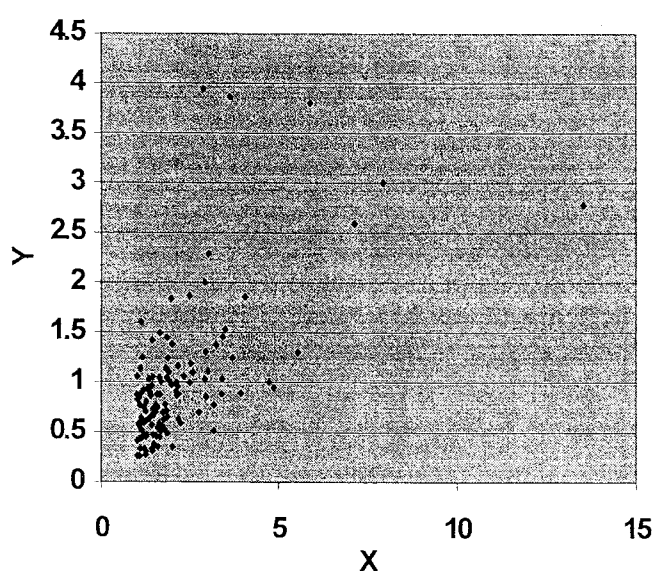


Fig.7.5 Population 5

