Available online at www.isas.org.in
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(2) 2009 117-122

Testing the Scale Parameter of an Exponential Distribution with known Coefficient of Variation in a Type II Right Censored Situation — Conditional Approach

C.D. Ravindran*
Central Institute for Research in Cotton Technology, Mumbai

(Received: December 2006, Revised: January 2008, Accepted: December 2008)

SUMMARY

The two-parameter exponential distribution $E(\mu, \theta)$, $\theta > 0$, is a well known probability model used for life-length studies owing to the useful description of observed variation it gives for many real life situations. In the context of life-length studies, the location parameter μ and the scale parameter θ respectively represent the minimum guaranteed life and the average excess life of an equipment or system. The parameters μ and θ are functionally unrelated and the statistical inference about these parameters make use of the existence of complete minimal sufficient statistics. This brings about a substantial simplification in the inferential problems. There, however, exist situations where the average life θ depends on the guaranteed life μ and the functionally independent nature of the parameters no longer hold, resulting in the loss of optimal properties of the statistics. In this situation, the two-parameter model reduces to a one-parameter model $E(a\theta, \theta)$, where 'a' is known. Ironically though the reduced model looks simplified with a single parameter θ, however, from the inference point of view, the problem of inference about θ becomes complicated. Several authors including the present author, have studied this kind of inference problem. In the present paper, the problem of testing a simple hypothesis about θ in the reduced model $E(a\theta, \theta)$ has been studied in the type II right censored situation from a conditional view point.

Keywords : Conditional UMP test, Minimal sufficient statistic, Power of unconditional test.

1. INTRODUCTION

The exponential distribution $E(a\theta, \theta)$ with p.d.f. given by

$$f(x\ ;\ \theta) = \frac{1}{\theta}\exp\left[\frac{-(x - a\theta)}{\theta}\right],\ x \geq a\theta \qquad (1.1)$$

where $\theta > 0$ is the unknown parameter and $a > 0$ is a known constant, has generated interest in the recent past owing to the scope it has offered to the existing procedures of statistical inference. For this distribution the coefficient of variation is known and is given by $100/(1 + a)$. This arises as a result of a functional relationship between the location (μ) and scale (θ) parameters of the original two-parameter exponential distribution $E(\mu, \theta)$. The interest for inferentialists comes from the fact that in the reduced model $E(a\theta, \theta)$, the inferential procedures instead of getting simplified becomes more intricate. The property of completeness enjoyed by the statistics in the case of $E(\mu, \theta)$ distribution no longer holds for that of $E(a\theta, \theta)$, as the standard theory of UMVUE in not applicable in this case.

In this background, two distinct approaches have been distinguished with respect to the choice of the reference set against which performances have to be

* *Corresponding author* : C.D. Ravindran
E-mail address : ravi_2612@rediffmail.com

evaluated. These are the unconditional and conditional approaches having their own merits/limitations. Naturally therefore, various workers have approached differently the inference problem of $E(a\theta, \theta)$. The attractive feature of the conditional viewpoint is that conditional models tends to be simpler than the original unconditional ones and frequently brings about simplification of theory (Lehmann 1986).

While Ebrahimi (1985), Ghosh and Razmpour (1984), Joshi and Nabar (1991) and Joshi and Sathe (1983) used the unconditional approach, Handa *et al.* (2002) and Samanta (1985) followed the conditional approach to study the inference problems. In general the unconditional approaches adopted by these workers produced adhoc, approximate or sub-optimal procedures whereas the conditional approaches produced conditionally optimal procedures, probably because the conditional procedures make use of ancillary information. In the next section, we show how this could be done.

For testing the simple hypothesis

$$H_0 : \theta = \theta_0 \text{ against } H_1 : \theta = \theta_1 \ (> \theta_0) \qquad (1.2)$$

Ebrahimi (1985) developed a test which was approximate because the distribution of the test statistic turned out to be complicated. Handa *et al.* (2002) using the conditional approach, developed a *conditional UMP test* for testing (1.2). They also investigated some interesting properties possessed by the distribution (1.1) like (i) the conditional superiority of the conditional test over its unconditional counterpart (Ebrahimi 1985), (ii) the property of conditional completeness, (iii) choice of ancillary statistics and (iv) large sample approximations. Their test, however, was based on a *complete sample* of size *n*.

There do exist numerous practical situations where complete samples are either unavailable or undesirable. For example, in life testing, fatigue testing and other kinds of tests of destructive nature, where data become available in an ordered manner, one can choose to discontinue experimentation after one has observed the first *r* observations. There are obvious advantages for choosing such a course, such as, one might be able to reach a decision in a shorter time or with fewer observations, than observing all items under test.

Let $X_{(1)} < X_{(2)} < \ldots < X_{(r)}$ denote the ordered statistics of a random sample of size *n* from the distribution (1.1) where $r \le n$. When only the first *r* observations are made or become available, the sample is usually termed a *right censored sample* (Epstein and Sobel (1954) and Tiku *et al.* (1986)). The objective of this paper is to extend the conditional UMP test developed by Handa *et al.* (2002) for complete samples to tests based on a right censored sample.

2. THE CONDITIONAL UMP TEST

It is well known that $(X_{(1)}, L)$ is a minimal sufficient statistic for θ, where

$$L = \sum_{i=2}^{r} (X_{(i)} - X_{(1)}) + (n-r)(X_{(r)} - X_{(1)})$$

Define

$$C = \frac{X_{(1)}}{X_{(1)} + r^{-1} L} \qquad (1.3)$$

Then *C* is an ancillary statistic and $(X_{(1)}, C)$ is also minimal sufficient. We now explore the existence of a conditional monotone likelihood ratio (MLR) for the family of conditional densities of $X_{(1)}$, given the ancillary *C*. For this purpose, we derive the conditional distribution of $X_{(1)}$, given *C*, in the following lemma.

Lemma 1. The conditional p.d.f. of $X_{(1)}$, given *C* is

$$f_{X_{(1)}|C}\left(x|c,\theta\right) = \begin{cases} \dfrac{\left[\dfrac{r+(n-r)c}{c\theta}\right]^r \exp\left[-\left\{\dfrac{r+(n-r)c}{c\theta}\right\}x\right] x^{r-1}}{J\left[\dfrac{a}{c}\{r+(n-r)c\}, r\right]}, & x \ge a\theta \\ 0, & \text{otherwise} \end{cases} \qquad (1.4)$$

where $J(\alpha, \beta) = \int_{\alpha}^{\infty} e^{-t} t^{\beta-1} dt = \tau(\beta) \Pr\left\{\chi^2(2\beta) > 2\alpha\right\}$

and $\tau(\beta)$ is the gamma function of β.

Proof. The joint p.d.f. of $X_{(1)}$ and *L* is given by

$$f_{X_{(1)},L}(x, y|\theta) = \frac{n}{\theta^r \tau(r-1)} \exp\left[-\frac{n(x-a\theta)}{\theta} + \frac{y}{\theta}\right] y^{r-2}, \quad x \ge a\theta, \ y > 0$$

Applying the transformation

$$U = X_{(1)},\ \ C = X_{(1)} \Big/ (X_{(1)} + r^{-1}L)$$

the joint p.d.f. of U and C is obtained as follows:

$$f_{U,C}(u,\ c/\theta)$$
$$= \frac{nr^{r-1}e^{na}(1-c)^{r-2}}{\theta^r c^r \tau(r-1)} \exp\left[-\left\{\frac{r+(n-r)c}{c\theta}\right\}u\right] u^{r-1},$$
$$u \geq a\theta,\ \ 0 < c < 1$$

This yields the marginal p.d.f. of C as

$$f_C(c) = \frac{nr^{r-1}e^{na}(1-c)^{r-2}}{\tau(r-1)\{r+(n-r)c\}^r} J\left[\left(\frac{r+(n-r)c}{c}\right)a,\ r\right],$$
$$0 < c < 1 \qquad (1.6)$$

The lemma now follows from (1.5) and (1.6).

Now in the next lemma, we prove that the family of conditional densities given by (1.4) has an MLR in $X_{(1)}$.

Lemma 2. The family of conditional densities $\{f_{X_{(1)}|C}(x/c,\theta),\ \theta > 0\}$ given by (1.4) has an MLR in $X_{(1)}$, given C.

Proof. For any $\hat{\theta} > \theta$, we have, on using Lemma 1:

$$\frac{f_{X_{(1)}|C}(x/c,\hat{\theta})}{f_{X_{(1)}|C}(x/c,\theta)}$$
$$= \frac{\left(\frac{r+(n-r)c}{c\hat{\theta}}\right)^r e^{-\left(\frac{r+(n-r)c}{c\hat{\theta}}\right)u} I(u > a\hat{\theta})}{\left(\frac{r+(n-r)c}{c\theta}\right)^r e^{-\left(\frac{r+(n-r)c}{c\theta}\right)u} I(u > a\theta)}$$
$$= \left(\frac{\theta}{\hat{\theta}}\right)^r e^{\frac{r+(n-r)c}{c}\left(\frac{1}{\theta} - \frac{1}{\hat{\theta}}\right)u} b(u)$$

where I(•) is the indicator function of the set (•) and

$$b(u) = I\,(u > a\hat{\theta}\,) \,/\, I\,(u > a\theta\,)$$
$$= \begin{cases} 1, & \text{if } u > a\hat{\theta} \\ 0, & \text{if } a\theta < u < a\hat{\theta} \end{cases}$$

Define $b(u) = 0$ if $u < a\theta$. Then it follows that $f_{X_{(1)}|C}(u|c,\hat{\theta}) \Big/ f_{X_{(1)}|C}(u|c,\theta)$ is a non-decreasing function of $X_{(1)}$ and the family of conditional densities given by (1.4) has an MLR in $X_{(1)}$ conditionally on C. The existence of a UMP conditional test for testing (1.2) follows by Theorem 2, pp. 78, Lehmann (1986). We state this result in the following theorem.

Theorem 1. The conditional size α test, given C, for testing

$$H_0\colon \theta \leq \theta_0 \ \text{ against } \ H_1\colon \theta > \theta_0$$

given by

reject H_0 if $X_{(1)} > K(c)$ is a UMP test.

3. THE POWER OF THE CONDITIONAL UMP TEST

The implementation of this UMP conditional test requires the computation of $K(c)$ as well as the test's power function. The next lemma presents these.

Lemma 3. For the testing problem stated in Theorem 1,

(i) for a UMP test of size α, $0 < \alpha < 1$, $K(c)$ is the solution of the equation

$$J\left[\left(\frac{r+(n-r)c}{c\theta_0}\right)K(c),\ r\right] = \alpha\, J\left[\left(\frac{r+(n-r)c}{c}\right)a,\ r\right] \qquad (1.7)$$

(ii) the power function of the UMP conditional test at θ is given by

$$\gamma(\theta|c) = \begin{cases} \dfrac{J\left[\left(\frac{r+(n-r)c}{c\theta}\right)K(c),\ r\right]}{J\left[\left(\frac{r+(n-r)c}{c}\right)a,\ r\right]}, & \text{if } \theta < \dfrac{K(c)}{a} \\ 1, & \text{if } \theta \geq \dfrac{K(c)}{a} \end{cases} \qquad (1.8)$$

where $K(c)$ is the solution of equation (1.7).

Proof.

(i) We have $P_r\,(X_{(1)} > K_{(c)} \mid c,\ \theta_0) = \alpha$

Using Lemma 1, we have

$$\int_{K(c)}^{\infty} \left(\frac{r+(n-r)c}{c\theta_0}\right)^r e^{-\left(\frac{r+(n-r)c}{c\theta_0}\right)u} u^{r-1}du$$

$$= \alpha\, J\left[\left(\frac{r+(n-r)c}{c}\right)a,\ r\right]$$

Change of variable by $t = \left(\frac{r+(n-r)c}{c\theta_0}\right)u$

yields the required result.

(ii) The power at θ is given by

$$\gamma(\theta|c) = \begin{cases} \Pr\left(X_{(1)} > K(c)|\ c,\theta\right), & \text{if } \theta < \dfrac{K(c)}{a} \\ \Pr\left(X_{(1)} > a\theta|\ c,\theta\right), & \text{if } \theta \geq \dfrac{K(c)}{a} \end{cases}$$

$$= \begin{cases} \dfrac{\int_{K(c)}^{\infty} \left(\frac{r+(n-r)c}{c\theta}\right)^r e^{-\left(\frac{r+(n-r)c}{c\theta}\right)u} u^{r-1}du}{J\left[\left(\frac{r+(n-r)c}{c}\right)\text{a, r}\right]}, & \text{if } \theta < \dfrac{K(c)}{a} \\ 1, & \text{if } \theta \geq \dfrac{K(c)}{a} \end{cases}$$

Transformation $t = \left(\frac{r+(n-r)c}{c\theta}\right)u$ yields (1.8).

4. COMPARISON OF CONDITIONAL AND UNCONDITIONAL TESTS

Comparison is done numerically between our proposed conditional test with the existing unconditional test of Ebrahimi (1985) for the censored case in terms of the power criterion. (The comparison of the tests in the case of complete samples has already been dealt with by Handa *et al.* (2002). Before carrying out the comparison, however, the critical point $K(c)$ of the conditional test is to be obtained from equation (1.7). Since the equation involves incomplete gamma functions, the solution for $K(c)$ has been obtained numerically by using the Newton-Raphson's method for which standard computer routines are available. Next, we have computed range of values of c for which the power of the conditional test dominates the power of the unconditional test. The power values were extensively computed for various values of the parameters for the UMP conditional test from (1.8) and for the unconditional test, from Theorem 1 of Ebrahimi (1985). The numerical comparison revealed that there exists an interval $(0, c')$ such that the power $\gamma(\theta|c)$ of the conditional test uniformly exceeds the power of $\gamma(\theta)$ of the unconditional test when $c \in (0, c')$ and the upper end point of $(0, c')$ moves closer to unity as a increases, implying that higher the value of a (which is the same as a smaller coefficient of variation), the more effective is an ancillary statistic in providing conditional inference about the parameter θ. It was also noted that lesser the value of c, more was the power of the conditional test.

Table 1 gives the value of c' for some chosen values of n, r, a, θ_0 and α. It also gives the values of

Table 1. Values of c' for interval $(0, c')$ of domination of conditional test over unconditional test

$\theta_0 = 1.0,\ \alpha = 0.05$

n	r	a	c'	c''
3	2	0.4	0.8	-
		0.7	*	-
	3	0.4	*	0.3
		0.7	*	0.5
5	2	0.4	*	-
		0.7	*	-
	3	0.4	*	-
		0.7	*	-
	4	0.4	*	-
		0.7	*	-
	5	0.4	*	0.3
		0.7	*	0.4
7	2	0.4	*	-
		0.7	*	-
	3	0.4	*	-
		0.7	*	-
	4	0.4	*	-
		0.7	*	-
	5	0.4	*	-
		0.7	*	-
	6	0.4	*	-
		0.7	*	-
	7	0.4	*	0.3
		0.7	*	0.4

* The whole interval (0,1)

Table 2. Comparison of power values of conditional and unconditional tests

$\theta_0 = 1.0,\ \theta_1 = 1.1(0.9)1.9,\ \alpha = 0.05$

n	r	a	c	Test	Power values at various values of θ_1								
					1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9
3	2	0.4	0.8	Conditional test	0.0753	0.1053	0.1392	0.1762	0.2156	0.2565	0.2983	0.3405	0.3827
			-	Unconditional test	0.0750	0.1041	0.1363	0.1707	0.2064	0.2427	0.2788	0.3144	0.3492
		0.7	0.9	Conditional test	0.0799	0.1170	0.1610	0.2110	0.2657	0.3244	0.3859	0.4494	0.5141
			-	Unconditional test	0.0781	0.1116	0.1494	0.1901	0.2325	0.2755	0.3182	0.3600	0.4005
	3	0.4	0.9	Conditional test	0.0770	0.1092	0.1454	0.1847	0.2258	0.2679	0.3102	0.3521	0.3931
			-	Unconditional test	0.0763	0.1073	0.1418	0.1789	0.2173	0.2563	0.2952	0.3333	0.3704
		0.7	0.9	Conditional test	0.0804	0.1181	0.1621	0.2111	0.2637	0.3188	0.3753	0.4322	0.4888
			-	Unconditional test	0.0788	0.1133	0.1524	0.1945	0.2383	0.2827	0.3268	0.3699	0.4114
5	2	0.4	0.9	Conditional test	0.7890	0.1145	0.1563	0.2033	0.2545	0.3090	0.3659	0.4244	0.4837
			-	Unconditional test	0.0775	0.1102	0.1469	0.1863	0.2273	0.2688	0.3102	0.3507	0.3900
		0.7	0.9	Conditional test	0.0891	0.1426	0.2114	0.2951	0.3928	0.5030	0.6242	0.7547	0.8927
			-	Unconditional test	0.0824	0.1224	0.1682	0.2179	0.2697	0.3219	0.3734	0.4231	0.4704
	3	0.4	0.9	Conditional test	0.0796	0.1159	0.1580	0.2045	0.2543	0.3061	0.3589	0.4119	0.4644
			-	Unconditional test	0.0830	0.1121	0.1503	0.1913	0.2340	0.2772	0.3202	0.3622	0.4027
		0.7	0.9	Conditional test	0.0876	0.1379	0.2007	0.2747	0.3583	0.4497	0.5469	0.6481	0.7518
			-	Unconditional test	0.0827	0.1232	0.1696	0.2200	0.2725	0.3254	0.3774	0.4276	0.4754
	4	0.4	0.9	Conditional test	0.0813	0.1198	0.1643	0.2131	0.2645	0.3173	0.3701	0.4222	0.4727
			-	Unconditional test	0.0795	0.1152	0.1556	0.1993	0.2446	0.2906	0.3361	0.3804	0.4231
		0.7	0.9	Conditional test	0.0872	0.1361	0.1959	0.2645	0.3401	0.4203	0.5033	0.5874	0.6711
			-	Unconditional test	0.0833	0.1247	0.1722	0.2239	0.2776	0.3317	0.3849	0.4361	0.4846
	5	0.4	0.9	Conditional test	0.0835	0.1254	0.1738	0.2267	0.2821	0.3382	0.3936	0.4473	0.4985
			-	Unconditional test	0.0811	0.1191	0.1624	0.2093	0.2581	0.3074	0.3560	0.4032	0.4484
		0.7	0.9	Conditional test	0.0876	0.1367	0.1958	0.2624	0.3342	0.4087	0.4839	0.5582	0.6304
			-	Unconditional test	0.0841	0.1267	0.1758	0.2292	0.2876	0.3405	0.3952	0.4477	0.4973
7	4	0.4	0.9	Conditional test	0.0837	0.1264	0.1769	0.2334	0.2940	0.3571	0.4212	0.4851	0.5477
			-	Unconditional test	0.0814	0.1198	0.1636	0.2110	0.2604	0.3103	0.3595	0.4072	0.4527
		0.7	0.9	Conditional test	0.0948	0.1587	0.2419	0.3433	0.4603	0.5901	0.7294	0.8753	1.0000
			-	Unconditional test	0.0840	0.1341	0.1890	0.2488	0.3108	0.3727	0.4328	0.4899	0.5431
	5	0.4	0.9	Conditional test	0.0850	0.1295	0.1817	0.2395	0.3005	0.3630	0.4252	0.4859	0.5442
			-	Unconditional test	0.0825	0.1225	0.1684	0.2182	0.2700	0.3222	0.3736	0.4232	0.4704
		0.7	0.9	Conditional test	0.0937	0.1545	0.2316	0.3227	0.4247	0.5344	0.6486	0.7645	0.8798
			-	Unconditional test	0.0875	0.1353	0.1911	0.2520	0.3150	0.3779	0.4389	0.4966	0.5503
	6	0.4	0.9	Conditional test	0.0870	0.1343	0.1898	0.2508	0.3146	0.3789	0.4420	0.5024	0.5595
			-	Unconditional test	0.0838	0.1258	0.1742	0.2268	0.2814	0.3364	0.3903	0.4421	0.4912
		0.7	0.9	Conditional test	0.0934	0.1527	0.2265	0.3115	0.4044	0.5016	0.6000	0.6972	0.7911
			-	Unconditional test	0.0881	0.1369	0.1939	0.2560	0.3204	0.3845	0.4465	0.5050	0.5594
	7	0.4	0.9	Conditional test	0.0893	0.1402	0.2000	0.2655	0.3335	0.4013	0.4668	0.5286	0.5859
			-	Unconditional test	0.0852	0.1295	0.1807	0.2364	0.2943	0.3523	0.4090	0.4631	0.5141
		0.7	0.9	Conditional test	0.0938	0.1531	0.2257	0.3079	0.3959	0.4858	0.5747	0.6603	0.7411
			-	Unconditional test	0.0888	0.1387	0.1972	0.2609	0.3268	0.3923	0.4555	0.5150	0.5700

c'' corresponding to the cases when $r = n$, *i.e.* when samples are complete. Table 2 presents the power comparison of the conditional and the unconditional tests for various levels of censoring r. For the conditional test, only the power values corresponding to $c = 0.9$ is shown, since for values of $c < 0.9$, the power automatically exceeds the power at $c = 0.9$.

5. CONCLUSION

It is clear from Table 2 that the power of the conditional test completely dominates the power of the unconditional test for *all values of the parameters* when the samples are censored. It is pointed out here that in the complete sample case also, where the comparison was with the most powerful (MP) unconditional test of Joshi and Nabar (1991), an interval $(0, c')$ was found such that the power of the conditional test dominated the power of the unconditional MP test. Thus, it has been possible to establish the supremacy of our conditional UMP test over the unconditional test of Ebrahimi (1985) for all values of c, thus demonstrating the effectiveness of the conditional test.

ACKNOWLEDGEMENT

This work came out as a result of the unpublished Ph.D. work submitted to the Indian Institute of Technology, New Delhi by the author. The author wishes to thank the IIT, New Delhi for giving all facilities and encouragement to do this work. He also wishes to thank Director, CIRCOT, Mumbai for providing keen interest, encouragement and facilities in the preparation of this paper.

REFERENCES

Ebrahimi, N. (1985). Estimating from censored samples the location of an exponential distribution with known coefficient of variation. *Cal. Stat. Assoc. Bull.,* **34,** 169-177.

Epstein, B. and Sobel, M. (1954). Some theorems relevant to life testing from an exponential distribution. *Ann. Math. Statist.,* **25,** 458-466.

Ghosh, M. and Razmpour, A. (1984). Estimation of the common location parameter of several exponentials. *Sankhya,* **A46**, 383-394.

Handa, B.R., Kambo, N.S. and Ravindran, C.D. (2002). Testing of scale parameter of the exponential distribution with known coefficient of variation: Conditional approach. *Comm. Statist. – Theory Methods,* **31(1)**, 73-86.

Joshi, S.M. and Nabar, S.P. (1991). Testing the scale parameter of the exponential distribution with known coefficient of variation. *Comm. Statist.—Theory Methods,* **20(2)**, 1129-1132.

Joshi, S. and Sathe, Y.S. (1983). On estimating the scale parameter of the exponential distribution with a known linear relation between the location and the scale parameter. *Naval Res. Logistics Quarterly,* **30**, 601-607.

Lehmann, E.L. (1986). *Testing Statistical Hypotheses*. John Wiley and Sons.

Samanta, M. (1985). On estimating the location parameter of an exponential distribution with known coefficient of variation. *Cal. Stat. Assoc. Bull.,* **34**, 43-49.

Tiku, M.L., Tan, W.Y. and Balakrishnan, N. (1986). *Robust Inference*. Marcel Dekker, Inc., New York.

Available online at www.isas.org.in
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(2) 2009 123-131

GARCH Nonlinear Time Series Analysis for Modelling and Forecasting of India's Volatile Spices Export Data

Ranjit Kumar Paul, Prajneshu* and Himadri Ghosh
Indian Agricultural Statistics Research Institute, New Delhi

(Received: June 2008, Revised: December 2008, Accepted: April 2009)

SUMMARY

Modelling and forecasting of India's spices export data set, which exhibits a volatile behaviour, is first attempted through the Box-Jenkins Autoregressive integrated moving average (ARIMA) approach. Subsequently, Generalized autoregressive conditional heteroscedastic (GARCH) nonlinear time-series model along with its estimation procedures are thoroughly studied. Lagrange multiplier test for testing presence of Autoregressive conditional heteroscedastic (ARCH) effects is also discussed. The GARCH model is employed for modelling and forecasting of the data. Comparative study of the fitted ARIMA and GARCH models is carried out from the viewpoint of dynamic one-step ahead forecast error variance along with Mean square prediction error (MSPE), Mean absolute prediction error (MAPE) and Relative mean absolute prediction error (RMAPE). The SAS and EViews, Ver. 4 software packages along with computer programs in C are used for data analysis. Superiority of GARCH model over ARIMA approach is demonstrated for the data under consideration. Possible use of more accurate forecasts obtained by GARCH methodology vis-à-vis ARIMA approach is briefly discussed.

Keywords: ARIMA, EViews software package, Generalized autoregressive conditional heteroscedastic model, Monthly export data of spices, SAS software package, Volatility.

1. INTRODUCTION

Spices are the most important commercial crops of our country. The important spices extensively grown in India are cardamom, pepper, chillies, turmeric, and ginger. With respect to production, consumption and export of spices, India ranks first in the World. The total area in India under these spices is over one million hectares, and these accounted for an annual export of about Rs. 3330 crores during the year 2006-07. In short, India commands a formidable position in the World spices trade with 47% share in volume and 40% in value. More than 150 value-added products of spices are currently available for export. The most important among these are spice oils and oleoresins. More than 70% of their total World supply is from India. The target set by Government of India is to increase the spices export by ten-folds in the next ten years. To achieve such an ambitious target, the twin goals of spice sector should be to enhance the annual growth rate from 13% to 20% and share of export of value-added spice products from 58% to 75%. As emphasized by Jaffee (2005), volatility seems to be the norm rather than the exception in international markets for spices due to the structure of the trade, climatic conditions, and the rapidity with which producers can respond to price changes. Proper monitoring and appropriate policy measures require efficient modelling and forecasting of spices time-series data.

The most widely used technique for analysis of time-series data is, undoubtedly, the Box Jenkins

* *Corresponding author* : Prajneshu
E-mail address : prajneshu@yahoo.co.in

Autoregressive integrated moving average (ARIMA) methodology. However, it is based on some crucial assumptions, like linearity, stationarity, and homoscedastic errors. Further, time-series data quite often exhibit features which can not be explained by ARIMA model, which is "linear". As an example, the famous time-series of average monthly sunspot numbers exhibits a cyclical behaviour in such a way that the series generally increases at a faster rate than it decreases. Similarly, asymmetric phenomenon arises with economic series, which tend to behave differently when the economy is moving into recession rather than when coming out of it. Many financial time-series show periods of stability, followed by unstable periods with high volatility. The loss in continuing to use the age-old ARIMA methodology is that this type of behaviour can not be explained satisfactorily, and so "nonlinear time-series models" are usually needed to describe data sets in which variance changes through time. The search for an appropriate model of this type would lead to a greater insight into the underlying mechanism. An excellent description of these and other related issues is given in Chatfield (2001).

During last two decades or so, the area of Nonlinear time-series modelling has been rapidly developing. The most promising parametric nonlinear time-series model has been the Autoregressive conditional heteroscedastic (ARCH) model, which was introduced by Engle (1982), and for which he was awarded the prestigious Nobel Prize in Economics in 2003. This entails a completely different class of models which is concerned with modelling volatility. The objective is not to give better point forecasts but rather to give better estimates of the variance which, in turn, allows more reliable forecast intervals leading to a better assessment of risk (Chatfield 2001). The ARCH model allows the conditional variance to change over time as a function of squared past errors leaving the unconditional variance constant. The presence of ARCH-type effects in financial and macro-economic time series is a well established fact. The combination of ARCH specification for conditional variance and the Autoregressive (AR) specification for conditional mean has many appealing features, including a better specification of the forecast error variance. Ghosh and Prajneshu (2003) employed AR(p)-ARCH(q)-in-Mean model for carrying out modelling and forecasting of volatile monthly onion price data. The AR-ARCH model has also been used as the basic "building blocks" for Markov switching and mixture models (See e.g. Lanne and Saikkonen 2003, and Wong and Li 2001). Various aspects of the family of mixtures of ARCH models have been thoroughly investigated by Ghosh *et al.* (2005, 2006).

However, ARCH model has some drawbacks. Firstly, when the order of ARCH model is very large, estimation of a large number of parameters is required. Secondly, conditional variance of ARCH(q) model has the property that unconditional autocorrelation function (Acf) of squared residuals, if it exists, decays very rapidly compared to what is typically observed, unless maximum lag q is large. To overcome these difficulties, Bollerslev (1986) proposed the Generalized ARCH (GARCH) model in which conditional variance is also a linear function of its own lags. This model is also a weighted average of past squared residuals, but it has declining weights that never go completely to zero. It gives parsimonious models that are easy to estimate and, even in its simplest form, has proven surprisingly successful in predicting conditional variances. Angelidis *et al*. (2004) used GARCH model for describing Value-at-Risk.

In this paper, our purpose is to thoroughly study the GARCH model and its estimation procedures. Subsequently, this model along with the Box Jenkins ARIMA model is applied to describe the volatility of monthly export of spices from India during the period April 2000 to August 2006. Finally, the performance of one-step ahead forecasting for three months, i.e. from September 2006 to November 2006 by both the models is examined.

2. DESCRIPTION OF MODELS

2.1 The ARIMA Model

The Autoregressive moving average (ARMA) model, denoted as ARMA(p, q), is given by

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + ... + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q} \quad (2.1)$$

or equivalently by

$$\varphi(B) y_t = \theta(B) \varepsilon_t \quad (2.2)$$

where

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - ... - \varphi_p B^p$$

and

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q$$

In the above, B is the backshift operator defined by $By_t = y_{t-1}$. A generalization of ARMA models, which incorporates a wide class of nonstationary time-series, is obtained by introducing "differencing" in the model. The simplest example of a nonstationary process which reduces to a stationary one after differencing is "Random Walk". A process $\{ y_t \}$ is said to follow Autoregressive integrated moving average (ARIMA), denoted by ARIMA(p, d, q), if $\nabla^d y_t = (1 - B)^d \varepsilon_t$ is ARMA(p, q). The model is written as

$$\varphi(B)(1 - B)^d y_t = \theta(B)\varepsilon_t \qquad (2.3)$$

where ε_t are identically and independently distributed as $N(0, \sigma^2)$. The integration parameter d is a nonnegative integer. When $d = 0$, the ARIMA(p, d, q) model reduces to ARMA(p, q) model.

2.2 The GARCH Model

The ARCH(q) model for the series $\{\varepsilon_t\}$ is given by

$$\varepsilon_t | \psi_{t-1} \sim N(0, h_t) \qquad (2.4)$$

Here ψ_{t-1} denotes information available up to time $t - 1$, and

$$h_t = a_0 + \sum_{i=1}^{q} a_i \varepsilon_{t-i}^2 \qquad (2.5)$$

where $a_0 > 0$, $a_i \geq 0$ for all i and $\sum_{i=1}^{q} a_i < 1$ are required to be satisfied to ensure nonnegativity and finite unconditional variance of stationary $\{\varepsilon_t\}$ series.

Bollerslev (1986) proposed the Generalized ARCH (GARCH) model in which conditional variance is also a linear function of its own lags and has the following form

$$h_t = a_0 + \sum_{i=1}^{q} a_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} b_j h_{t-j} \qquad (2.6)$$

A sufficient condition for the conditional variance to be positive is

$$a_0 > 0,\ a_i \geq 0,\ i = 1, 2, ..., q;\ b_j \geq 0, \quad j = 1, 2, ..., p$$

The GARCH (p, q) process is weakly stationary if and only if $\sum_{i=1}^{q} a_i + \sum_{j=1}^{p} b_j < 1$. The most popular GARCH model in applications is the GARCH(1, 1) model. To express GARCH model in terms of ARMA model, denote $\eta_t = \varepsilon_t^2 - h_t$. Then from eq. (2.6)

$$\varepsilon_t^2 = a_0 + \sum_{i=1}^{Max(p,q)} (a_i + b_i)\varepsilon_{t-i}^2 + \eta_t + \sum_{j=1}^{p} b_j \eta_{t-j} \qquad (2.7)$$

Thus a GARCH model can be regarded as an extension of the ARMA approach to squared series $\{ \varepsilon_t^2 \}$.

2.3 Estimation of Parameters

Estimation of parameters for ARIMA model is generally done through Nonlinear least squares method. Fortunately, several software packages are available for fitting of ARIMA models. In this paper, SAS, Ver. 9.1 software package is used. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) values for ARIMA model are computed by

$$AIC = T'\log(\sigma^2) + 2(p + q + 1) \qquad (2.8)$$

and

$$BIC = T'\log(\sigma^2) + (p + q + 1)\log T' \qquad (2.9)$$

where T' denotes the number of observations used for estimation of parameters and σ^2 denotes the Mean square error.

In order to estimate the parameters of GARCH model, Method of maximum likelihood is used. The loglikelihood function of a sample of T observations, apart from constant, is

$$L_T(\theta) = T^{-1} \sum_{t=1}^{T} \left(\log h_t + \varepsilon_t^2 h_t^{-1} \right)$$

where

$$h_t = a_0 + \sum_{i=1}^{q} a_i y_{t-i}^2 + \sum_{j=1}^{p} b_j\ h_{t-j}$$

If $f(.)$ denotes the probability density function of ε_t, generally, maximum likelihood estimators are derived by minimizing

$$L_T(\theta) = T^{-1}\sum_{t=v}^{T}\left(\log\sqrt{\tilde{h}_t} - \log\ f(\varepsilon_t/\sqrt{\tilde{h}_t})\right)$$

where $\tilde{h}_t$ is the truncated version of h_t (Fan and Yao 2003). For heavy tailed error distribution, Peng and Yao (2003) proposed Least absolute deviations estimation (LADE), which minimizes $\sum_{t=v}^{T}\left|\log\varepsilon_t^2 - \log(h_t)\right|$, where $v = p + 1$, if $q = 0$ and $v > p + 1$, if $q > 0$. Fan and Yao (2003) and Straumann (2005) have given a good description of various estimation procedures for conditionally heteroscedastic time-series models.

The Akaike information criterion (AIC) and Bayesian information criterion (BIC) values for GARCH model with Gaussian distributed errors are computed by

$$\text{AIC} = \sum_{t=1}^{T}\left(\log\tilde{h}_t + \varepsilon_t^2\tilde{h}_t^{-1}\right) + 2(p + q + 1) \qquad (2.10)$$

and

$$\text{BIC} = \sum_{t=1}^{T}\left(\log\tilde{h}_t + \varepsilon_t^2\tilde{h}_t^{-1}\right) + 2(p + q + 1)\log(T - v + 1) \qquad (2.11)$$

where T is the total number of observations.

Evidently, the likelihood equations are extremely complicated. Fortunately, the estimates can be obtained by using a software package, like EViews, SAS, SPLUS GARCH, GAUSS, TSP, MATLAB, and RATS. In the present investigation, the Gaussian maximum likelihood estimation procedure available in EViews software package, Ver. 4 is used for data analysis. Further, AIC and BIC values for ARIMA and GARCH models are computed separately by writing computer programs in C.

2.4 Testing for ARCH Effects

Let $\varepsilon_t = y_t - \phi y_{t-1}$ be the residual series. The Lagrange multiplier (LM) test for squared series $\{\varepsilon_t^2\}$ may be used to check for conditional heteroscedasticity. The test is equivalent to usual F-statistic for testing $H_0 : a_i = 0$, $i = 1, 2, ..., q$ in the linear regression

$$\varepsilon_t^2 = a_0 + a_1\varepsilon_{t-1}^2 + ... + a_q\varepsilon_{t-q}^2 + e_t,\ t = q + 1,..., T \qquad (2.12)$$

where e_t denotes the error term, q is the prespecified positive integer, and T is the sample size. Let $SSR_0 = \sum_{t=q+1}^{T}\left(\varepsilon_t^2 - \bar{\omega}\right)^2$, where $\bar{\omega} = \sum_{t=q+1}^{T}\varepsilon_t^2/T$ is sample mean of $\{\varepsilon_t^2\}$, and $SSR_1 = \sum_{t=q+1}^{T}\hat{e}_t^2$, where $\hat{e}_t$ is the least square residual of eq. (2.12). Then, under H_0, the ARCH-LM test statistic, viz.

$$F = \frac{(SSR_0 - SSR_1)/q}{SSR_1/(T - q - 1)} \qquad (2.13)$$

follows asymptotically the chi-squared distribution with q degrees of freedom.

3. MODELLING OF INDIA'S SPICES EXPORT DATA

All-India data of monthly export of spices during the period April 2000 to November 2006 are obtained from Indiastat (www.indiastat.com) available at I.A.S.R.I., New Delhi and the same are exhibited in Fig. 1. From the total 80 data points, first 77 data points corresponding to the period April 2000 to August 2006 are used for building the model and remaining are used for validation purpose. A perusal of the data shows that, during the period from April 2004 to February 2006, these varied between Rs 143 crores and Rs 189 crores. Then the spices export suddenly jumped almost 80% to the level of Rs 345 crores in March 2006, which was followed

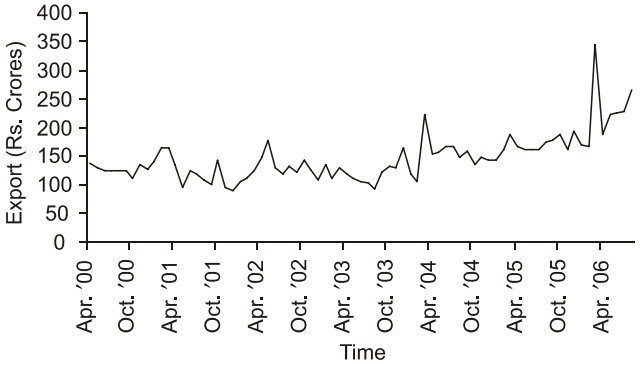

Fig. 1. Data of monthly spices export from India

by a sudden dip to as low as Rs 188 crores in the very next month. All this clearly shows that volatility was present during March 2006. Similar type of presence of volatility was noticed at several other time-epochs, like May 2001, August 2001, and June 2002.

3.1 Fitting of ARIMA Model

A perusal of estimated autocorrelation functions (acf) of original series, reported in Table 1, shows that it decays very slowly, implying thereby that this series may be differenced. Analytically, this issue may be resolved by applying the unit root test, proposed by Dickey and Fuller (1979) for parameter ρ in the auxiliary regression

$$\Delta_1 y_t = \rho\, y_{t-1} + \alpha_1 \Delta_1 y_{t-1} + \varepsilon_t \qquad (3.1)$$

which is derived from the AR(2) model, viz.

$$(1 - \varphi_1 L - \varphi_2 L^2) y_t = \varepsilon_t \qquad (3.2)$$

by expressing the associated autoregressive polynomial in L as

$$1 - \varphi_1 L - \varphi_2 L^2 = (1 - \varphi_1 - \varphi_2)L + (1 - L)(1 - \alpha_1 L) \qquad (3.3)$$

where $\alpha_1 = -\varphi_2$. Using (3.3) in (3.2), we get

$$(1 - L)(1 - \alpha_1 L) y_t = \rho L y_t + \varepsilon_t \qquad (3.4)$$

where $\rho = \varphi_1 + \varphi_2 - 1$. Now, presence of unit root, i.e. $L = 1$ in the autoregressive polynomial implies that the condition for nonstationarity is $1 - \varphi_1 - \varphi_2 = 0$, i.e. $\varphi_1 + \varphi_2 = 1$. Further, region of stationarity is $\varphi_1 + \varphi_2 < 1$. Thus, the unit root test reduces to testing H_0: $\rho = 0$ against H_1: $\rho < 0$. In the present situation, $\hat{\rho}$ is computed as 0.005. Since calculated value of t-statistic, i.e. 0.212 is found to be greater than the tabulated value of t-statistic at 5% level of significance, i.e. –1.95 (Franses, 1998, Page 82), therefore H_0 is not rejected at 5% level and so $\rho = 0$. Thus, there is presence of one unit root and so differencing is required until the acf shows an interpretable pattern with only a few significant autocorrelations. On taking the first difference of the original series, it is seen that only a few autocorrelations, reported in Table 1, are high making it easier to select the order of the model. On differencing the original series twice, it is seen that the sum of the autocorrelations of double differenced series is –0.507, which implies that the series is overdifferenced (Franses, 1998, Page 50). This suggests that only one differencing would be more appropriate.

Table 1. Sample autocorrelation functions (acf) and partial autocorrelation functions (pacf) of the original and differenced series

Lag	acf of the series	pacf of the series	acf of the differenced series	pacf of the differenced series
1	0.550	0.550	–0.498	–0.498
2	0.525	0.319	–0.048	–0.393
3	0.539	0.267	0.188	–0.065
4	0.405	–0.022	–0.197	–0.181
5	0.457	0.147	0.079	–0.110
6	0.303	–0.148	0.042	–0.040
7	0.274	–0.023	–0.009	0.069
8	0.251	–0.039	–0.008	0.049
9	0.211	0.053	0.000	0.040
10	0.202	0.001	–0.058	–0.064
11	0.214	0.121	–0.007	–0.120
12	0.239	0.102	0.105	0.014
13	0.182	–0.031	–0.078	–0.009
14	0.196	–0.006	0.044	0.033
15	0.175	–0.035	–0.014	0.000
16	0.165	–0.013	–0.018	0.034
17	0.162	–0.020	–0.001	–0.007
18	0.138	0.031	0.040	0.047
19	0.109	–0.038	0.005	0.059
20	0.089	–0.009	–0.035	0.016
21	0.095	0.022	0.018	–0.005
22	0.083	0.019	–0.009	0.007
23	0.089	0.016	–0.113	–0.175
24	0.163	0.161	0.306	0.222

The appropriate ARIMA model is chosen on the basis of minimum Akaike information criterion (AIC) and Bayesian information criterion (BIC) values. Using eqs. (2.8) and (2.9), the AIC and BIC values, which are respectively computed as 521.29 and 532.95, the ARIMA(1, 1, 1) model is selected for modelling and forecasting of India's spices export data. The estimates of parameters of above model are reported in Table 2.

Further, the residual error variance for the fitted ARIMA model is computed as 867.762. The graph

Table 2. Estimates of parameters along with their standard errors for fitted ARIMA(1, 1, 1) model

Parameter	Estimate	Standard error
AR1	–0.100	0.159
MA1	0.696	0.119
Constant	1.468	0.966

of fitted model along with data points is exhibited in Fig. 2. Evidently, the fitted ARIMA(1, 1, 1) model is not able to capture successfully the volatility present at various time-epochs, like October 2001; May 2002; March 2004; and March 2006.

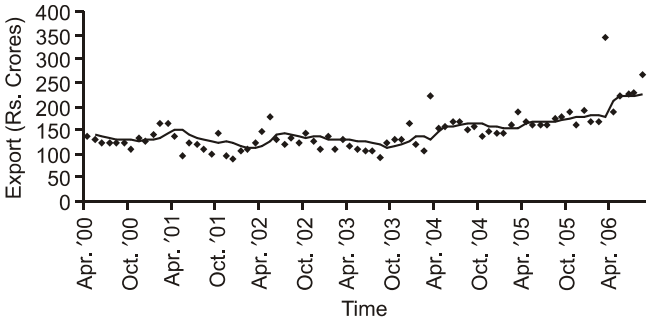

Fig. 2. Fitted ARIMA(1, 1, 1) model along with data points

3.2 Fitting of GARCH Model

On investigating autocorrelations of squared residuals of the fitted ARIMA(1,1,1) model, reported in Table 3, it is found that the autocorrelation is highest at lag 24, which is 0.265. The ARCH-LM test statistic at lag 24 computed using equations (2.12) and (2.13) is 37.48, which is significant at 5% level. But it is not reasonable to apply ARCH model of order 24 in view of the enormously large number of parameters. Therefore, the parsimonious GARCH model is applied. The AR(1)-GARCH(1, 1) model is selected on the basis of minimum AIC and BIC values. The estimates of parameters of the above model along with their corresponding standard errors in brackets ( ) using Method of maximum likelihood with Gaussian distributed error terms are

$$y_t = \underset{(33.692)}{157.99} + \underset{(0.087)}{0.829}\, y_{t-1} + \varepsilon_t$$

where $\varepsilon_t = h_t^{1/2}\eta_t$, and h_t satisfies the variance equation

$$h_t = \underset{(237.058)}{1427.855} + \underset{(0.277)}{0.354}\, \varepsilon_{t-1}^2 + \underset{(0.206)}{0.509}\, h_{t-1}$$

Using eqs. (2.10) and (2.11), the AIC and BIC values for fitted AR(1)- GARCH(1,1) model are respectively computed as 479.77 and 521.97.

Table 3. Sample autocorrelation functions (acf) and partial autocorrelation functions (pacf) of the squared residuals of the ARIMA (1, 1, 1) series

Lag	acf of the squared residuals series	pacf of the squared residuals series
1	–0.015	–0.015
2	–0.045	–0.045
3	–0.030	–0.031
4	–0.041	–0.044
5	0.009	0.005
6	–0.023	–0.027
7	–0.007	–0.010
8	–0.022	–0.027
9	–0.025	–0.027
10	–0.025	–0.031
11	–0.029	–0.035
12	0.028	0.020
13	–0.021	–0.028
14	–0.023	–0.028
15	–0.014	–0.020
16	–0.029	–0.035
17	–0.005	–0.016
18	–0.030	–0.039
19	–0.015	–0.026
20	–0.021	–0.033
21	–0.024	–0.035
22	–0.001	–0.015
23	–0.022	–0.034
24	0.265	0.254

To study the appropriateness of fitted GARCH model, autocorrelation functions of standardized residuals and squared standardized residuals are computed and the same are reported in Table 4. It is found that, in both situations, the autocorrelation functions are insignificant at 5% level, thereby confirming that the mean and variance equations are correctly specified. Conditional standard deviation for

Table 4. Autocorrelation functions of the standardized residuals and squared standardized residuals for fitted GARCH(1,1) model

Lag	acf of standardized residuals	Q-Statistic	Probability	acf of squared standardized residuals	Q-Statistic	Probability
1	–0.093	0.672	---	0.157	1.906	---
2	0.017	0.694	0.405	–0.009	1.913	0.167
3	0.222	4.589	0.101	0.018	1.937	0.380
4	–0.014	4.604	0.203	–0.113	2.957	0.398
5	–0.014	4.621	0.328	–0.041	3.093	0.542
6	0.083	5.192	0.393	0.152	4.995	0.416
7	0.138	6.784	0.341	0.092	5.707	0.457
8	0.009	6.791	0.451	–0.050	5.924	0.549
9	–0.030	6.867	0.551	–0.069	6.336	0.610
10	–0.025	6.922	0.645	–0.135	7.927	0.541
11	0.064	7.288	0.698	–0.109	8.995	0.533
12	0.019	7.321	0.773	0.048	9.202	0.603
13	–0.002	7.321	0.836	0.009	9.210	0.685
14	0.218	11.784	0.545	–0.054	9.482	0.736
15	0.052	12.039	0.603	–0.028	9.558	0.794
16	0.021	12.084	0.673	–0.127	11.116	0.744

fitted model is plotted in Fig. 3. Further, graph of fitted model along with data points is exhibited in Fig. 4. Obviously, the fitted GARCH model is able to capture the volatility present in the data set.

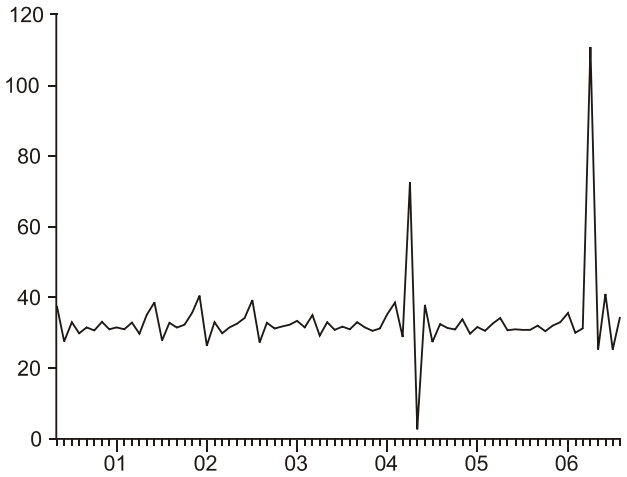

Fig. 3 Conditional standard deviation of fitted AR(1)-GARCH(1,1) model

4. FORECASTING OF INDIA'S SPICES EXPORT DATA

One-step ahead forecasts of export of spices along with their corresponding standard errors inside the brackets ( ) for the months of September 2006 to November 2006 in respect of above fitted models are reported in Table 5. In view of the assumption of homoscedasticity of error terms in ARIMA approach, the one-step ahead forecast error variance remains constant. A perusal indicates that, for fitted GARCH model, all the observed values lie within one standard error of their forecasts. However, this attractive feature

Table 5. One-step ahead forecasts of export of spices (in Rs. Crores) for fitted models

Months	Observed values	Forecasts by	
		ARIMA (1, 1, 1)	AR(1)-GARCH (1, 1)
Sep. '06	270.91	235.67 (29.61)	247.14 (40.93)
Oct. '06	232.59	240.27 (29.61)	231.89 (48.17)
Nov. '06	286.21	241.50 (29.61)	265.68 (33.31)

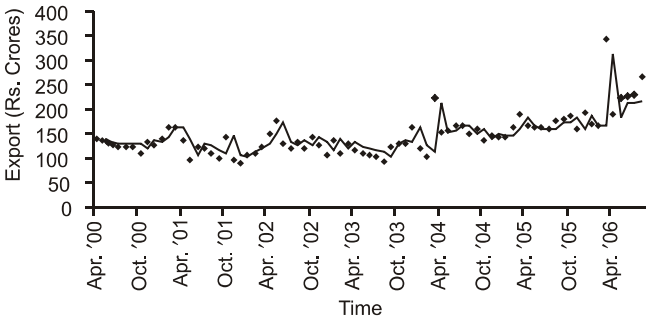

Fig. 4 Fitted AR(1) – GARCH (1,1) model along with data points

does not hold for fitted ARIMA model. Further, for GARCH model, it may be noted that the magnitude of one-step ahead forecast error at a time-epoch is also reflected in the magnitude of corresponding forecast error variance at next time-epoch. For example, when one-step ahead forecast error (i.e. 23.77, being the difference of observed value 270.91 and forecast value 247.14) and corresponding forecast error variance during September 2006 (i.e. 40.93) are large, one-step ahead forecast error variance for October 2006 (i.e. 48.17) is also large. But when one-step ahead forecast error during October 2006 (i.e. 0.70, being the difference of observed value 232.59 and forecast value 231.89) is small, corresponding forecast error variance for November 2006 (i.e. 33.31) is also relatively small. It may be noticed that while periods of strong turbulence caused large fluctuations in Indian spices export, these were often followed by relative calm and slight fluctuations. Further, while most volatility is embedded in the random error, its variance depends on previously realized random errors with large errors being followed by large errors and small by small. Thus, the fitted GARCH model is capable of explaining volatility in the underlying phenomenon. This is in contrast to the ARIMA model wherein the random error is assumed to be constant over time.

The Mean square prediction error (MSPE) values and Mean absolute prediction error (MAPE) values for fitted GARCH model are respectively computed as 18.14 and 15.00, which are found to be lower than the corresponding ones for fitted ARIMA model, viz. 33.17 and 29.02 respectively. Further, a comparative study of forecasts of monthly spices export by above discussed two models is carried out on the basis of their Relative mean absolute prediction error (RMAPE) values defined as

$$\text{RMAPE} = \frac{1}{6}\sum_{i=1}^{6}\left\{\left|y_{t+i} - \hat{y}_{t+i}\right| / y_{t+i}\right\} \times 100$$

The RMAPE values for fitted ARIMA (1,1,1) and AR(1)-GARCH(1,1) models are respectively computed as 32.46 and 10.82. The lower values of all the three statistics, viz. MSPE, MAPE, and RMAPE reflect superiority of GARCH approach for forecasting purposes also.

The more realistic forecast intervals for India's spices export data obtained through GARCH approach could be of immense help to planners in formulating appropriate strategies. This type of information would go a long way in arriving at the appropriate decisions on several issues, like Quantities of spices in future to be exported and quantities to be earmarked for domestic consumption, Whether to impose ban on exports at various points of time, and Whether or not to impose export duty and how much in case export of spices is allowed. This would enable the planners to take appropriate policy decisions from time to time well in advance in order to meet the targets set for Indian spices export. These, in turn, would also benefit the farmers in production of optimum quantities of spices. All this would ultimately result in efficient management of India's spices sector export scenario on a sound statistical basis.

5. CONCLUDING REMARKS

It has been shown that for Indian spices export time-series data, the usual assumption of homoscedasticity of error terms is not satisfied. For modelling as well as forecasting of this data, the GARCH nonlinear time-series model has performed better than the well-known Box-Jenkins ARIMA model. Therefore, for data sets exhibiting volatility, ARIMA approach should be abandoned in favour of GARCH methodology in order to obtain more accurate forecasts and changing forecast interval lengths. The methodology advocated in this paper can also be used for forecasting other volatile data sets.

ACKNOWLEDGEMENT

Authors are grateful to the referee for valuable comments.

REFERENCES

Angelidis, T., Benos, A. and Degiannakis, S. (2004). The use of GARCH models in VaR estimation. *Stat. Meth.*, **1**, 105-128.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *J. Econ.,* **31**, 307-327.

Chatfield, C. (2001). *Time Series Forecasting*. Chapman and Hall, U.S.A.

Dickey, D.A. and Fuller, W.A. (1979). Distribution of the estimators for the autoregressive time series with a unit root. *J. Amer. Statist. Assoc.*, **74**, 427-431.

Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica,* **50**, 987-1008.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer, U.S.A.

Franses, P.H. (1998). *Time Series Models for Business and Economic Forecasting*. Cambridge University Press, U.K.

Ghosh, H. and Prajneshu (2003). Nonlinear time series modelling of volatile onion price data using AR(p)-ARCH(q)-in-Mean. *Cal. Stat. Assoc. Bull.*, **54,** 231-247.

Ghosh, H., Sunilkumar, G. and Prajneshu (2005). Nonlinear time-series modelling: A mixture-ARCH approach. *J. Ind. Soc. Agril. Statist.*, **59,** 209-216.

Ghosh, H., Iquebal, M.A. and Prajneshu (2006). On mixture nonlinear time-series modelling and forecasting for ARCH effects. *Sankhya*, **68**, 111-129.

Huang, D., Wang, H. and Yao, Q. (2008). Estimating GARCH models: When to use what? *Econ. J.,* **11**, 27-38.

Jaffee, S. (2005). *Delivering and Taking the Heat: Indian Spices and Evolving Product and Process Standards.* Agriculture and Rural Development Discussion Paper 19. The World Bank, U.S.A.

Lanne, M. and Saikkonen, P. (2003). Modeling the U.S. short-term interest rate by mixture autoregressive processes. *J. Fin. Econ.*, **1**, 96-125.

Peng, L. and Yao, Q. (2003). Least absolute deviations estimation for ARCH and GARCH models. *Biometrika,* **90**, 967-975.

Straumann, D. (2005). *Estimation in Conditionally Heteroscedastic Time Series Models*. Springer, Germany.

Wong, C.S. and Li, W.K. (2001). On a mixture autoregressive conditional heteroscedastic model. *J. Amer. Statist. Assoc.*, **96**, 992-995.

Available online at www.isas.org.in
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(2) 2009 133-137

A Note on Alternative Estimators for Multi-Character Surveys

Raghunath Arnab[1*], Sarjinder Singh[2] and P.A.E. Serumaga-Zake[3]
[1]*University of Botswana, Botswana*
[2]*The University of Texas at Brownsville and Southmost.*
[3]*Department of Statistics, North-West University, Mafikeng Campus*

(Received: June 2008, Revised: March 2009, Accepted: April 2009)

SUMMARY

The problems of estimating the population total in multi-character surveys in varying probability sampling schemes when the measure of size is not well-related to the study variables, have been considered by Rao (1966), Scott and Smith (1969) and Arnab (2001). In the present note, their results are extended for a wider class of superpopulation models and sampling designs.

Keywords: Auxiliary information, Model design unbiased estimator, Multi-character surveys, Optimal estimator, PPSWR sampling, Superpopulation model.

1. INTRODUCTION

In large-scale surveys, we generally estimate population parameters like totals, means and variances for more than one character at a time. In such a survey if a sample is selected by a varying probability sampling scheme using an auxiliary variable x as a measure of size, then the resulting sampling design may yield efficient estimators for those characters which are well-related to the auxiliary variable but may not provide efficient estimators for the characters which are poorly related to the auxiliary variable. Rao (1966) first addressed the requirement for the adjustments of the conventional estimators in such a multicharacter survey and provided with some alternative estimators for estimation of a finite population total under various sampling schemes when the correlation between the study and auxiliary variable is very low. The alternative estimators, proposed by Rao (1966), fare better than the conventional estimators under the following superpopulation model:

$$\text{Model } M1 : E_{M1}(y_i) = \mu,\ V_{M1}(y_i) = \sigma^2 \text{ and } C_{M1}(y_i, y_j) = 0 \text{ for } i \neq j \quad (1)$$

where, μ, $\sigma^2 (> 0)$ are unknown model parameters and E_{M1}, V_{M1} and C_{M1} denote respectively the expectation, variance and covariance with respect to the model $M1$. Following Rao (1966), Scott and Smith (1969), Bansal and Singh (1985), Kumar and Agarwal (1997), Mangat and Singh (1992-93) and Singh and Horn (1998), among others also suggested some alternative estimators under the PPSWR sampling scheme. Arnab (2001) extended Rao's (1966) results for an arbitrary varying probability sampling scheme and showed that Rao's (1966) results could be derived from his results as special cases. For the sake of clarity, let us describe Rao (1966), Arnab (2001) and Scott and Smith (1969) results relevant to our present discussion as follows.

1.1 Estimators due to Rao (1966), Arnab (2001) and Scott and Smith (1969)

Let $U = \{1, ..., i, ..., N\}$ be a finite population of N units and y_i (x_i) be the value of the study (auxiliary)

* *Corresponding author* : Raghunath Arnab
E-mail address : arnabr@mopipi.ub.bw

variable for the i^{th} unit of the population and $Y(X)$ be their total. Here x_i's are assumed to be known and positive for every $i \in U$. Let a sample s of size n be selected from U by a varying probability sampling scheme using x_i as a measure of size for the i^{th} unit. Rao (1966), Arnab (2001) and Scott and Smith (1969) alternative estimators are given below.

1.1.1 Rao's (1966) estimators

The conventional estimators for a finite population total Y under PPSWR, πps and Rao-Hartley-Cochran (1963, RHC) sampling schemes are respectively given by

$$t_{pps} = \frac{1}{n}\sum_{i\in s} n_i(s)\frac{y_i}{p_i} \tag{2}$$

$$t_{hte}(\pi ps) = \sum_{i\in s}\frac{y_i}{\pi_i} = \sum_{i\in s}\frac{y_i}{np_i} \tag{3}$$

and

$$t_{rhc} = \sum_{i\in s} y_i \frac{P_i}{p_i} \tag{4}$$

where $p_i = x_i/X$, $n_i(s)$ = frequency of the i^{th} unit in s, π_i = inclusion probability for the i^{th} unit and P_i = sum of p_j's for the group containing the i ($\in s$)th unit for selection of sample under RHC sampling scheme.

Rao (1966) showed that the alternative estimators

$$t_{pps}(1) = \frac{1}{n}\sum_{i\in s} n_i(s)y_i = N\,\bar{y}_n t_0(s) = N\sum_{i\in s} y_i/n \quad = N\,\bar{y}_s$$

and $t_{rhc}(1) = N\sum_{i\in s} y_i P_i$ are unbiased for Y under model $M1$ and more efficient than the corresponding conventional estimators t_{pps}, t_{hte} (πps) and t_{rhc}.

The Murthy's (1957) estimator for PPSWOR sampling scheme is given by

$$t^{*}_{mur} = \frac{1}{p(s)}\sum_{i\in s} y_i\, p(s\,|\,i) \tag{5}$$

where $p(s)$ and $p(s|i)$ denote respectively the probability of selection of an unordered sample s based on PPSWOR sampling scheme and the conditional probability of selection s given that the unit i was chosen on the first draw. Rao (1966) proposed an alternative estimator of $t^{*}_{mur}(2)$ (which is t^{*}_{mur} with n = 2) as

$$t_{mur}(2) = \frac{N}{2-p_i-p_j}\{(1-p_j)y_i + (1-p_i)y_j\} \tag{6}$$

The estimator $t_{mur}(2)$ is inconsistent but unbiased under model M_1. Rao (1966) did not prove theoretically whether or not the proposed alterntive estimator $t_{mur}(2)$ is superior to the conventional estimator $t^{*}_{mur}(2)$. However, he showed empirically the superiority of $t_{mur}(2)$ over $t^{*}_{mur}(2)$.

1.1.2 Arnab's (2001) estimators

Let P_n be the class of fixed effective size n sampling design and C be the class of linear homogeneous unbiased estimators for Y consisting of estimators of the form

$$t(s) = \sum_{i\in s} b_{si}\, y_i \tag{7}$$

where b_{si}'s are constants free from y_i's satisfying the unbiasedness condition

$$\sum_{s\supset i} b_{si}\, p(s) = 1 \quad \forall\, i \in U \tag{8}$$

Arnab (2001) showed that the alternative estimators $t_0(s) = N\,\bar{y}_s$ fares better than any estimator belonging to C in the sense that

$$E_{M1}V_p\left(t_0(s)\right) \le E_{M1}V_p\left(t(s)\right) \;\forall\; p \in P_n, t(s) \in C \tag{9}$$

From equation (9), we can establish the following inequalities

$$E_{M1}V_p\left(t_0(s)\right) \le E_{M1}V_p\left(\sum_{i\in s}\frac{y_i}{\pi_i}\right),\; E_{M1}V_p\left(t^{*}_{mur}\right)$$

and also

$$E_{M1}V_p\left(t_0(s)\right) \le E_{M1}V_p\left(t_{rhc}(1)\right) \le E_{M1}V_p\left(t_{rhc}\right)$$

1.1.3 Scott and Smith's (1969) estimators

Scott and Smith considered a class C^{*} of linear homogeneous model design unbiased estimators of the population total Y based on a sampling design $p \in P_n$ of n distinct units. The class C^{*} consists of estimators of the form

$$t(s) = \sum_{i \in s} b_{si} y_i$$

satisfying the model-design unbiasedness condition

$$\sum_s p(s) \sum_{i \in s} b_{si} y_i = N \qquad (10)$$

Scott and Smith (1969) proved that

$$E_{M1}(MSE\ (t_0(s))) = E_{M1}\ E_p\ (t_0(s) - Y)^2$$

$$\leq E_{M1}(MSE(t(s))) = E_{M1} E_p (t(s) - Y)^2$$

Here E_p denotes expectation with respect to design p.

2. PROPOSED ESTIMATOR UNDER MODEL M2

In this present note we have showed that Rao (1966) and Arnab (2001)'s results can be extended further for a wider superpopulation model given below.

Model $M2$: $E_{M2}(y_i) = \mu$, $V_{M2}(y_i) = \sigma_i^2 = \sigma^2 v(x_i)$

and $C_{M2}(y_i, y_j) = 0$ for $i \neq j$

where $v(x_i)$ is a function of x_i only. Various forms of the variance function $v(x_i)$ specially $v(x_i) = x_i^g$ with $g \geq 0$ are referred to by Cassel *et al.* (1971), and Chaudhuri and Stenger (1992) among others. We have also extended the Scott and Smith's (1969) result by showing that their result is valid also for the wider classes of sampling designs $P_n^* (\supset P_n)$ consisting of n units which may not necessarily be distinct.

Theorem 1. $E_{M2} V_p \left(t_{pps}(1)\right) \leq E_{M2} V_p (t_{pps})$

Proof.

$$E_{M2} V_p(t_{pps}) = E_{M2} V_p \left(\frac{1}{n} \sum_{i \in s} n_i(s) \frac{y_i}{p_i} \right)$$

$$= E_{M2} \left(\frac{1}{n} \left(\sum_{i=1}^{N} \frac{y_i^2}{p_i} - Y^2 \right) \right)$$

$$= \frac{1}{n} \left(\sum_{i=1}^{N} \sigma_i^2 \left(\frac{1}{p_i} - 1 \right) \right) + \mu^2 V_p \left(\frac{1}{n} \sum_{i \in s} n_i(s) \frac{1}{p_i} \right) \qquad (11)$$

$$E_{M2} V_p[t_{pps}(1)] = E_{M2} \left(\frac{1}{n} \left(\sum_{i=1}^{N} \frac{z_i^2}{p_i} - Z^2 \right) \right)$$

where $z_i = y_i p_i$ and $z = \sum_{i=1}^{N} z_i$

$$E_{M2} V_p[t_{pps}(1)] = \frac{1}{n} \sum_{i=1}^{N} \sigma_i^2 p_i (1 - p_i) \qquad (12)$$

From (11) and (12), we get

$$E_{M2} V_p(t_{pps}) - E_{M2} V_p(t_{pps}(1))$$

$$= \sum_{i=1}^{N} \sigma_i^2 \frac{(1 + p_i)}{p_i} (1 - p_i)^2 + \mu^2 V_p \left(\frac{1}{n} \sum_{i \in s} n_i(s) \frac{1}{p_i} \right) \geq 0$$

Theorem 2. $E_{M2} V_p(t_0(s)) \leq E_{M2} V_p(t(s))\ \forall\ t(s) \in C$, $p \in P_n$, if σ_i^2 is a decreasing function of π_i.

Proof. $V_p(t(s)) = E_p(t(s))^2 - Y^2$

$$= \sum_{i=1}^{N} y_i^2 \left(\sum_{s \supset i} b_{si}^2 p(s) - 1 \right) + \sum_{i \neq}^{N} \sum_{j=1}^{N} y_i y_j \left(\sum_{s \supset i, j} b_{si} b_{sj} p(s) - 1 \right)$$

and

$$E_{M2} V_p(t(s)) = \sum_{i=1}^{N} \sigma_i^2 \left(\sum_{s \supset i} b_{si}^2 p(s) - 1 \right) + \mu^2 V_p \left(\sum_{i \in s} b_{si} \right)$$

$$\geq \sum_{i=1}^{N} \sigma_i^2 \left(\sum_{s \supset i} b_{si}^2 p(s) - 1 \right)$$

$$\geq \sum_{i=1}^{N} \sigma_i^2 \left(\frac{1}{\pi_i} - 1 \right) \qquad (13)$$

$\left(\sum_{s \supset i} b_{si}^2 p(s) \ \geq \frac{\left(\sum_{s \supset i} b_{si} p(s) \right)^2}{\sum_{s \supset i} p(s)} = \frac{1}{\pi_i} \right.$ follows from the unbiasedness condition (8))

$$V_p(t_0(s)) = \frac{N^2}{n^2} E_p \left(\left(\sum_{i \in s} y_i \right)^2 - \left(\sum_{i=1}^{N} y_i \pi_i \right)^2 \right)$$

$$= \frac{N^2}{n^2} \left(\sum_{i=1}^{N} \pi_i (1 - \pi_i) y_i^2 - \sum_{1 \neq}^{N} \sum_{j=1}^{N} (\pi_i \pi_j - \pi_{ij}) y_i y_j \right) \qquad (14)$$

Equation (14) yields

$$E_{M2}V_p(t_0(s)) = \frac{N^2}{n^2}\sum_{i=1}^{N}\pi_i(1-\pi_i)\sigma_i^2 + \frac{N^2}{n^2}\mu^2\left(\sum_{i=1}^{N}\pi_i(1-\pi_i) - \sum_{i\neq}^{N}\sum_{j=1}^{N}(\pi_i\pi_j - \pi_{ij})\right)$$

$$= \frac{N^2}{n^2}\sum_{i=1}^{N}\pi_i(1-\pi_i)\sigma_i^2 \qquad (15)$$

(noting $\sum_{i=1}^{N}\pi_i = n$ and $\sum_{i\neq}^{N}\sum_{j=1}^{N}\pi_{ij} = n(n-1)$)

Finally from (13) and (15), we get

$$E_{M2}V_P(t(s)) - E_{M2}V_p(t_0(s)) \geq \sum_{i=1}^{N}\sigma_i^2\left(\frac{1}{\pi_i}-1\right) - \frac{N^2}{n^2}\sum_{i=1}^{N}\pi_i(1-\pi_i)\sigma_i^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}q_i\left(\pi_i - \frac{n}{N}\right)$$

$$= Cov(q_i, \pi_i)$$

where

$$q_i = -\frac{\sigma_i^2}{n^2}\left(\frac{1}{\pi_i}-1\right)(n+N\pi_i)$$

$$= -\frac{\sigma_i^2}{n^2}\left(n\left(\frac{1}{\pi_i}-1\right) + N(1-\pi_i)\right) \qquad (16)$$

Now if σ_i^2 is a decreasing function of π_i, then q_i will be an increasing function of π_i since $n\left(\frac{1}{\pi_i}-1\right) + N(1-\pi_i)$ is a decreasing function of π_i. In this situation $Cov(q_i, \pi_i)$ becomes positive.

Corollary 1. For an IPPS sampling design where $\pi_i = np_i$ and for the model with $M2$, $\sigma_i^2 = \sigma^2 x_i^g$, σ_i^2/π_i becomes a decreasing function of π_i if $g \leq 1$. In this case $E_{M2}V_p(t_0(s)) \leq E_{M2}V_p(t(s))$.

In particular if $\sigma_i^2 = \sigma^2$, Theorem 2 reduces to inequality (9).

Theorem 3. For a sampling design $p \in P_n^*$ and $t(s) \in C^*$

$$E_{M1}(MSE(t(s)) = E_{M1}E_p(t(s) - Y)^2 \geq \sigma^2 N\left(\frac{N}{\gamma} - 1\right)$$

$$= E_{M1}E_{p_0}\left(t_0(s) - Y\right)^2$$

where $\gamma = E_p(\gamma_s)$ = expected effective sample size $= \sum_s \gamma_s p(s)$ and p_0 is a fixed effective size sampling design with Prob$\{\gamma_s = \gamma\} = 1$.

Proof.

$$E_{M1}(MSE(t(s))) = E_{M1}E_P(t(s) - Y)^2$$

$$= E_pE_{M1}(t(s) - Y)^2$$

$$= \sigma^2\sum_s p(s)\left(\sum_{i\in s} b_{si}^2 + N - 2\sum_{i\in s} b_{si}\right) + \mu^2\sum_s p(s)\left(\sum_{i\in s} b_{si} - N\right)^2 \qquad (17)$$

and

$$\sum_s p(s)\left(\sum_{i\in s} b_{si}^2 + N - 2\sum_{i\in s} b_{si}\right)$$

$$= \sum_s p(s)\sum_{i=1}^{N} I_{si}\left(b_{si}-1\right)^2 + N - \gamma \qquad (18)$$

Further the model-design unbiased condition (10) yields

$$\sum_s p(s)\sum_{i=1}^{N} I_{si}\left(b_{si}-1\right)^2 = \sum_s\sum_{i=1}^{N} I_{si}p(s)(b_{si}-1)^2$$

$$\geq \frac{\left(\sum_s\sum_{i=1}^{N} I_{si}p(s)(b_{si}-1)\right)^2}{\sum_s\sum_{i=1}^{N} I_{si}p(s)}$$

$$= \frac{(N-\gamma)^2}{\gamma} \qquad (19)$$

Finally using (17), (18) and (19), we get

$$E_{M1}(MSE(t(s))) \geq N\left(\frac{N}{\gamma} - 1\right)\sigma^2 \qquad (20)$$

Equality in equation (20) holds for a sampling strategy based on fixed effective size sampling design p_0 satisfying Prob$\{\gamma_s = \gamma\} = 1$ and an estimator $t(s)$ with $b_{si} = N/\gamma_s = N/\gamma$.

Remark 1: Scott and Smith's (1969) assertions of non-existence of the lower bound given in (20) for a with replacement sampling design is clearly incorrect.

ACKNOWLEDGEMENTS

The authors are grateful to the referee and editor for their valuable suggestions that led to an improvement of the paper.

REFERENCES

Arnab, R. (2001). Estimation of a finite population total in varying probability sampling for multi-character surveys. *Metrika,* **54**, 159-177.

Bansal, M.L. and Singh. R. (1985). An alternative estimator for multiple characteristics in PPS sampling. *J. Statist. Plann. Inf.,* **11**, 313-320.

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1977). *Foundation of Inference in Survey Sampling.* John Wiley and Sons, New York.

Chaudhuri, A. and Stenger, H. (1992). *Survey Sampling Theory and Methods.* Marcel Deckker, Inc.

Hansen, M.H., Hurwitz, W.N. (1943). On theory of sampling from finite populations. *Ann. Math. Statist.,* **14**, 433-362.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc*., **47**, 663-685.

Kumar, P. and Agarwal, S.K. (1997). Alternative estimators for the population totals in multiple characteristic surveys. *Comm. Statist.—Theory Methods,* **26**, 2527-2537.

Mangat, N.S. and Singh, R. (1992-93). Sampling with varying probabilities without replacement : A review. *Aligarh J. Stat.,* **12&13**, 75-105.

Murthy, M.N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhya*, **18**, 379-390.

Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1963). A simple procedure of unequal probability without replacement. *J. Roy. Statist. Soc.,* **B24**, 482-491.

Rao, J.N.K. (1966). Alternative estimator for PPS sampling for multiple characteristics. *Sankhya,* **A28**, 47-60.

Scott, A. and Smith, T.M.F. (1969). A note on estimating secondary characteristics in multivariate surveys. *Sankhya,* **A31**, 497-498.

Singh, S. and Horn, S. (1998). An alternative estimator for multi-character surveys. *Metrika*, **48**, 99-107.

Available online at www.isas.org.in
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(2) 2009 139-149

Assessing Stability of Crop Varieties with Incomplete Data

B.M.K. Raju[1*], V.K. Bhatia[2] and L.M. Bhar[2]
[1]*Central Research Institute for Dryland Agriculture, Hyderabad*
[2]*Indian Agricultural Statistics Research Institute, New Delhi*

(Received: March 2008, Revised: April 2009, Accepted: April 2009)

SUMMARY

Joint regression has been very popular among plant breeders to evaluate stability of crop varieties tested under multi location-year trials. The plant breeders often finish their investigation for stability with Eberhart and Russell model (1966) though the component of deviation from linear regression is found significant for many varieties. Consequently, in such situations one cannot do ranking of all the genotypes tested with regard to stability. Eventually pair-wise comparisons with respect to stability can be made only in a subset of genotypes, whose deviations from linear regression are found not significant. This paper aims to emphasize the limitations of Eberhart and Russell model in evaluating stability of a set of varieties tested and suggest plant breeders alternative measures of stability when Eberhart and Russell model fails to comment on the stability of a sufficiently large number of varieties tested. Another problem of plant breeders that this paper also attempts, is dealing with the situation of the stability analysis when some cells in two-way table of genotype × environments are blank. This paper examines methods cited in literature to handle incomplete data situations and brings out their practical relevance in the current generation of computers. An attempt has been made to develop handy computational algorithms wherever required and compares various procedures with respect to their capabilities in evaluating stability of the varieties.

Keywords: Eberhart and Russell model, Multi location-year trials, Stability, Stability variance.

1. INTRODUCTION

In developing countries like India, where the number of small and marginal farmers with small holdings is very high, stable yields minimize risk and ensure sustainable food supply. One of the plant breeders' aim has been to develop cultivars that produce stable yields across a range of environments. Environments may be locations or years or combinations of both.

The existence of interaction reflecting differences among varieties in their ability to maintain performance over a wide range of environmental conditions has long been recognized to exist (Finlay and Wilkinson 1963). This ability, which is an important property of a variety, is usually referred to as the sensitivity or adaptability of a variety. The basic ANOVA model for two-way crossed classification with interaction serves to obtain a rough idea about the partition of variance over different terms. However, it identifies the interaction as a source but does not analyze it since the interaction here is modeled by a separate, additive parameter for each combination of genotype by environment coarsely and un-parsimoniously. No attempt is made at the interpretation of this interaction, leaving the causes of interaction.

As an alternative to linear formulations of interactions, multiplicative formulations may be chosen in an attempt to quantify the variety's contribution to genotype × environment interaction. These

[*] *Corresponding author* : B.M.K. Raju
E-mail addresses : bmkraju@yahoo.com, bmkraju@gmail.com

multiplicative formulations permit the interpretation of interaction as differential genotypic sensitivity to environmental variable(s). Whenever the information on external environmental characteristics such as weather parameters and soil characteristics are available, it may be accommodated in the basic ANOVA model for making an attempt to interpret the interaction. This technique consists of regression of the estimated interactions of each variety on to the measured external environmental variables to obtain the linear sensitivities. However, it is difficult to obtain data on natural environments that comply with the data structure of varieties and properly explain variations in agricultural value of crop varieties. In such cases one may consider the regression of observed yield on the observed environmental mean yield. Finlay and Wilkinson (1963) reasoned that the average yield of a large group of genotypes can be used to describe a complex natural environment, without the complexities of defining or analyzing the important edaphic and seasonal factors. Environment averages, or their deviations from the overall average, are generally used as environmental indices. The resulting regression coefficient may be interpreted as linear sensitivity of the variety to environmental change. This technique was first used by Yates and Cochran (1938) and later by Finlay and Wilkinson (1963) and Eberhart and Russell (1966). This technique is popularly known as Joint Regression, as the joint effect of all the genotypes is used as explanatory regression variable.

Though this technique is very popular among plant breeders, it has got certain limitations. There is a need to elaborate these limitations and suggest alternative measures of stability that do not suffer from such limitations. Another problem faced by plant breeders is non-availability of data on all locations-years and varieties, which makes the data set obtained from Multi Environment Testing (MET) unbalanced. It may be incidental or accidental. The list of varieties being tested changes over the years and not all varieties are tested in all the environments since the genotypes change from year to year. As new varieties become available, older ones become obsolete which makes the data set unbalanced. Some causes for the accidental imbalances are non-germination, damage of crop on account of pests and diseases and floods etc. Literature cites some methods that can handle incomplete data situations. But it is again an issue for the plant breeders to choose the best technique for a given situation. Hence, there is a need to study the existing methodologies to bring out their practical relevance in the current generation of computers and to develop handy computational algorithms for evaluating stability of the varieties. Section 2 focuses on limitations of Eberhart and Russell (1966) model and highlights alternative measures of stability. The subsequent sections elicit on the comparison of various methodologies under incomplete data situations.

2. STABILITY ANALYSIS FOR BALANCED DATA

2.1 Eberhart and Russell (1966) model

Eberhart and Russell (1966) proposed an observational formulation for the Joint Regression context. The model proposed by Eberhart and Russell (1966) is written as

$$y_{ij} = \alpha_i + \beta_i e_j + \delta_{ij}$$

where

y_{ij} is the performance of i-th genotype at the j-th environment ($i = 1, \ldots, K$; $j = 1, \ldots, N$) averaged over R replications

α_i is the mean of i-th genotype over all the environments

e_j is the environmental index for the j-th environment which is obtained as the mean of all genotypes at the j-th environment minus the general mean.

β_i is the regression coefficient measuring the linear sensitivity of i-th genotype to environment change.

δ_{ij} is the 'deviation from regression' of the i-th genotype in the j-th environment

Testing for the significance of genotype environment interaction

The significance of either $G \times E$ (linear) mean squares or pooled deviation mean squares or both when tested against average error confirms the presence of GE interaction. If the latter alone is significant then no useful prediction is possible from this approach. If both are significant then the practical usefulness of the predictions depends on the significance of former relative to the latter.

Stability and adaptability

A genotype with unit regression coefficient i.e. $\beta_i = 1$ and the mean squared deviation not significantly

different from zero ($\overline{S}_{d_i}^2 = 0$) is said to be stable. Significance of $\overline{S}_{d_i}^2$ from zero invalidates the linear prediction. If $\overline{S}_{d_i}^2$ is not significantly different from zero, the performance of the genotype for a given environment may be predicted. Accordingly, a genotype whose performance can be predicted is said to be stable and it also helps in choosing genotypes for specific adaptation.

Eberhart and Russell model analysis for groundnut data

The data used in this study has been provided by Regional Agricultural Research Station (RARS), Palem of Acharya N.G. Ranga Agricultural University, Andhra Pradesh. The data was an outcome of multi location trials of released and pre-released varieties of groundnut conducted at research stations situated in different agro-climatic zones of Andhra Pradesh. The data was consisted of 15 varieties of groundnut viz., TPT-1, TPT-2, Girnar-1, ICG (FDRS)-4, ICG (FDRS)-10, K-134, SVGS-1, TCGS-1, TCGS-3, ICGV-86699, Kadiri-3, ICGS-11, ICGS-44, JL-24 and TMV-2. These are designated as G-1 to G-15 respectively. These varieties were grown in Kharif-1990 and Kharif-1991. The locations used for these trials were 14. The 6 locations, namely, RARS-Tirupati, ARS-Utukur, ARS-Darsi, RARS-Nandyal, ARS-Seethampet, RARS-Palem were used for Kharif-1990 as well as Kharif-1991. They are designated as E-1 to E-6 for Kharif-1990 and E-9 to E-14 for Kharif-1991. The 2 locations, namely, ARS-Kadiri and RARS-Jagitial were used in Kharif-1990 only. These were designated as E-7 and E-8. Remaining 6 locations, namely, ARS-Ananthapur, ARS-Peddipalli, ARS-Peddapuram, RARS-Yellamanchili, ARS-Ragolu, ARS-Vizayanagaram were used in Kharif 1991 only. These were designated as E-15 to E-20. The structure of environments is as under.

Location-Year combination is treated as environment and accordingly 20 environments are designated as E-1 to E-20. The experiments were laid in Randomized Block Design (RBD) with 3 replications. The pod yields were expressed as kg/ha. The mean data over the replicates for the 15 genotypes and 20 environments is given in Raju (2002).

The stability statistics of Eberhart and Russell's model are presented in Table 1. The results revealed that there was significant difference among the genotypes indicating wider genetic diversity among the genotypes. Genotype × Environment (linear) and pooled deviation were found to be significant when tested against pooled error. It indicated significant Genotype × Environment interaction. Genotype × Environment (linear) interaction was found to be not significant when tested against pooled deviation which implies that the genotypes do not differ for their regression on environmental index and overwhelming portion of G × E interaction is of non-linear type, which ultimately makes the behaviour of genotypes unpredictable. On examining the significance of deviation from linear regression for the 15 genotypes in Table 1, all the deviations are significant at 1% level except genotype-7 and genotype-14. The deviation for the genotype-14 is not significant and the regression coefficient β_i is around unity (0.921) and as such it is regarded as stable variety. Similarly, the deviation for genotype-7 is not significant at 1% level and the coefficient of linear sensitivity is very close to unity, hence this can also be regarded as stable variety. Genotype-6 tops with respect to the average yield over the environments. However, the significance of deviation from linear regression makes its behaviour unpredictable over the environments and one may not be able to comment on its stability from Eberhart and Russell model's point of view.

Location Year	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Kharif 1990	E-1	E-2	E-3	E-4	E-5	E-6	E-7	E-8	×	×	×	×	×	×
Kharif 1991	E-9	E-10	E-11	E-12	E-13	E-14	×	×	E-15	E-16	E-17	E-18	E-19	E-20

Limitation of Eberhart and Russell (1966) model with regard to making comments on the stability of each genotype

Eberhart and Russell (1966) model proves to be a good tool for understanding the nature and type of GE interaction exhibited by the data set in the sense that whether a dominating portion of it, is linear or non linear type. The scope of stability parameters of Eberhart and Russell (1966) model (β_i and $\overline{S}^2_{d_i}$) is limited as it may not be possible to comment on the stability of each of the genotypes tested. Whenever the component of deviation from linear regression is found to be significant for a genotype, then it is not possible to make any comments about the stability of that genotype. A genotype may possibly be stable, but due to the fact that its interaction with environments is not of linear type, one becomes handicapped to make any comments on its stability. Further, if there is no information about the stability of some genotypes, one cannot make any comparison among all the genotypes with respect to stability. The scope of the investigation

Table 1. Stability analysis results for balanced data

Eberhart and Russell Model Statistics				Stability Variance Statistics	
Source	df	MS	$\hat{\beta}_i$	Stability variance	Stability rank
Genotypes	14	254686 **	-	-	-
Env + Gen × Env	285		-	-	-
Env (linear)	1		-	-	-
Gen × Env (linear)	14	62925 ^{NS}	-	-	-
Pooled deviation	270	90839	-	-	-
G-1	18	58480 **	1.034	57030	6
G-2	18	40652 **	0.961	37716	3
G-3	18	165738 **	1.096	177785	13
G-4	18	211461 **	1.100	228151	14
G-5	18	227958 **	1.110	247177	15
G-6	18	51799 **	1.109	54365	5
G-7	18	35235 *	0.998	31153	2
G-8	18	57064 **	0.918	57951	7
G-9	18	83747 **	1.074	86689	10
G-10	18	127073 **	1.095	135561	12
G-11	18	52538 **	0.910	53564	4
G-12	18	59875 **	0.942	59582	8
G-13	18	73167 **	0.833	84863	9
G-14	18	30754 ^{NS}	0.921	28965	1
G-15	18	87042 **	0.899	92239	11
Average error	560	20168			

* Significant at 5% level of significance ** Significant at 1% level of significance

will in that case be limited only to a subset of genotypes tested, whose deviations from linear regression are found to be not significant. Pair-wise comparisons with respect to stability can be made in that subset only. Ranking of all the genotypes tested with regard to stability will not be possible. In these situations one thus explores some stability measures on the lines of stability variance given by Shukla (1972). This, however, is not based on linear regression model. As a result, some limitations which are inherent in linear regression model, can be overcome. Nevertheless it becomes inevitable to look for alternatives when component of deviation from linear regression is found significant for more number of genotypes. From the results in Table 1, it is seen that the component of deviation from linear regression is significant at 1% level of significance for all genotypes except G-7 and G-14. For such data sets, it is cautioned that plant breeders may not end up their investigation for stability with Eberhart and Russell model alone.

2.2 Stability Variance — An Alternative Measure of Stability

Stability variance of *i*-th variety given by Shukla (1972) measures the variance of interaction residuals of i-th variety. The genotype with smallest stability variance was the most stable among the genotypes tested. The genotype with second smallest stability variance was the second most stable among the genotypes tested.

Stability variance of *i*-th variety given by Shukla (1972) is

$$\hat{\sigma}_i^2 = \frac{1}{(K-1)(K-2)(N-1)}\left[K(K-1)W_i - \sum_{s=1}^{K} W_s\right]$$

where $W_i = \sum_j (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} - \bar{y}_{..})^2$ is the Wricke's ecovalence for the i-th genotype. It is shown that for balanced data Shukla's estimator is a MINQUE (Minimum Norm Quadratic Unbiased Estimator) of σ_i^2. It is obvious that Shukla's estimator of stability variance is equivalent to Wricke's ecovalence W_i for ranking purposes.

Shukla model analysis for groundnut data

Stability variances were computed for the 15 genotypes of groundnut. The genotypes were ranked with respect to their stability. The results are presented in Table 1. Genotype-14 is found to have maximum stability in pod yield whereas Genotype-5 is found to have least stability variance among the 15 varieties tested. Obviously there is no restriction with this stability variance measure while making stability comparisons among the varieties tested. This stability measure can capture nonlinear interactions too. This measure permits to make stability comparisons among the 15 varieties tested unlike the Eberhart and Russell (1966) model, where one can only make stability comparisons between varieties 7 and 14.

3. STABILITY ANALYSIS WITH INCOMPLETE GENOTYPE BY ENVIRONMENT DATA

The two stability approaches (i) Joint Regression, and (ii) Stability Variance for incomplete data situation are discussed as under.

3.1 Joint Regression Approach

When the yields of some of the genotypes are not available or are not reliable, then the orthogonality of the design is not satisfied and bias is introduced in the observed varietal means. The comparison based on these means is likely to favour the varieties which happen to be exposed to better than average environmental conditions. Hence before proceeding to evaluate stabilities, such compensation needs to be made in the means for the environments in which particular varieties are not present. This section describes two such procedures, namely, (i) Joint Regression with Fitcon estimates, and (ii) Modified Joint Regression. Though it is established that the latter one is a more generalized one, but to establish the superiority of the latter the details of methodology is outlined. An iterative algorithm has been developed for the latter procedure, which would be very handy for the programmers attempting to evaluate it.

3.1.1 Joint regression with Fitcon estimates

Fitcon Analysis: The usual method of obtaining the aforesaid compensation or adjustment is to use a fitting constants technique, described by Patterson (1978), for the additive model

$$y_{ij} = \alpha_i + e_j + \varepsilon_{ij} \qquad (3.1)$$

where

y_{ij} is the (average) yield of i-th variety in j-th environment.

α_i is the mean of i-th variety.

e_j is the effect of j-th environment.

ε_{ij} is the random error, distributed normally with mean zero and a constant variance.

For estimating the parameters α_i and e_j, we have to minimize the residual sum of squares $\sum_{i,j}\left(y_{ij}-\alpha_i-e_j\right)^2\delta_{ij}$ with respect to both α_i and e_j; noting that the weights δ_{ij}, introduced to obtain the incomplete data set-up are such that

$$\delta_{ij}=\begin{cases}1 \text{ if } y_{ij} \text{ is present in the data}\\ 0 \text{ if } y_{ij} \text{ is missing}\end{cases}$$

The iterative algorithm to solve for the parameters of (3.1) is as under

$$\hat{\alpha}_i=\overline{y}'_{i.}+\sum_{j^*}\frac{e_j}{n_i} \tag{3.2}$$

$$\hat{e}_j=\overline{y}'_{.j}-\sum_{i}\frac{\delta_{ij}\alpha_i}{n_j} \tag{3.3}$$

where $\overline{y}'_{i.}$, $\overline{y}'_{.j}$ are the means based on the existing n_i and n_j observations for the i-th variety and j-th environment respectively. Adjustments for these estimates depend on each other's final estimates. Summation over j^* is for those environments where the i-th variety is found to be absent.

Firstly, start the iteration by considering the trial value $\overline{y}'_{i.}$ for $\hat{\alpha}_i$ in (3.3) giving rise to a set of e_j values. Substitution of these values in (3.2) gives rise to revised estimates of α_i's. These are then substituted in (3.3) to get the revised estimates of e_j's. This cycle is continued till we reach more or less stable values for α_i and e_j.

During 1970s, computation of inverse of matrices of higher dimension was indeed a difficult task. Majority of researchers, during that period, were busy in deriving numerical techniques that could yield an approximate solution to normal equations. In those days, the above iterative algorithm could prove to be very handy. But it becomes redundant as on today in the light of advanced software and hardware technology. The current generation of computers can invert a matrix of any dimension in no time. With incomplete data set considered in this study, subroutine of SAS statistical software namely *lsmeans*[*] produced almost same results as produced by Patterson's (1978) Fitting constants technique.

The stabilized Fitcon estimates for e_j can be used for Joint Regression. The linear sensitivity for each variety can be estimated by regressing the existing y_{ij}'s on the Fitcon estimates of e_j's as shown below

$$y_{ij} = \tau_i + \beta_i \hat{e}_j + \delta_{ij}$$

However, the estimates of τ_i are not the same as the estimates of α_i except for the varieties that are present in all the environments or that have unit sensitivities. The discrepancy arises from the fact that the adjustment given in (3.2) is made to be of same degree for every variety; i.e. there is no allowance for varieties differing in their response or sensitivity to environmental effect. When such differences are expected, it is better to incorporate the parameter β_i in the adjustment as

$$\hat{\alpha}_i=\overline{y}'_{i.}+\beta_i\left(\sum_{j^*}\frac{e_j}{n_i}\right)$$

This leads to the consideration of the non-additive model

$$y_{ij} = \alpha_i + \beta_i e_j + \varepsilon_{ij}$$

Digby (1979) proposed this improved adjustment in his modified Joint Regression analysis for incomplete variety by environment data.

3.1.2 Modified Joint Regression

The model considered by Digby (1979) is

$$y_{ij} = \alpha_i + \beta_i\, e_j + \varepsilon_{ij} \tag{3.4}$$

Minimization of residual sum of squares $\sum_{i,j}\left(y_{ij}-\alpha_i-\beta_i e_j\right)^2\delta_{ij}$ with respect to parameters α_i, β_i and e_j leads to the following normal equations

[*] *Ismeans routine of SAS statistical software under ANOVA procedure produces the least square means of effects specified.*

$$\sum_j \delta_{ij} y_{ij} = \alpha_i \sum_j \delta_{ij} + \beta_i \sum_j \delta_{ij} e_j \qquad (3.5)$$

$$\sum_j \delta_{ij} e_j y_{ij} = \alpha_i \sum_j \delta_{ij} e_j + \beta_i \sum_j \delta_{ij} e_j^2 \qquad (3.6)$$

$$\sum_i \delta_{ij} \beta_i y_{ij} = \sum_i \delta_{ij} \beta_i \alpha_i + e_j \sum_i \delta_{ij} \beta_i^2 \qquad (3.7)$$

where δ_{ij} has the same meaning as given earlier.

Since equations (3.5), (3.6) and (3.7) are not linearly independent, they are to be solved subject to the constraint $\sum_j e_j = 0$. To solve for the parameters of (3.4) from (3.5), (3.6) and (3.7) subject to $\sum_j e_j = 0$, the following iterative algorithm is proposed

Step 1: Set the β_i's equal to one, which reduces the equations (3.5) and (3.7), the solutions of which ($\hat{\alpha}_i$ and $\hat{e}_j$) subject to constraint $\sum_j e_j = 0$ can be obtained from the iterative algorithm given in the equations (3.2) and (3.3)

Step 2: Substitute the estimates of e_j in equations (3.5) and (3.6) and obtain the estimates of β_i

Step 3: Substitute the estimates of β_i in equations (3.5) and (3.7). Treat β_i as fixed

Step 3a: Estimate α_i as $\hat{\alpha}_i = \overline{y}'_{i.} + \beta_i \left(\sum_{j^*} \frac{e_j}{n_i} \right)$

Step 3b: Using the estimate of α_i obtained in step 3a, solve for the estimates of e_j as

$$\hat{e}_j = \frac{\sum_i \delta_{ij} \beta_i y_{ij} - \sum_i \delta_{ij} \beta_i \alpha_i}{\sum_i \delta_{ij} \beta_i^2}$$

Step 3c: Go to step 3a, till there is convergence in $\hat{\alpha}_i$ and $\hat{e}_j$

Step 4: Go to step 2, till there is convergence in $\hat{\alpha}_i$ $\hat{\beta}_i$ and $\hat{e}_j$

This algorithm is very handy for the programmers doing the analysis work.

3.2 Stability Variance Approach

Piepho (1994) proposed a procedure for estimating stability variance σ_i^2, when some cells in two-way table are empty. It is outlined as under

Let $x_{srj} = y_{sj} - y_{rj}$

and $V_{s-r}^2 = \frac{1}{N-1} \left[\sum_j x_{srj}^2 - \frac{(\sum_j x_{srj})^2}{N} \right]$

where N is the number of environments in which the genotypes s, r are grown together.

We know that $E[V_{s-r}^2] = \sigma_s^2 + \sigma_r^2$ where $s = 1, 2,\ldots, (K-1)$ and $r > s$.

In order to estimate σ_i^2, the method of moments may be employed where the sample moments are equated to population moments. Replacement of $E[V_{s-r}^2]$ by V_{s-r}^2 may lead to the following system of equations, to be solved for σ_i^2

$$V_{s-r}^2 = \sigma_s^2 + \sigma_r^2 \qquad [s = 1, 2,\ldots, (K-1) \text{ and } r > s]$$

There are $K(K-1)/2$ different equations in K unknowns, so that for $K > 3$ there are more equations than there are unknowns. Grubbs' estimates are the least squares solutions of these equations (Jaech 1985).

Formally the system of equations can be represented in matrix notation as

$$\mathbf{Q}\sigma = \mathbf{V} \qquad (3.8)$$

where σ is a K dimensional vector of σ_i^2's, $\boldsymbol{V}$ is $K(K-1)/2$ dimensional vector of V_{s-r}^2's and $\boldsymbol{Q}$ is a $K(K-1)/2 \times K$ matrix with elements 0 and 1, that picks the appropriate σ_i^2's.

$\boldsymbol{Q'Q}$ has full rank and thus can be inverted.

The solution of equation (3.8) is

$$\tilde{\tilde{\sigma}} = (\boldsymbol{Q'Q})^{-1}\boldsymbol{Q'V} \qquad (3.9)$$

Grubbs' estimates are unbiased. If we take expectation on both sides of equation (3.9)

$$E[\tilde{\tilde{\sigma}}] = E[(\boldsymbol{Q'Q})^{-1}\boldsymbol{Q'V}] = E[(\boldsymbol{Q'Q})^{-1}\boldsymbol{Q'Q}\sigma]$$
$$= E[\boldsymbol{I}\sigma] = \sigma$$

For 2 genotypes *s* and *r*, we can compute V^{2}_{s-r} as long as they are grown together in atleast two environments. In this case, the 2 genotypes *s* and *r* are said to be connected. To obtain a unique solution of equation (3.8), we require that there be atleast *K* connected pairs of genotypes as we need atleast as many equations as there are unknowns. Also each genotype must be connected to atleast one other genotype.

Comparison of potential of various methods for incomplete data

The potential of a given method (for incomplete data) may be judged by its ability to reproduce the stability/sensitivity rank order given by the method with complete data. The Coefficient of Spearman's Rank Correlation between rank orders displayed by balanced and unbalanced data using that method may be used to quantify the potential. Comparison of various methods can be done with the help of their computed potentials.

Empirical study with unbalanced data

To evaluate the methodologies described for missing data, unbalancedness is created by eliminating 20 cells at random in the 15 × 20 matrix of genotype by environment yields. This unbalanced data has been analysed for Patterson's fitting constants, Joint Regression with Fitcon estimates and Digby's modified Joint Regression analysis with the iterative algorithms described earlier. The resulting parameter estimates are obtained and are given in Table 3.

On comparison of the unadjusted varietal means and the varietal means obtained from Patterson's Fitcon method, it is seen from equation (3.2) that if the variety of interest is absent in the positive environments, the adjustment is made upwards and vice versa. The sign and amount of adjustment is determined by the sum of environmental effects in which the variety of interest is absent. This sum of environmental effects can be obtained from Table 2. In this way the adjustment for variety-1 is made upwards which is absent in the environment-17 having the effect 457.83. Similarly, the correction for variety-2 is positive whereas the adjustment for variety-8 is negative, which is absent in E-5 and E-10. The adjustment for variety-6 is zero as it is present in all the environments.

Table 2. Estimated Environment effects with incomplete data

Environment	Fitcon	Modified Joint Regression
1	180.66	175.42
2	–359.02	–364.43
3	1361.18	1368.64
4	667.00	657.58
5	–308.73	–314.38
6	–138.20	–126.95
7	–187.75	–192.01
8	–912.38	–900.51
9	–47.85	–46.62
10	–918.90	–906.07
11	808.69	825.25
12	348.50	338.16
13	–446.28	–447.78
14	368.05	359.58
15	–83.74	–83.66
16	776.38	795.08
17	457.83	436.28
18	–535.52	–539.29
19	–657.34	–661.97
20	–372.60	–372.31

If one wants to compare the varietal means obtained from Joint Regression with Fitcon estimates and Patterson's Fitcon means, one has to study the adjustment given by Patterson's Fitcon method for the unadjusted means and the improvement offered by the Joint Regression with Fitcon estimates to Fitcon means by allowing the varieties to differ in their sensitivities to the environmental effect.

$$\hat{\alpha}_{i_{(f)}} = \overline{y}'_{i.} + \sum_{j^*} \frac{\hat{e}_j}{n_i}$$

$$\hat{\alpha}_{i_{(jf)}} = \overline{y}'_{i.} + \beta_i \left(\sum_{j^*} \frac{\hat{e}_j}{n_i} \right)$$

Table 3. Estimated variety parameters with incomplete data

Variety	Unadjusted mean	Fitcon mean	Joint Regression with Fitcon estimates		Modified Joint Regression analysis		Stability Variance Statistics	
			Mean	Sensitivity	Mean	Sensitivity	Stability variance	Stability rank
1	1478.68	1502.78	1503.32	1.023	1502.13	1.021	43976	6
2	1250.05	1321.69	1321.53	0.998	1321.95	0.998	39826	4
3	1422.79	1420.27	1420.03	1.094	1420.11	1.091	172119	13
4	1243.22	1296.39	1299.84	1.065	1301.22	1.076	235535	14
5	1325.37	1309.12	1307.24	1.116	1306.76	1.125	256158	15
6	1692.75	1692.75	1692.75	1.115	1692.75	1.115	57442	7
7	1481.58	1474.31	1474.29	1.003	1474.90	1.001	35030	3
8	1351.72	1283.52	1288.37	0.929	1288.90	0.927	27559	2
9	1390.53	1380.65	1379.82	1.083	1379.60	1.082	92553	10
10	1459.35	1401.81	1400.34	1.026	1399.55	1.030	109499	12
11	1338.50	1362.73	1361.63	0.955	1362.52	0.955	43863	5
12	1415.78	1472.20	1467.19	0.911	1465.98	0.908	59030	8
13	1480.58	1473.31	1474.50	0.836	1475.02	0.832	85740	9
14	1437.47	1456.85	1455.62	0.937	1455.16	0.935	26814	1
15	1324.79	1276.77	1282.45	0.882	1283.37	0.874	101612	11

where $\bar{y}'_{i.}$ is unadjusted mean; $\hat{\alpha}_{i_{(f)}}$ is Fitcon mean; and $\hat{\alpha}_{i_{(jf)}}$ is the mean obtained from the Joint Regression with Fitcon estimates.

One may identify the following 6 cases to study the adjustment offered to the Fitcon means by the Joint Regression with Fitcon estimates.

1. When $\sum_{j^*} \hat{e}_j$ is negative (unadjusted means are corrected downwards) and $\beta_i < 1$: The adjustment to the Fitcon means are positive, e.g. variety-8.
2. When $\sum_{j^*} \hat{e}_j$ is negative and $\beta_i > 1$: Fitcon means are corrected downwards, e.g. variety-5.
3. When $\sum_{j^*} \hat{e}_j$ is positive (unadjusted means are corrected upwards) and $\beta_i < 1$: The adjustment to the Fitcon means is negative, e.g. variety-12.
4. When $\sum_{j^*} \hat{e}_j$ is positive and $\beta_i > 1$: Fitcon means are corrected upwards, e.g. variety-4.
5. When $\sum_{j^*} \hat{e}_j$ is zero, i.e. the variety of interest is absent in none of the environments: The adjustment given by Fitcon means as well as the adjustment given by the Joint Regression with the Fitcon estimates to the unadjusted means are zero e.g. variety-6.
6. When $\beta_i \cong 1$ i.e. the sensitivity is close to unity: The adjustment made to the Fitcon means by the Joint Regression with Fitcon estimates would be negligible, as Joint Regression with Fitcon estimates reduces to the Fitcon method, e.g. variety-7, variety-2.

One may also compare the means obtained from Digby's Modified Joint Regression with unadjusted (observed) varietal means, Fitcon means and the means obtained from Joint Regression with Fitcon estimates.

Unadjusted mean vs mean obtained from Digby's modified Joint Regression

If $(\sum_{j^*} \hat{e}_j)_D$ is positive, the unadjusted mean is corrected upwards and vice versa, where $(\sum_{j^*} \hat{e}_j)_D$ is the sum of environment effects in which the variety of interest is absent and obtained from Digby's modified Joint Regression analysis.

Fitcon means vs Digby's modified Joint Regression means

The improvement in varietal means offered Digby's modified Joint Regression over Patterson's Fitcon is determined by the quantity

$$\beta_i \left(\frac{\sum_{j^*} \hat{e}_j}{n_i} \right)_D - \left(\frac{\sum_{j^*} \hat{e}_j}{n_i} \right)_P = Q \text{ (say)}$$

where $\left(\frac{\sum_{j^*} \hat{e}_j}{n_i} \right)_P$ is adjustment offered by Patterson's Fitcon method to the unadjusted mean. If $Q > 0$ then Fitcon means are corrected upwards and vice versa.

Joint Regression with Fitcon estimates vs modified Joint Regression

In Joint Regression with Fitcon estimates, the e_j's are merely unweighted means of $(y_{ij} - \alpha_i)$ for those varieties present in the j-th environment, whereas modified Joint Regression estimates e_j's as weighted means of $(y_{ij} - \alpha_i)$, the weights being proportional to the varietal sensitivities. Hence, the weighted means used by the iterative analysis are more appropriate.

For variety-15, $Q = -41.42 - (-48.02) = 6.6$. Hence the adjustment offered by modified Joint Regression to the Fitcon mean for variety-15 is +6.6.

Comparison between Joint Regression with Fitcon estimates and modified Joint Regression and stability variance

Joint Regression with Fitcon estimates and modified Joint Regression techniques are compared with respect to their ability to assess the sensitivity rank order obtained with balanced data. The association between the two has been quantified by Spearman's rank correlation. On observing the same sensitivity rank order with Joint Regression with Fitcon estimates and Digby's modified Joint Regression, the correlation of this rank order with the one obtained from balanced data using Eberhart and Russell (1966) model is 0.95. The result re-establishes that choice between Joint Regression with Fitcon estimates and modified Joint Regression is less critical when the varieties do not differ much in regard to sensitivities to environmental index.

Stability variance for the 15 genotypes of groundnut was evaluated using the methodology proposed by Piepho (1994) for incomplete two-way data. The genotypes were ranked with respect to their stability. The corresponding results are presented in Table 3. The rank correlation between the stability rank orders displayed by complete and incomplete data is found to be 0.9429, which is reasonably good concordance. So the researchers and plant breeders can use the measure of Stability Variance safely even in the case of incomplete data. This clearly indicated that stability variance measure is a robust measure of stability of crop variety.

As far as choice between Joint Regression and Stability Variance is concerned, it is to be kept in mind that stability rank order displayed by Joint Regression (based on deviation of sensitivity coefficient from 1) is not necessarily the true stability rank order of varieties. It becomes true stability rank order only when the components of deviation from linear regression are found to be not significant for all the tested varieties. This kind of problem, however, does not arise in case of Stability Variance measure.

4. CONCLUSION

In view of the above results, it is concluded that it is always better to employ the Stability Variance measure to evaluate the stability of a set of genotypes when the component of deviation from linear regression is found significant for sufficiently large number of varieties with Eberhart and Russell model. For incomplete data situation also it is preferable to use Stability Variance approach in place of Joint Regression as it gives stability rank order rather than conditional sensitivity rank order.

ACKNOWLEDGEMENTS

Authors are very much thankful and grateful to referees for their constructive suggestions for the improvement of the contents of the paper.

REFERENCES

Digby, P.G.N. (1979). Modified joint regression analysis for incomplete variety × environment data. *Cambridge J. Agril. Sci.,* **93**, 81-86.

Eberhart, S.A. and Russell, W.A. (1966). Stability parameters for comparing varieties. *Crop Sci*., **6**, 36-40.

Finlay, K. and Wilkinson, G.N. (1963). The analysis of adaptation in a plant-breeding programme. *Austr. J. Agric. Res.,* **14**, 742-754.

Jaech, J.L. (1985). *Statistical Analysis of Measurement Errors.* Wiley, New York.

Patterson, H.D. (1978). Routine least squares estimation of variety means in incomplete tables. *J. National Instt. Agril. Botany*, **14**, 385-400.

Piepho, H.P. (1994). Missing observations in the analysis of stability. *Heredity*, **72**, 141-145.

Raju, B.M.K. (2002). A study on AMMI model and its biplots. *J. Ind. Soc. Agril. Statist*., **55(3)**, 297-322.

Shukla, G.K. (1972). Some statistical aspects of partitioning genotype environmental components of variability. *Heredity*, **29**, 237-245.

Yates, F. and Cochran, W.G. (1938). The analysis of groups of experiments. *Cambridge J. Agril. Sci.,* **28**, 556-580.

Available online at www.isas.org.in
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(2) 2009 151-157

Spatial Smoothing Technique in Field Experiments

C.T. Jose[1*], Ravi Bhat[1], B. Ismail[1] and S. Jayasekhar[2]
[1]*Central Plantation Crops Research Institute, Regional Station, Vittal, Karnataka*
[2]*Mangalore University, Mangalagangotri, Karnataka*

(Received: June 2007, Revised: March 2009, Accepted: July 2009)

SUMMARY

We generally use block designs in field experiments to control the experimental error due to positional variations. The underlying assumption in classical block designs that the homogeneity of experimental area within the block may not satisfy always, particularly when the block size is large. Also we may not know in advance the soil fertility gradient and other factors influencing the response variable to divide the experimental area into homogeneous blocks. We propose spatial smoothing technique to estimate/eliminate positional effect in field experiments. We have considered a semiparametric regression model with treatment effect as the parametric component and the positional effect as the nonparametric spatial function. The only assumption about the positional effect is that it is a smooth spatial function. The proposed method is also extended for the analysis of data in the presence of sudden shifts in the spatial function (positional effect). The method is illustrated through both simulated as well as field experimental data.

Keywords: Nonparametric regression, Design of experiments, Positional effect, Semiparametric regression, Jump regression surface.

1. INTRODUCTION

Experimental error or the unexplained variation is the main concern in field experimentation technique. We generally use block designs in field experiments to control the experimental error due to positional variations. The underlying assumption in classical block designs regarding the homogeneity of experimental area within the block may not satisfy always, particularly when the block size is large. Field experiments with perennial tree crops require large experimental area, and it is grown mainly in hilly areas where getting large homogeneous area is difficult. Also we may not know in advance the soil fertility gradient and other factors influencing the response variable to divide the experimental area into homogeneous blocks. Gilmour *et al.* (1997) suggested the covariance modeling technique to tackle this problem. In the present study, nonparametric spatial modeling technique has been used to estimate/eliminate the positional effect in agricultural field experiments. The treatment effect is taken as the parametric component and the positional effect (covariate) is taken as a spatial (bivariate) nonparametric function. The only assumption about the positional effect is that it is a smooth spatial function. The field experiments with perennial or tree crops require large experimental area and it is difficult to get large homogeneous blocks to conduct experiments particularly in farmer's field. In many situations, the soil characters or the environmental variables have some sudden changes in the field or in other words, the spatial function representing the positional effect may have some jumps or discontinuities. The method is extended for the analysis of data in the presence of

[*] *Corresponding author* : C.T. Jose
E-mail address : ctjos@yahoo.com

sudden jumps in the spatial function. The proposed method is applied to both the simulated as well as field experimental data to see its performance.

2. MODEL SETTINGS AND ESTIMATORS

The semiparametric regression model considered for the study is given by

$$\boldsymbol{Y} = \mu + \boldsymbol{X}\beta + f(U, V) + \varepsilon \qquad (1)$$

where $Y = [y_1\ y_2 \ldots . y_n]^T$ is the observation vector, μ is the general mean, $X = [x_1\ x_2 \ldots x_n]^T$ is the design matrix, $\beta = [\beta_1 \beta_2 \ldots \beta_p]^T$ is the treatment effect vector, $f(U,V)=[f(u_1,v_1)\ \ldots\ f(u_n,v_n)]^T$ is the nonparametric spatial function representing the positional effect and ε is the independently and identically distributed (iid) random error vector with mean zero. It is assumed that $f(U,V)$ is a smooth function. The backfitting algorithm of Hastie and Tibshirani (1990) is used to compute the estimates for the semiparametric regression model. The backfitting estimators for β and f are equivalent to

$$\hat{\mu} = \bar{Y}\ ,\ \hat{\beta} = (\boldsymbol{X}^T(I - S)\ \boldsymbol{X})^{-1}\ \boldsymbol{X}^T(I - S)\ (\boldsymbol{Y} - \hat{\mu}\ )$$

$$\text{and}\ \ \hat{f}\ = \boldsymbol{S}(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta} - \hat{\mu}\ )$$

where, S is the smoothing matrix derived using local linear regression (Ruppert and Wand 1994). Let S_{uv} be the row of the smoother matrix correspond to the smoother vector S_{uv}^T evaluated at the observation point $(u,v) = (u_1,v_1), (u_2,v_2), \ldots , (u_n,v_n)$. Then

$$\boldsymbol{S} = [S_{u_1 v_1} \ldots S_{v_n v_n}]^T$$

where, $\boldsymbol{S}_{uv}^T = e_1^T (Z_{uv}^T W_{uv} Z_{uv})^{-1} Z_{uv}^T W_{uv}$

with $\mathbf{Z}_{uv} = \begin{bmatrix} 1 & (u_1 - u) & (v_1 - v) \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & (u_n - u) & (v_n - v) \end{bmatrix}$, $e_1^T = [1\ 0\ 0]$

and $\boldsymbol{W}_{uv} = diag\left\{ K\left[\left(\frac{u_1 - u}{h_1}\right), \left(\frac{v_1 - v}{h_2}\right)\right], \ldots, K\left[\left(\frac{u_n - u}{h_1}\right), \left(\frac{v_n - v}{h_2}\right)\right]\right\}$ for some bivariate kernel functions K and bandwidths h_1 and h_2 .

Under the assumption that the treatments are allotted at random to the spatial locations, it can be shown that $\hat{\beta}$ is asymptotically unbiased and its asymptotic variance is $\sigma^2(\boldsymbol{X}^T \boldsymbol{X})^{-1}$ which is same as when the model is fully parametric (Opsomer and Ruppert 1999). An estimate of σ^2 is given by

$$\hat{\sigma}_1^2 = \frac{1}{(n - p - 1 - trace(S))}\left[\boldsymbol{Y} - \hat{\mu} - \boldsymbol{X}\hat{\beta} - \hat{f}\right]^T \times \left[\boldsymbol{Y} - \hat{\mu} - \boldsymbol{X}\hat{\beta} - \hat{f}\right]$$

The variance of $\hat{\beta}$ is estimated by

$$\hat{V}(\hat{\beta}) = PP^T \hat{\sigma}_1^2$$

where, $P = (\boldsymbol{X}^T (I - \boldsymbol{S})\ \boldsymbol{X}\ )^{-1}\ \boldsymbol{X}^T(I - \boldsymbol{S})$. The significance of the positional effect f is tested using the lack of fit statistic by comparing parametric and nonparametric models (Hart 1997).

Under the null hypothesis that the positional effect $f\,(U, V) = 0$, the mean residual sum of squares obtained by fitting model (1) is given by

$$\hat{\sigma}_0^2 = (Y - \hat{\mu})^T\ [(\mathrm{I} - \boldsymbol{X}(\boldsymbol{X}^T\ \boldsymbol{X})^{-1}\ \boldsymbol{X}]^T \times [(\mathrm{I} - \boldsymbol{X}(\boldsymbol{X}^T\ \boldsymbol{X})^{-1}\ \boldsymbol{X}]\ (\boldsymbol{Y} - \hat{\mu})/(n - p - 1)$$

The lack of fit test statistic is given by

$$R_1 = \frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2}$$

The statistic R_1 asymptotically follows an F distribution with $(n - p - 1)$, $[n - p - 1 -$ trace$(S)]$ degrees of freedom and it can be used for testing the significance of the positional effect.

Additive model for positional effect: In many situations, the number of experimental units is comparatively small and estimating the spatial function using the bivariate smoother will be inadequate. In such situations, bivariate additive model can be fitted instead of the two dimensional spatial function used in model (1). By using bivariate additive function, model (1) becomes

$$\boldsymbol{Y} = \mu + \boldsymbol{X}\beta + f_1(U) + f_2(V) + \varepsilon \qquad (2)$$

where, f_1 and f_2 are the univariate nonparametric functions representing the positional effect in the U and V directions and it is assumed that $\sum f_1(u_i) = \sum f_2(v_i) = 0$.

Let $\boldsymbol{M_1}$ and $\boldsymbol{M_2}$ are the centered smoother matrices corresponding to $\boldsymbol{U}$ and $\boldsymbol{V}$. The backfitting algorithm will provide an explicit solution to the above semiparametric regression model and the estimates are given by

$$\hat{\mu} = \bar{Y},\ \hat{\beta} = (\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{Q})(\boldsymbol{Y}-\hat{\mu})$$

and

$$\hat{f} = \hat{f}_1 + \hat{f}_2 = \boldsymbol{Q}(\boldsymbol{Y}-\hat{\mu}-\boldsymbol{X}\hat{\beta})$$

The matrix $\boldsymbol{Q}$ and the estimates $\hat{f}_1$ and $\hat{f}_2$ are obtained by solving the set of equations

$$\begin{bmatrix} I & M_1 \\ M_2 & I \end{bmatrix}\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \begin{bmatrix} M_1 \\ M_2 \end{bmatrix}\left(Y-\hat{\mu}-\boldsymbol{X}\hat{\beta}\right)$$

$$\hat{f}_1 = \{I-(M_1M_2)^{-1}(1-M_1)\}\ (Y-\hat{\mu}-\boldsymbol{X}\hat{\beta})$$

$$= Q_1\left(Y-\hat{\mu}-\boldsymbol{X}\hat{\beta}\right)$$

$$\hat{f}_2 = \{I-(M_2M_1)^{-1}(1-M_2)\}(\boldsymbol{Y}-\hat{\mu}-\boldsymbol{X}\hat{\beta})$$

$$= Q_2\left(\boldsymbol{Y}-\hat{\mu}-\boldsymbol{X}\hat{\beta}\right)$$

and $\boldsymbol{Q} = \boldsymbol{Q}_1 + \boldsymbol{Q}_2$

An estimate of σ^2 under model (2) is given by

$$\hat{\sigma}_2^2 = \frac{1}{(n-p-trace(Q))}\left[\boldsymbol{Y}-\hat{\mu}-\boldsymbol{X}\hat{\beta}-\hat{f}\right]^T \times\left[\boldsymbol{Y}-\hat{\mu}-\boldsymbol{X}\hat{\beta}-\hat{f}\right]$$

The significance of the positional effect f is tested using the lack-of-fit test statistic

$$R_2 = \frac{\hat{\sigma}_0^2}{\hat{\sigma}_2^2}$$

The test statistic R_2 asymptotically follows an F distribution with (n – p – 1), [n – p – trace (**Q**)] degrees of freedom and it can be used for testing the significance of the positional effect. An approximate α-level point wise confidence band around the estimated function f is given by $\hat{f}(u_i,v_i) \pm z_{\alpha/2}\hat{\sigma}_2\sqrt{[\boldsymbol{QQ}^T]_{ii}}$ for $i = 1,\ldots,n$, where, $[\boldsymbol{QQ}^T]_{ii}$ represents the element in the ii^{th} position of the matrix $[\boldsymbol{QQ}^T]$.

The variance of $\hat{\beta}$ is estimated by

$$V(\hat{\beta}) = \boldsymbol{PP}^T\hat{\sigma}_2^2$$

where, $\boldsymbol{P} = (\boldsymbol{X}^T(I-\boldsymbol{Q})\boldsymbol{X})^{-1}\boldsymbol{X}^T(I-\boldsymbol{Q})$

Choice of bandwidth: The procedure described above involves two smoothing parameters h_1 and h_2. The choice of bandwidth parameters is very crucial in smoothing technique. We have used the cross-validation technique (Hardle 1990) to obtain the optimum bandwidths. Let y_i, $i = 1,\ldots,n$ are the observations and $\hat{y}_{(i)h_1h_2}$ be its leave-one-out estimate (estimate without using the i^{th} observation) with h_1 and h_2 as bandwidths. Then the cross-validation score is defined by

$$CV(h_1, h_2) = \frac{1}{n}\left[\sum_{i=1}^{n}(y_i - \hat{y}_{(i)h_1, h_2})^2\right]$$

The values of h_1 and h_2 which minimize $CV(h_1, h_2)$ can be used as the bandwidths for estimating the regression model.

3. JUMPS IN THE SPATIAL FUNCTION

Sometimes the soil characters or environmental variables have some sudden changes in the field or in other words, the spatial function representing the positional effect has some jumps or discontinuities. In such situations the procedure given in Section 2 needs to be modified. Let us first define a jump in the spatial function f as follows:

$$f(u, v) = g(u, v) + \Delta(u)\, I_{v>c(u)},\ (u, v) \in [0, 1]^2 \quad (3)$$

where, $g(u, v)$ is the continuous part, $c(u)$ denotes the jump location curve and $\Delta(u)$, is the jump magnitude function. The functions g and Δ are assumed to be smooth. The jump location curve $c(u)$ is assumed to have first order derivative. Note that the jump location curve $c(u)$ divides the entire experimental area into two parts. Under the assumption that the treatments are randomly distributed to the entire experimental area, initial estimates for μ and β are given by

$$\hat{\mu} = \bar{Y} \qquad \hat{\beta} = \left[\boldsymbol{X}^T(I-\boldsymbol{S})\boldsymbol{X}\right]^{-1}\boldsymbol{X}^T(1-\boldsymbol{S})\left(\boldsymbol{Y}-\hat{\mu}\right)$$

Let $\boldsymbol{Y}^* = \boldsymbol{Y} - \hat{\mu} - \boldsymbol{X}\hat{\beta}$

We have used the method of Jose and Ismail (2001) to estimate the jump location curve. Define the set $Q_i(u, v)$, $i = 1, ..., 4$ as the set of points in the i^{th} quadrant with respect to the point (u, v). At any point (u, v), consider the following two kernel weighted least squares (minimization) problem:

Minimize

$$\sum_{i=1}^{n} \left\{ y_i^{\#} - b_0 - b_1(u-u_i) - b_2(v-v_i) - a_0(u,v) I \left[(u_i, v_i) \in Q_1(u,v)\right] \right\}^2 I\left[(u_i, v_i) \in Q_1(u,v) \cup Q_3(u,v)\right] K_i \quad (4)$$

Minimize

$$\sum_{i=1}^{n} \left\{ y_i^{\#} - b_0 - b_1(u-u_i) - b_2(v-v_i) - a_0(u,v) I\left[(u_i, v_i) \in Q_2(u,v)\right] \right\}^2 I\left[(u_i, v_i) \in Q_2(u,v) \cup Q_4(u,v)\right] K_i \quad (5)$$

where, $K_l = K\left[\left(\frac{u-u_l}{h_1}\right), \left(\frac{v-v_l}{h_2}\right)\right]$ is some bivariate kernel function.

If the slope of the jump location curve at any $(u, v) \in c$ is negative, then for small bandwidths h_1 and h_2, the points in $Q_1(u, v)$ and $Q_3(u, v)$ will be in the opposite sides of c. Similarly, if the slope of c at (u, v) is positive, the points in $Q_2(u, v)$ and $Q_4(u, v)$ will be in the opposite sides of $c(u)$. The estimates of $a_0(u, v)$ obtained by solving the least squares problems (4) and (5) corresponding to the point (u, v) are denoted by $\hat{a}_{01}(u, v)$ and $\hat{a}_{02}(u, v)$ respectively. Among these two estimates, the estimate with maximum absolute value is denoted by $\hat{a}_0(u, v)$. Then the estimate of the jump location curve is given by

$$\hat{c}(u) = \underset{v \in [h_2,\, 1-h_2]}{\arg\max} \left|\hat{a}_0(u,v)\right|$$

and $\hat{a}_0(u, \hat{c}(u))$ is the estimate of the jump size function $\Delta(u)$ which divides the experimental area into two disjoint sets, say A and B. The estimate $\hat{f}$ of the spatial function f on both sides of $\hat{c}(u)$ can be estimated separately based on the observations on either sides of $\hat{c}(u)$ by the method of kernel weighted local linear regression (Ruppert and Wand 1994).

Let Y_A, Y_B; X_A, X_B; f_A, f_B and S_A, S_B are the observation vectors, design matrices, positional effect vectors and the smoother matrices correspond to the sets A and B respectively. The final estimate of the treatment vector and the rearranged spatial function f^* are given by

$$\hat{\beta}^* = [\boldsymbol{X}^{*T}(I - \boldsymbol{S}^*)\boldsymbol{X}^*]^{-1} \boldsymbol{X}^{*T}(I - \boldsymbol{S}^*)(\boldsymbol{Y}^* - \mu^*)$$

$$\hat{f}^* = \boldsymbol{S}^* \left[\boldsymbol{Y}^* - \mu^* - \boldsymbol{X}^* \hat{\beta}^*\right]$$

where $\boldsymbol{S}^* = \begin{bmatrix} S_A & 0 \\ 0 & S_B \end{bmatrix}$, $\boldsymbol{Y}^* = \begin{bmatrix} Y_A \\ Y_B \end{bmatrix}$, $\mu^* = \begin{bmatrix} \overline{Y}_A \\ \overline{Y}_B \end{bmatrix}$,

$$\boldsymbol{X}^* = \begin{bmatrix} X_A \\ X_B \end{bmatrix}, \boldsymbol{f}^* = \begin{bmatrix} f_A \\ f_B \end{bmatrix}$$

An estimate of the error variance σ^2 is given by

$$\hat{\sigma}^{*2} = \frac{1}{(n - p - 1 - trace(S^*))} [\boldsymbol{Y}^* - \mu^* - \boldsymbol{X}^* \hat{\beta} - \hat{f}^*]^T \times [\boldsymbol{Y}^* - \mu^* - \boldsymbol{X}^* \hat{\beta} - \hat{f}^*]$$

The variance of $\hat{\beta}$ is estimated by

$$V(\hat{\beta}^*) = \boldsymbol{P}^* \boldsymbol{P}^{*T} \hat{\sigma}^{*2}$$

where $P^* = [\boldsymbol{X}^{*T}(I - \boldsymbol{S}^*)\, \boldsymbol{X}^*]^{-1} \boldsymbol{X}^{*T} (\boldsymbol{I} - \boldsymbol{S}^*)$

The above method can be extended to a more general case that the jump location curve does not have the explicit functional form given in (3). Assume that the jump location curve $c(\ )$ induces a partition of the field into disjoint subsets A and B. Then the spatial function f can be defined as

$$f(u, v) = g(u, v) + \Delta(u, v)\, I_B(u, v),\ (u, v) \in [0,1]^2$$

where, g and Δ are smooth functions and $inf\ |\Delta(u, v)| > 0$ for all $(u, v) \in c$. As discussed above obtain $\hat{a}_0(u, v)$ for all $(u, v) \in (h_1, 1 - h_1) \times (h_2, 1 - h_2)$. Note that $|\hat{a}_0(u, v)|$ near c are significantly larger than the others. An estimate of c can be constructed by the

maximin method suggested by Muller and Song (1994). Find the curve that maximizes the minimum of $|\hat{a}_0(u, v)|$ along curves in Γ; that is,

$$\hat{c} = \arg\max_{\phi \in \Gamma} \left[\min_{(u,v)\in\phi} \left| \hat{a}_0(u,v) \right| \right]$$

where, Γ is a sufficiently rich class of candidate boundaries, containing c. Once the jump location curve is estimated, the positional effect on both sides of the estimated jump location curve can be obtained separately.

4. SIMULATION STUDY

A simulation study is carried out to see the performance of the proposed method. We considered the following model for the simulation study

$$\boldsymbol{Y} = \boldsymbol{X}\beta + f(\boldsymbol{U}, \boldsymbol{V}) + \varepsilon \qquad (6)$$

where Y is the n × 1 observation vector, X is the n × n design matrix, β is the k × 1 treatment effect vector which is taken as β'= [−2 −2 0 4], $f(u,v)=2\{2+sin[2(u+v])\}$ and the random error vector ε follows $N(0, \sigma^2)$. The spatial locations of the n observations are obtained by dividing the region [0, 1] × [0, 1] equally and each treatment is allotted randomly to n/k spatial locations. Based on the above, 100 sets of data are simulated for different values of n (100, 400, 900) and σ (0.5, 1.0). The bivariate kernel function considered is $K(u, v)=0.75^2(1 - u^2)(1 - v^2)$ which is the product of two Epanechnikov kernels. The treatment effect vector $\beta^{\mathrm{T}} = [\beta_1\ \beta_2\ \beta_3\ \beta_4]$, the bivariate function f and the error variance σ^2 are estimated using the method given in Section 2. The Average Mean Squared Errors (AMSE) of the estimated values of σ, β and f with the true values of 100 sets of simulated data for different values of n (100, 400, 900) and σ (0.5, 1.0) are given in Table 1. The AMSE of the estimated parameters are calculated as follows:

$$\text{AMSE of, } \hat{\sigma} = \frac{1}{100}\sum_{i=1}^{100}(\sigma - \hat{\sigma}_{(1)})^2$$

$$\text{AMSE of } \hat{\beta}_j = \frac{1}{100}\sum_{i=1}^{100}(\beta_j - \hat{\beta}_{j(i)})^2 \text{ , j =1,..., 4}$$

$$\text{AMSE of } \hat{f} = \frac{1}{100}\sum_{i=1}^{100}\frac{1}{n} \sum_{j=1}^{n}[f(u_j, v_j) - f_{(i)}(u_j, v_j)]^2$$

where, $\hat{\sigma}_{(i)}, \hat{\beta}_{j(i)}$ and $\hat{f}_{(i)}(u_j, v_j)$ are the estimated values of σ, β_j and $f(u_j, v_j)$ corresponding to the i^{th} simulated data set. It can be observed that the AMSE of the estimates are converging to zero as n increases or in other words, the estimated values are converging to the true values as n increases. This indicates the consistency of the estimates. The MSE varies with change in the choice of bandwidths. The optimum bandwidth (bandwidth corresponds to the minimum MSE) will depend on the curvature of the function. The optimum bandwidth for estimating the regression model is obtained based on the cross validation technique given in Section 2.

Table 1. Average Mean Squared Errors (AMSE) of the estimated values with the true values of the simulated data (Model 6)

σ	n	MSE of the estimates multiplied by 100						
		$X\hat{\beta} + \hat{f}$	$\hat{\sigma}$	$\hat{f}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
0.5	100	3.90	0.18	3.51	0.99	0.95	0.65	0.93
	400	1.42	0.03	1.27	0.21	0.15	0.12	0.14
	900	0.82	0.01	0.73	0.08	0.07	0.08	0.08
1.0	100	9.57	0.28	6.28	1.55	1.50	1.34	1.42
	400	3.37	0.12	2.95	0.45	0.41	0.38	0.43
	900	1.34	0.05	1.16	0.17	0.20	0.22	0.19

The performance of the proposed method in the case of sudden shift or jump in the spatial function is illustrated through a simulation study. For this, the regression model (6) is modified as

$$\boldsymbol{Y} = \boldsymbol{X}\beta + f_1(\boldsymbol{U}, \boldsymbol{V}) + \varepsilon \qquad (7)$$

where Y, X , β and ε are as defined in model (6). The bivariate regression function $f_1(u, v)$ is taken as a jump regression surface of the following form

$$f_1(u, v) = 2\{2 + \sin[2(u + v)]\} + [1 + 2\sin(1 + 2u)]I_{ve\ 0.6\ \sin(1+2u)} \qquad (u, v) \in [0,1]^2$$

Based on the above, one set of data is simulated for $n = 900$ and $\sigma = 0.40$, the treatment vector $\beta' = [-2.0\ -2.0\ 0.0\ 4.0]$, the spatial function $f_1(u, v)$ the jump location function $c(u) = 0.6\ \sin(1 + 2u)$ and the jump magnitude function $\Delta(u) = 1 + 2\sin(1 + 2u)$. The treatment effect vector β, the error variance σ^2, the jump location curve $c(u)$ and jump magnitude function $\Delta(u)$ are estimated using the method given in Section 3. The estimated values of β and σ are respectively $\hat{\beta} = [-1.95\ -1.97\ -0.06\ 3.99]$ and $\hat{\sigma} = 0.44$ which are very close to the true values. The jump location function and jump magnitude function are obtained by smoothing the point wise estimates of the jump location curve and jump size function. The estimated and true values of the jump location curve and jump magnitude function are shown in Fig. 1 and 2 respectively. It can be noted that the estimated and true values are very close.

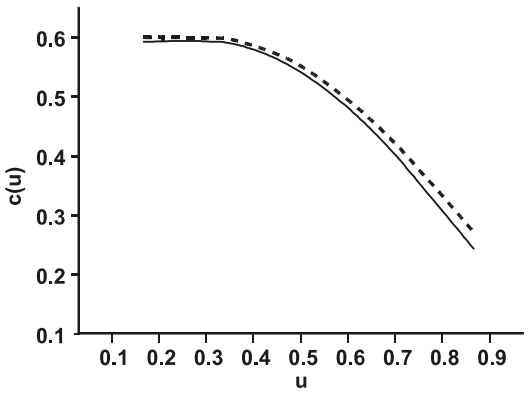

Fig. 1. Estimated (dotted line) and true values (solid line) of the jump location function

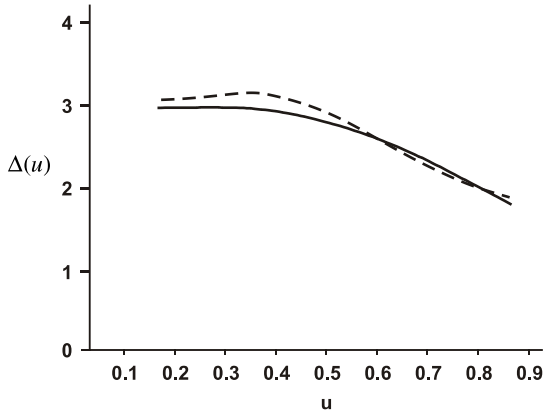

Fig. 2. Estimated (dotted line) and true values (solid line) of the jump magnitude function

5. FIELD APPLICATION

The proposed spatial technique is applied to the data of irrigation cum fertilizer trial of cocoa + areca mixed cropping system at CPCRI Regional Station, Vittal and it has been compared with the traditional method of eliminating the positional effect by blocking the experimental area. The experiment was laid out in randomized block design with 12 treatment combinations, 4 replications and 6 trees per plot. The

Table 2. Estimated parameters with standard errors of the field experiment

	Proposed Method		Method of blocking	
Treatments	μ+β	SE	μ + β	SE
1	7.57	1.15	6.18	1.33
2	11.80	1.14	11.47	1.33
3	7.26	1.14	6.38	1.33
4	8.82	1.15	7.94	1.33
5	11.97	1.12	11.91	1.33
6	9.45	1.13	8.73	1.33
7	13.79	1.12	14.45	1.33
8	12.12	1.13	12.16	1.33
9	11.38	1.15	11.84	1.33
10	14.08	1.14	14.84	1.33
11	14.10	1.14	15.38	1.33
12	14.97	1.14	16.03	1.33
MSE	32.02		42.22	

Note: μ+β is the sum of the estimated values of general mean and treatment effect after eliminating the positional/block effect

main objective of the experiment is to compare the effect of different treatments on the yield of cocoa. Four years cumulative yield data has been taken as the study variable. A total of 288 experimental cocoa trees were planted at a spacing of 2.7m × 5.4m. Estimated parameters (general mean + treatment effect) with standard errors and the mean squared errors (MSE) of cumulative yield data of cocoa after eliminating the positional/block effect through both the methods are given in Table 2. There is a significant reduction in the MSE of the proposed method than the traditional method for comparing the treatment effect. We have used MATLAB package to develop programmes for the simulation study and the data analysis.

6. CONCLUSION

We generally use block designs to eliminate positional effect in field experiments. In many situations, the underlying assumption of homogeneity with in the block may not be true. In the present study, a method is proposed to eliminate the positional effect nonparametrically and the only assumption about the positional effect is that it is a smooth spatial (bivariate) function. The method is also extended for the analysis of data in the presence of sudden shifts in the spatial function (positional effect). The proposed method is useful when there is no advance information about the field conditions to divide the experimental area into homogeneous blocks.

REFERENCES

Gilmour, A.R., Cullis, B.R. and Verbyla, A.P. (1997). Accounting for natural and extraneous variation in analysis of field experiments. *J. Ag. Biol. Environ. Stat.,* **2**, 269-273.

Hardle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press.

Hart, J.D. (1997). *Nonparametric Smoothing and Lack-of-fit Tests*. Springer Verlag, New York.

Hastie,T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Jose, C.T. and Ismail, B. (2001). Nonparametric inference on jump regression surface. *J. Nonparametric Stat.*, **13**, 791-813.

Muller, H.G. and Song, K. (1994). Maximin estimation of multidimensional boundaries. *J. Mult. Anal.*, **50**, 265-281.

Opsomer, J.D. and Ruppert, D. (1999). A root-n consistent estimator for semiparametric additive modeling. *J. Comput. Graph. Stat.*, **8**, 715-732.

Ruppert, D. and Wand, M.P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.*, **22**, 1346-1370.

Available online at www.isas.org.in
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(2) 2009 159-164

Modelling for Growth Pattern in Crossbred Cattle along with Autocorrelation

Surendra Singh[1], L.M. Bhar[1], A.K. Paul[1*], Satish Chand Sharma[2] and N. Sivaramane[1]
[1]*Indian Agricultural Statistics Research Institute, New Delhi*
[2]*J.V. College, Baraut, Bagpat*

(Received: October 2007, Revised: April 2009, Accepted: July 2009)

SUMMARY

Research on cattle growth is one of the important studies in the animal sciences. In the present study data on body weight were taken from birth to an age of 36 months for double cross Friesian × Sahiwal (F×S) and tripple cross Friesian × Sahiwal × Hariana (F×S×H) cattle. The data is found to contain heteroscedasticity of error variance and autocorrelation. The data in the study were unequal spaced which made impossible to use convention autocorrelation technique and hence a technique has been developed for finding out autocorrelation for unequal spaced data. Two nonlinear models, Logistic and Gompertz models are fitted to estimate growth rate and other parameters. Models are modified incorporating heterogeneity of error variance along with autocorrelation. Both the models were fitted under homoscedastic error structure and heteroscedastic error structure along with autocorrelation for comparison. It is found that parameters estimated under heteroscedastic error structure along with autocorrelation are better than the models fitted under homoscedastic error structure. Growth rate was found to be better for F×S×H breed than F×S breed. Maturity weight was found to be more for F×S breed than F×S×H breed. The results shows that Gompertz model outperformed Logistic model and correcting the models under homoscedastic error structure to heteroscedastic error structure has greatly improved the estimates.

Keywords: Homoscedastic, Heteroscedastic, Autocorrelation, Crossbred, Non-linear model, cattle growth, Logistic model, Gompertz model.

1. INTRODUCTION

Growth is a complex biological process that must be evaluated carefully if a profitable combination of growth and efficiency is to be realized. Knowledge relating to birth weight, mature weight, maturing rate and the point of inflection of the growth curve in various breeds and crosses is useful for cattle growth in various breeds so that producers can select breed combinations that will produce the most efficient growth pattern for their operations.

The growth pattern of cattle has been mainly studied under homoscedastic error structure, reported in the literature. But data of cattle growth generally violate assumption of homoscedasticity, i.e., error have common variance. Therefore, the purpose of this study is to compare the growth pattern under homoscedastic and heteroscedastic error structure along with autocorrelation.

Growth models are used to predict rates and change in the shape of the organism. They can be applied in determining the food requirements so as to get a desired growth. The estimated parameters of growth function can evaluate various growth characteristics of animal. Comparison of nonlinear models for weight age data in cattle has been done

[*] *Corresponding author* : A.K. Paul
E-mail address : amrit_66@rediffmail.com

under homoscedastic error structure ((Brown *et al.* (1972), Brown *et al*. (1976) and Kolluru *et al*. (2003)). A number of such nonlinear models are available, but comparisons of models are needed to find most appropriate one. Such comparisons were made among weight–age models for animals. Kolluru (2000) studied only Logistic model under heteroscedastic error condition for cattle growth. Therefore, there is a need to study other models also. Two models are taken for the present study; these are Logistic and Gompertz models. When heterogeneity of variance is evident, ordinary least square estimate of parameters may be inefficient as well as when weights are collected over time for each cattle, serial correlation is often present. In the present study Logistic and Gompertz models are fitted incorporating heteroscedasticity of variance along with autocorrelation.

2. MATERIAL AND METHODS

2.1 Data Description

Data used in the study were collected from history sheets of cattle from birth to 36 months of age from military dairy farms at Dehradun for Friesian × Sahiwal breed for 40 cattle. For Friesian × Sahiwal × Hariana breed data for 35 cattle were taken for comparing the growth pattern.

Let us consider the following model

$$y_j = f(\boldsymbol{X}_j, \beta) + e_j \qquad (2.1)$$

where y_j is the j^{th} observation, X_j is covariate vector, β is parameter vector, e_j is the error term and f is a non-linear function. In the present study two non-linear models have been considered viz. Logistic and Gompertz Models. The functional forms of these models are as follows

(i) Logistic model: $f(t, \beta) = \dfrac{\beta_1}{1+\beta_2 exp(-\beta_3 t)}$

(ii) Gompertz model: $f(t, \beta) = \beta_1 \exp(-\beta_2 e^{-\beta_3 t})$

Here in $f(\boldsymbol{X}_j, \beta)$, the covariate vector $\boldsymbol{X}_j$ is replaced by t, which is the only covariate in the model. Usually, it is assumed that (i) the errors e_j have zero means, (ii) the errors e_j are uncorrelated, (iii) the errors e_j have common variance and (iv) the errors e_j are normally distributed. In case of the animal growth data, many a times, the above assumption are violated; errors are generally correlated and do not have common variance. In the present study we, therefore, fitted these models considering non-constant and correlated error variances. However, for the sake of comparison, models are also fitted under homoscedastic error variance.

2.2 Models with Heteroscedastic Error Structure along with Autocorrelation

2.2.1 Heteroscedastic error structure

In the present study data revealed heteroscedasticity of error variance. Heteroscedasticity of variance is tested by Rank correlation test

$$R_e = 1 - \{6\Sigma d_i^2 / (n\,(n^2 - 1))\}$$

where, d_i = difference between the ranks of corresponding value of y_i and e_i. A high rank correlation suggests the presence of heteroscedasticity.

Let us consider the model as given in (2.1). As mentioned earlier assumption of constant intra-individual response variance is violated frequently for growth data. Generally, growth data often exhibit constant coefficient of variation rather than constant variance (Davidian and Giltinan 1995); that is, variance is proportional to the squares of the mean response. In this case, a more appropriate assumption is

$$E(y_j) = f(\boldsymbol{X}_j, \beta),\ V(y_j) = \sigma^2\,[f(\boldsymbol{X}_j, \beta)]^2 \qquad (2.2)$$

where, σ a scale parameter, is the coefficient of variation. Under such situation, where variance is nonconstant across the response range, it is assumed that the variances of y_j are known up to a constant of proportionality, (Davidian and Giltinan 1995), that is,

$$V(y_j) = \sigma^2/w_j \qquad (2.3)$$

for some constants w_j, $j = 1, \ldots, n$. This type of setting might arise in the case where the responses y_j are themselves averages of w_j uncorrelated replicate measurements, with all such measurements having common variance σ^2. Under this model, except for the multiplicative constant σ^2, variance is known up to the value of the regression parameter β, which appears through the mean response. An obvious approach is thus to take advantage of the functional form for a variance to construct estimated weights, replacing β by a suitable estimate, and to apply the weighted least squares idea.

The OLS estimator $\hat{\beta}_{OLS}$ is a natural choice to use for construction of estimated weights. An estimator for β that takes into account the assumed mean-variance

relationship may be obtained by forming estimated weights (Davidian and Giltinan 1995)

$$\hat{w}_j = \frac{1}{f^2(\boldsymbol{X}_j, \hat{\boldsymbol{\beta}}_{OLS})} \tag{2.4}$$

2.2.2 Auto-correlated structure

In the present study, the presence of autocorrelation in the data was checked by Durbin–Watson test. The Durbin–Watson statistic is given by

$$d = \sum_{u=2}^{n} (e_u - e_{u-1})^2 / \sum_{u=1}^{n} e_u^2$$

which ranges between 0 and 4. Value of d near 2 indicates no autocorrelation, a value towards 0 indicates positive autocorrelation, while the value towards 4 indicates negative autocorrelation. When there is evidence for the presence of autocorrelation, we find the value of auto correlation and model is fitted accordingly.

Correlation among observations on a given individual (cattle) is more likely to be present in this context (weight). In many cases, a systematic pattern of correlation is evident, which may be characterized accurately by a relatively simple model. To accommodate intra-individual correlation, a description of the assumed correlation pattern among the elements of e (error vector) is made. This assumption will be simple if the observations are taken at equal intervals. But the situation is quite complex when the unequally spaced observations are considered in the present study. Suppose that

$$\text{Corr } (e) = \Gamma(\alpha) \tag{2.5}$$

where the correlation matrix $\Gamma(\alpha)$ is a function of a vector of correlation parameters α. The choice of a suitable correlation matrix depends on the nature of the repeated measurement factor. As a special case where the repeated observations are taken over time, standard models for serial correlation patterns are available, i.e, the autoregressive (AR) model of order one. For definiteness, the observations are assumed to be indexed in the order in which they were collected. In the simplest case where the n repeated measurements are equally spaced in time, if the correlation between two adjacent observation is α, then the correlation between any two measurements j_1 and j_2 is given by

$$\text{Corr } \left(e_{j_1}, e_{j_2}\right) = \alpha^{|j_1 - j_2|} \tag{2.6}$$

The AR(1) correlation pattern may be generalized to accommodate situations where the observations are not equally spaced (see, Liang and Zeger (1986) and Chi and Reinsel (1989)) . If j_1 and j_2 are measurements taken at times t_{j_1} and t_{j_2} respectively where $j_1 \neq j_2$ then

$$\text{Corr } (e_{j_1}, e_{j_2}) = \alpha^{|t_{j_1} - t_{j_2}|} \tag{2.7}$$

This may be expressed by the correlation matrix as

$$\Gamma(\alpha) = \begin{bmatrix} 1 & \alpha^{(t_2 - t_1)} & \alpha^{(t_3 - t_1)} & . & . & . & \alpha^{(t_n - t_1)} \\ & 1 & \alpha^{(t_3 - t_2)} & . & . & . & \alpha^{(t_n - t_2)} \\ & & . & . & & & . \\ & & & . & . & & . \\ & & & & . & . & . \\ & & & & & . & \alpha^{(t_n - t_{n-1})} \\ & & & & & & 1 \end{bmatrix} \tag{2.8}$$

The method of estimation of α is described in the next sub-section.

2.2.3Auto-correlation for unequally spaced observations

When observations are taken on subjects at arbitrary time points, there must be an underlying continuous time process (Jones 1981, Diggle 1988, Chi and Reinsel 1989). For equally spaced observations, there may or may not be an underlying continuous time process. Unequally spaced observations differ from equally spaced observations with some missing observations in that there is no basic sampling interval. The mathematical model for a continuous time AR(1) process, denoted as CAR(1), (Jones and Boadi-Boateng 1991), is given by

$$\frac{\mathrm{d}}{\mathrm{dt}} \varepsilon(t) + \alpha \varepsilon(t) = G\eta(t) \tag{2.9}$$

where G is a constant, G^2 is variance per unit time, $\eta(t)$ is continuous time 'white noise', $\Sigma(t)$ is error at t^{th} time and α is the correlation between two adjacent

observations. A model for $\eta(t)$ is a differential equation given by

$$\eta(t) = \frac{d}{dt}\omega(t) \tag{2.10}$$

where $\omega(t)$ is Brownian motion or a Wiener process.

Combining equations (2.9) and (2.10), we get

$$d\varepsilon(t) + \alpha\varepsilon(t)dt = G d\omega(t) \tag{2.11}$$

This equation is solved by integration. A solution of (2.11) with the random input removed is given by (see, Jones 1993)

$$\frac{d}{dt}\varepsilon(t) + \alpha\varepsilon(t) = 0 \tag{2.12}$$

If (2.12) is integrated from time t_1 to time t_2, the solution is a prediction

$$\varepsilon(t_2) = \exp\{-\alpha(t_2 - t_1)\}\varepsilon(t_1) \tag{2.13}$$

The solution, (2.13) is now in the form of a discrete time AR(1) process with an auto regression coefficient α. The equation (2.13) can be generalized as

$$\varepsilon(t_n) = \exp\{-\alpha(t_n - t_{n-1})\}\varepsilon(t_{n-1}) \tag{2.14}$$

Fitting equation (2.14) by nonlinear modeling techniques, one can estimate α.

2.2.4 Computational aspects

Once the weights are calculated by (2.4) by using OLS estimator $\hat{\beta}_{OLS}$, they form the weight matrix for all observations. Let us denote this matrix as W, a diagonal matrix whose diagonal elements are the weights estimated through (2.4). Then the variance-covariance for y under heteroscedastic model is

$$\text{Cov}(y) = \sigma^2 W \tag{2.15}$$

For the model with heteroscedastic error variance along with auto-correlation, the variance-covariance matrix of y becomes

$$\text{Cov}(y) = \sigma^2 W^{1/2} \Gamma(a) W^{1/2} = \Sigma \text{ (say)} \tag{2.16}$$

where $\Gamma(\alpha)$ is as given in (2.8). After estimating α by the model (2.14), it is incorporated in (2.16).

Now applying Generalized Least Squares principle, the one-step ahead estimates of parameters can be obtained by minimizing

$$(y - f(X, \beta_{OLS}))' \Sigma^{-1} (y - f(X, \beta_{OLS})) \tag{2.17}$$

After getting new estimates of β, weights are again estimated and another set of parameters are estimated through (2.17). The process is continued till the values of β converges. Final value of estimates of β is represented as β_{GLS}.

2.3 Measure of Model Adequacy

The empirical comparison of models can be made using with goodness of fit statistics such as RMSE and RMAPE. Lower the values of RMSE and RMAPE better are the models. It is concluded that the model which has minimum RMSE and RMAPE gives better parameters of the fitted model. For calculating the RMSE and RMAPE following formulas are used.

Root mean squared error (RMSE)

$$= \left[\sum_{i=1}^{n} \frac{\left(Y_i - \hat{Y}_i\right)^2}{n - p}\right]^{1/2} \tag{2.18}$$

Relative mean absolute percentage error (RMAPE)

$$= \frac{1}{n}\sum_{i=1}^{n} \frac{|Y_i - Y|}{Y_i} \times 100 \tag{2.19}$$

where Y_i is original value, $\hat{Y}_i$ is predicted values or estimated value and n is the total number of observations, p is the number of parameters.

3. RESULTS AND DISCUSSION

Models are first fitted under homoscedastic error structure. For this purpose SAS package Version 9.1 has been used. In the present study the data revealed heteroscedasticity of error variance as rank correlation is found to be 0.6920 for F×S×H breed and 0.6210 for F×S breed. In the present study data also have auto-correlation, when checked using Durbin–Watson test. The Durbin–Watson statistic (d) is found to be 0.8120

Table 1. Auto-correlation of different breed by different model

Name of Breed	Model	Auto-correlation
F×S	Logistic	0.1254
F×S	Gompertz	0.1330
F×S×H	Logistic	0.4448
F×S×H	Gompertz	0.4319

Table 2. Parameter estimates of models under homoscedastic error variance and heteroscedastic error variance along with auto-correlation of Friesian × Sahiwal breed at Dehradun station

Parameters	Under homoscedastic error variance		Under heteroscedastic error variance along with autocorrelation	
	Logistic model	Gompertz model	Logistic model	Gompertz model
β_1	354.4000 (15.6193)	382.5000 (15.7104)	299.9152 (3.3298)	328.5822 (1.7363)
β_2	6.6555 (0.8399)	2.3190 (0.0966)	9.4134 (0.1504)	2.5333 (0.0064)
β_3	0.1534 (0.0142)	0.0927 (0.0072)	0.2643 (0.0035)	0.1377 (0.0009)
Goodness of fit statistics				
RMSE	16.7320	11.3109	0.1210	0.0908
RMAPE	21.4587	13.4961	12.9276	8.1148
Autocorrelation	-	-	0.1254	0.1330

Note: Figures in the brackets indicate standard errors.

for F×S×H breed and 0.8394 for F×S breed. α was estimated by fitting equation (2.14) by nonlinear modeling technique. The value of α (auto-correlation) is obtained through NLIN option of SAS procedure. The estimated values of autocorrelation for different breeds are given in Table 1.

For fitting models under heteroscedastic error structure along with autocorrelation, program is written in SAS/IML language.

It can be observed from Table 2 that for F×S breed at Dehradun farm RMSE (11.3109) is less for Gompertz model than RMSE (16.7320) by Logistic model and similarly RMAPE(13.4961) is Less for Gompertz model than RMAPE(21.4587) by Logistic model, this shows that results of Gompertz model are better than logistic model under homoscedastic error condition. The data of the breed are having heteroscedasticity of variance and autocorrelation which was tested by Durbin Watson test. Auto correlation for this breed is found to be 0.1254 and 0.1330 of Logistic and Gompertz models respectively.When the results under homoscedastic error structure and heteroscedastic error structure along with autocorrelation are compared it is found that RMSE and RMAPE under heteroscedastic error structure with auto correlation are found less than homoscedastic error structure for both models, this shows that results for heteroscedastic error structure are better than homoscedastic error structure.

From Table 3 it is observed that for F×S×H breed RMSE (23.9571) by Gompertz model is less than RMSE (24.3640) by logistic model under homoscedastic error structure and RMAPE (7.1689) by Gompertz model is less than RMAPE (13.3725) by Logistic model, so results of Gompertz model are better than Logistic model under homoscedastic error structure. RMSE (0.2942) and RMAPE (7.2071) of Logistic model and RMSE (0.2975) and RMAPE (4.3462) of Gompertz model under heteroscedastic error structure along with auto correlation are less than RMSE and RMAPE under homoscedastic error structure. This shows that results under heteroscedastic error structure with autocorrelations are better than homoscedastic error structure. Mature weight is found to be more for F × S breed than mature weight of F × S × H breed under homoscedastic error structure where as mature weight found more for F × S × H breed than F × S breed under heteroscedastic error structure along with autocorrelation. Growth rate was found to be better for F × S × H than F × S breed.

Table 3. Parameter estimates of different models under homoscedastic error variance and heteroscedastic error variance along with auto-correlation of Friesian × Sahiwal × Hariana breed at Dehradun station

Parameters	Under homoscedastic error variance		Under heteroscedastic error variance along with autocorrelation	
	Logistic model	Gompertz model	Logistic model	Gompertz model
β_1	327.8000 (17.2990)	349.2000 (23.9764)	362.7186 (8.6022)	475.8925 (5.2318)
β_2	8.5121 (1.9712)	2.5636 (0.2814)	13.3927 (1.4666)	2.9731 (0.1243)
β_3	0.1949 (0.0262)	0.1156 (0.0176)	0.2429 (0.0179)	0.0941 (0.0100)
Goodness of fit statistics				
RMSE	24.3640	23.9571	0.2942	0.2975
RMAPE	13.3725	7.1689	7.2071	4.3462
Autocorrelation	-	-	0.4448	0.4319

Note: Figures in the brackets indicate standard errors.

REFERENCES

Brown, J.E., Brown C.J. and Butts W.T. (1972). A discussion of the aspects of weight, mature weight and rate of maturing in hereford and angus cattle. *J. Anim. Sci*., **34**, 525.

Brown, J.E., Fitzhugh, H.A. and Cartwright, T.C. (1976). A comparison of nonlinear models for describing weight-age relationships in cattle. *J. Anim. Sci.*, **43**, 810-818.

Chi, E.M. and Reinsel, G.C. (1989). Models for longitudinal data with random effects and AR(1) errors. *J. Amer. Statist. Assoc.,* **84**, 452-459.

Davidian, D. and Giltinan, M. (1995). *Non Linear Models for Repeated Measurement Data*. Chapman Hall, London.

Diggle , P.J. (1988). An approach to the analysis of repeated measurement data. *Biometrics,* **45**, 1255-1258.

Jones, R.H. and Boadi-Boateng, F. (1991). Unequally spaced longitudinal data with AR(1) serial correlation. *Biometrics,* **47**, 161-175.

Jones, R.H. (1993). *Longitudinal Data with Serial Correlation: A State-Space Approach.* Chapman and Hall, London.

Kolluru, R. (2000). On some aspect of growth patterns of crossbred cattle. Unpublished M.Sc. Thesis, Indian Agricultural Research Institute, New Delhi.

Kolluru, R., Rana, P.S. and Paul, A.K. (2003). Modelling for growth pattern in crossbred cattle. *J. Anim. Sci.,* **73(10),** 1174-1179.

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika,* **73,** 13-22.

SAS (1990). SAS Users' Guide Version 6. SAS Institute Incorporation, U.S.A.

Available online at www.isas.org.in
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(2) 2009 165-174

Generation of Linear Trend-Free Designs for Factorial Experiments

Susheel Kumar Sarkar, Krishan Lal*, Rajender Parsad and V.K. Gupta
Indian Agricultural Statistics Research Institute, New Delhi

(Received: February 2008, Revised: May 2009, Accepted: July 2009)

SUMMARY

The present article deals with generation of designs for 2^k factorial experiments (without and with confounding) that permit the estimation of main effects free from linear trend effects present in the experimental units. The proposed method exploits the use of component wise product of vectors to generate linear trend-free for main effects designs for two-level factorials. The procedure of identifying two- and three- factor interactions that are linear trend-free has also been incorporated in the method. The method has also been extended to generate designs for confounded factorial experiments with any number of factors k (≥ 3) that are linear trend-free for main effects and identify two- and three- factor interactions for which the design is linear/nearly linear trend-free.

Keywords: Factorial experiments, Linear trend-free designs, Orthogonal polynomials, Run orders, Systematic designs.

1. INTRODUCTION

Designs for factorial experiments are very popular among the agricultural, biological and industrial experimenters. Designs for factorial experiments are used for identifying the important main effects and interactions. However, the experimental units in these designs may exhibit a trend over space or time. Such situations occur in agricultural experiments when there is a slope, in the field and there is sequential application of treatments to the same experimental unit over different time periods. Trends may also occur in the experimental units when the land is irrigated and the nutrients are supplied by the fertilizers but because of the slope, the distribution of nutrients is not uniform. In such experimental situations, a common polynomial trend within experimental units is likely to occur. The trend may be represented by a polynomial of appropriate degree smaller than the block size. The presence of trends among the experimental units within a block affects the inference problem on main effects and interactions of interest. In the presence of trends among the experimental units it may be desirable to estimate the main effects and interactions of interest free from trend effects. Generally, we consider the presence of a linear trends among the experimental units within a block. In the presence of linear trends among the experimental units within a block of a factorial experiment, it is desired to allocate the treatment combinations to experimental units in such a manner that the main effects and interactions of interest are estimated free from the linear trend effects. Such designs are called as *linear trend-free designs for factorial experiments* for estimating the effects of interest and the ordered application of treatments to experimental units is called *run order*.

Bradley and Yeh (1980) gave a rigorous treatment to the theory of trend-free block designs. Yeh and Bradley (1983) discussed the existence of trend-free block designs for specified trend polynomials under a homoscedastic model. For further research on trend-free

*Corresponding author : Krishan Lal
E-mail address : klkalra@iasri.res.in

block designs, one may refer to Dhall (1986), Bradley and Odeh (1988), Stufken (1988), Chai and Majumdar (1993), Jacroux *et al.* (1995, 1997), Majumdar and Martin (2002), Lal *et al.* (2005, 2007), among others. For factorial set up of treatments, Daniel and Wilcoxon (1966) developed plans for sequencing the treatment combinations of a two-level full or fractional factorial to experimental units. Cheng and Jacroux (1988) further introduced trend-free run orders of two-level designs using group theory. John (1990) introduced a fold over method and discussed the trend-free properties of systematic run orders based on their method. Coster (1993) extended the work of Coster and Cheng (1988) to mixed level factorials. For the construction of trend-free fractional factorial and response surface designs using computer, readers may visit the website of Nguyen (http://designcomputing.net/gendex/rat/). The approach used to generate trend-free designs is to minimize a proper objective function using the random seed.

The present article attempts to generate linear trend-free designs for full factorials both without and with confounding. The search is restricted to two-level factorials. The designs are generated by making use of a computer-aided search because this approach is relatively easy and fast as compared to generating designs through an algebraic treatment. The search is restricted to linear main effects trend-free designs but the search also identifies 2- and 3- factor interactions that are linear trend-free/nearly linear trend-free in these designs. The method is illustrated with suitable examples. Section 2 deals with general description of linear trend-free designs for two-level factorial experiments. Section 3 develops the method for generation of linear trend-free 2^k full factorial experiments in which main effects are estimated free from linear trend effects and 2- and 3- factor interactions are identified that are estimable free from linear trend effects. The method is illustrated in Section 4. Section 5 modifies the method to generate confounded factorial experiments that are linear trend-free for main effects and identifies 2- and 3-factor interactions that are also estimable free from linear trend effects. The working of the algorithm for generating linear trend-free confounded factorial designs has been described in Section 6. Catalogues of linear trend-free 2^k factorial experiments for (k = 3, …, 7) are available with the authors.

2. LINEAR TREND-FREE DESIGNS FOR TWO LEVEL FACTORIAL EXPERIMENTS

In factorial experiments, generally the interest of the experimenter is to estimate the lower order factorial effects precisely. In general, the treatment combinations are applied to experimental units randomly. If it is known or assumed that experimental units exhibit a linear trend over space or time, then it is advantageous to choose a systematic run order so that the run order is linear trend-free for main effects and is linear/nearly linear trend-free for lower order interactions. For a design of 2^k factorial experiment the total number of treatment combinations is $n = 2^k$ and the model for factorial experiment with linear trend effect conducted in a single replication is

$$y = \boldsymbol{U}\beta + \boldsymbol{T}\theta + \boldsymbol{e} \qquad (1)$$
$$E(\boldsymbol{e}) = \boldsymbol{0},\ D(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$$

where $\boldsymbol{y}$ is a $n \times 1$ vector of observations, $\boldsymbol{U}$ is an $n \times n$ design matrix and β is an n-component vector of general mean and treatment effects (treatment combinations written in lexicographic order from the left), $\boldsymbol{T}$ is the vector of coefficients of the first degree orthogonal polynomial of order n, θ is a regression coefficient and $\boldsymbol{e}$ is $n \times 1$ vector of independently and identically normally distributed errors. In a 2^k factorial experiment, the total number of treatment combinations is even. So vector $\boldsymbol{T}$ is

$$\boldsymbol{T} = (-(n-1), -(n-3), ..., -3, -1, 1, 3, (n-3), (n-1)) \qquad (2)$$

We may redefine

$$\mathbf{U} = \left[\mathbf{1} \vdots \boldsymbol{X}_1 \vdots \boldsymbol{X}_2 \vdots \cdots \vdots \boldsymbol{X}_s \vdots \cdots \vdots \boldsymbol{X}_k\right],$$

where $\mathbf{1}$ is an n-component vector of ones, $\mathbf{X}_s$ is an $n \times \binom{k}{s}$ matrix of coefficients of the s-factor factorial effects, s = 1, 2, ..., k. Based on general definition given by Bradley and Yeh (1980), all the s-factor factorial effects are linear trend-free if $\boldsymbol{X}_s'\boldsymbol{T} = \mathbf{0}$. Here $\mathbf{0}_t$ denotes a t-component vector of all zeros. For example, a design for factorial experiment would be linear trend free for main effects if $\boldsymbol{X}_1'\boldsymbol{T} = \mathbf{0}$.

It may not be possible always to generate a design for factorial experiment that is linear trend-free for all

the effects of interest. This provides a motivation to go for nearly linear trend-free designs for some of the effects of interest. The condition given by Chai (1995) for a balanced incomplete block design to be nearly linear trend-free block design can be generalized for the factorial experiments. In a factorial experiment, the condition for the $n \times 1$ column vector of coefficients of contrast of interest, say $\mathbf{A}$, to be nearly linear trend-free is

$$\mathbf{0} < \boldsymbol{A}'\boldsymbol{T} \leq n \tag{3}$$

Our interest is to obtain designs for factorial experiments in which contrasts for the main effects are estimated free from linear trend effects and to identify/search two and three factor interactions that are estimable free from linear trend effect. To obtain such designs, we begin with following lemmas:

Lemma 2.1 {Cheng and Jacroux (1988)}. In any run order of a complete 2^k design, the number of mutually orthogonal linear trend-free factorial contrasts (main effects or interaction) are at most $2^k - k - 1$.

Let $s_1, s_2, \ldots, s_k$ denote the coefficients of the estimates of the main effect contrasts, defined in Section 3 as I3.

Lemma 2.2 {Cheng and Jacroux (1988)}. For $p = 1, \ldots, k$, let s_p denote the vector of coefficients of the estimated contrast of main effect of factor p and $\circ$ denote the component-wise product of vectors. For example, for $s_1' = (-1, 1, -1, 1, -1, 1, -1, 1)$,

$s_2' = (-1, -1, 1, 1, -1, -1, 1, 1)$ and

$s_3' = (-1, -1, -1, -1, 1, 1, 1, 1)$,

$s_1 \circ s_2 \circ s_3 = (-1, 1, 1, -1, 1, -1, -1, 1)$.

Then the component wise product

$$\prod_{p=1}^{t+1} \circ s_{j_p} = s_{j_1} \circ s_{j_2} \circ \cdots \circ s_{j_{t+1}}$$

is orthogonal to $\boldsymbol{T}_0, \boldsymbol{T}_1, \ldots, \boldsymbol{T}_t$, where $\boldsymbol{T}_i$ is the vector of coefficients of i^{th} order orthogonal polynomial for number of design runs n, $i = 0, 1, \ldots, t$.

Using these two lemmas, the algorithm to develop the factorial experiments that are linear trend-free for main effects is given in Section 3 and working of the algorithm is given in Section 4.

3. GENERATION OF TREND-FREE DESIGNS FOR FACTORIAL EXPERIMENTS

We shall henceforth denote by AL1 the algorithm that generates designs for complete factorial experiment in single replication that are linear trend-free for estimation of main effects and also identifies the 2- and 3- factor interactions that are estimable free from linear trend effect. AL1 is described in the sequel.

I. The Method AL1

I1 Let the number of factors be k; number of treatment combinations be $n = 2^k$, n is even.

I2 Generate an $n \times k$ array with $n/2$ symbols as +1 and $n/2$ symbols as −1 in each column. The generation of columns is explained in the following steps.

I3 The j^{th} column of the array contains 2^{j-1} replications of the symbols −1 and +1 alternatively; $j = 1, 2, \ldots, k$. Thus, the first column contains symbols −1 and +1 alternatively and the k^{th} column contains $n/2$ times symbol −1 and $n/2$ times symbol +1.

I4 In the generated $n \times k$ array, the k columns correspond to the coefficients of the contrasts corresponding to the k main effects.

I5 The n rows of $n \times k$ matrix represent n treatment combinations in lexicographic order.

Now we describe the steps to convert the full factorial generated in I3 into a linear trend-free for main effects design by using the steps given below. Let $s_1, s_{2, \ldots,} s_k$ denote the k columns of the array with each column denoting the coefficients of contrasts for main effects. We shall denote by $\prod_{j=1}^{k} \circ s_j$ the component wise product of symbols. For example, if $k = 2$, then $s_1' = (-1\ +1\ -1\ +1)$, $s_2' = (-1\ -1\ +1\ +1)$ and $(s_1 \circ s_2)' = (+1\ -1\ -1\ +1)$.

I6 Case I: k is odd.

For $i = 1, 2, \ldots, k-1$, define $A_i = \prod_{j(\neq i)=1}^{k} \circ s_j$, and $A_k = \prod_{j=1}^{k} \circ s_j$. Then

$$\boldsymbol{X} = [A_1 \quad A_2 \quad \cdots \quad A_i \quad \cdots \quad A_k]$$

is the required linear trend-free design for main effects for a 2^k factorial experiment.

I7 Case II: k is even.

For $i = 1, 2, ..., k$, define $A_i = \prod_{j(\neq i)=1}^{k} \circ s_j$. Then

$$\boldsymbol{X} = [A_1 \quad A_2 \quad \cdots \quad A_i \quad \cdots \quad A_k]$$

is the required linear trend-free design for main effects for a 2^k factorial experiment.

We further describe steps to identify linear trend-free two-factor and three-factor interactions in the design generated in I6 or I7.

I8 From the linear trend-free for main effects design generated in I6 or I7 generate a new design matrix $n \times \left[k + \binom{k}{2} + \binom{k}{3} \right]$ given by $\mathbf{Z} = [\boldsymbol{X} \quad \boldsymbol{X}^{(2)} \quad X^{(3)}]$. Here $\boldsymbol{X}^{(u)}$, $u = 2, 3$ contains columns corresponding to the coefficients of the contrasts of all the $\binom{k}{u}$ u-factors interactions obtained from X.

I9 For $u = 2, 3$, identify the columns in $\boldsymbol{X}^{(u)}$ that are linear trend-free. Then the corresponding u-factor interactions are linear trend-free. Further, identify the columns from the remaining columns in $\boldsymbol{X}^{(u)}$ that satisfy the condition in (3). Then the corresponding u-factor interactions are nearly linear trend-free. The remaining columns are not trend-free.

Remark 3.1: The algorithm AL1 is infact a combination of steps provided by Chen and Jacroux (1988) as stated above in Lemma 2.2 and clarification regarding odd and even number of factors given by Hinkelman and Jo (1998).

4. WORKING OF AL1

Consider the problem of constructing a linear trend-free design for main effects for a 2^4 factorial experiment with four factors as A, B, C and D. Using the method AL1, first obtain an 16×4 array using step I3. The four columns of the array correspond to

Table 4.1(a)

s_1	s_2	s_3	s_4	Treatment combinations
–1	–1	–1	–1	(1)
1	–1	–1	–1	*a*
–1	1	–1	–1	*b*
1	1	–1	–1	*ab*
–1	–1	1	–1	*c*
1	–1	1	–1	*ac*
–1	1	1	–1	*bc*
1	1	1	–1	*abc*
–1	–1	–1	1	*d*
1	–1	–1	1	*ad*
–1	1	–1	1	*bd*
1	1	–1	1	*abd*
–1	–1	1	1	*cd*
1	–1	1	1	*acd*
–1	1	1	1	*bcd*
1	1	1	1	*abcd*

Table 4.1(b)

A	*B*	*C* Design *X*	*D*	Treatment combinations
–1	–1	–1	–1	(1)
–1	1	1	1	*bcd*
1	–1	1	1	*acd*
1	1	–1	–1	*ab*
1	1	–1	1	*abd*
1	–1	1	–1	*ac*
–1	1	1	–1	*bc*
–1	–1	–1	1	*d*
1	1	1	–1	*abc*
1	–1	–1	1	*ad*
–1	1	–1	1	*bd*
–1	–1	1	–1	*c*
–1	–1	1	1	*cd*
–1	1	–1	–1	*b*
1	–1	–1	–1	*a*
1	1	1	1	*abcd*

the coefficients of the contrasts of the four main effects. The full description of the generated array is given below as Table 4.1(a). We denote the contrasts of main effects as s_1, s_2, s_3, and s_4.

In this example k (=4) is even. So we use Step I7 to generate linear trend-free design for main effects. This design is given in Table 4.1(b) as Design $\boldsymbol{X}$.

In Design $\boldsymbol{X}$, all the main effects are linear trend-free; some other factorial effects, however, may also be linear trend-free (Lemma 2.1). Since the interest of the experimenter is to estimate lower order interactions more precisely, we search for two and three factor interactions that are linear trend-free by using steps I8 and I9 of AL1. Using step I8, we first generate design matrix $\boldsymbol{Z}$ for main effects, two-factor and three factor interactions. The design $\boldsymbol{Z}$ is given in Table 4.2.

Using now Step I9 we identify two-factor and three-factor interactions that are linear trend-free or are nearly linear trend-free. It is found that all the six two-factor interactions *AB*, *AC*, *AD*, *BC*, *BD* and *CD* are linear trend-free. On the other hand, only one three factor interaction, *BCD*, satisfies condition (3) and is, therefore, nearly linear trend-free. Other three-factor interactions are neither linear nor nearly linear trend-free.

Table 4.3 gives an example of a design $\boldsymbol{Z}$ for 2^5 factorial which is linear trend-free for all main effects, all two-factor and three-factor interactions.

Table 4.2. Design Z for 2^4 factorial

$\boldsymbol{Z}$													
X				$X^{(2)}$						$X^{(3)}$			
A	*B*	*C*	*D*	*AB*	*AC*	*AD*	*BC*	*BD*	*CD*	*ABC*	*ABD*	*ACD*	*BCD*
–1	–1	–1	–1	1	1	1	1	1	1	–1	–1	–1	–1
–1	1	1	1	–1	–1	–1	1	1	1	–1	–1	–1	1
1	–1	1	1	–1	1	1	–1	–1	1	–1	–1	1	–1
1	1	–1	–1	1	–1	–1	–1	–1	1	–1	–1	1	1
1	1	–1	1	1	–1	1	–1	1	–1	–1	1	–1	–1
1	–1	1	–1	–1	1	–1	–1	1	–1	–1	1	–1	1
–1	1	1	–1	–1	–1	1	1	–1	–1	–1	1	1	–1
–1	–1	–1	1	1	1	–1	1	–1	–1	–1	1	1	1
1	1	1	–1	1	1	–1	1	–1	–1	1	–1	–1	–1
1	–1	–1	1	–1	–1	1	1	–1	–1	1	–1	–1	1
–1	1	–1	1	–1	1	–1	–1	1	–1	1	–1	1	–1
–1	–1	1	–1	1	–1	1	–1	1	–1	1	–1	1	1
–1	–1	1	1	1	–1	–1	–1	–1	1	1	1	–1	–1
–1	1	–1	–1	–1	1	1	–1	–1	1	1	1	–1	1
1	–1	–1	–1	–1	–1	–1	1	1	1	1	1	1	–1
1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 4.3 Linear trend-free for main effects, 2-factors and 3-factors interactions design for Complete Factorial Experiment

	A	B	C	D	E
abcd	1	1	1	1	–1
ae	1	–1	–1	–1	1
be	–1	1	–1	–1	1
cd	–1	–1	1	1	–1
ce	–1	–1	1	–1	1
bd	–1	1	–1	1	–1
ad	1	–1	–1	1	–1
abce	1	1	1	–1	1
de	–1	–1	–1	1	1
bc	–1	1	1	–1	–1
ac	1	–1	1	–1	–1
abde	1	1	–1	1	1
ab	1	1	–1	–1	–1
acde	1	–1	1	1	1
bcde	–1	1	1	1	1
(1)	–1	–1	–1	–1	–1
e	–1	–1	–1	–1	1
bcd	–1	1	1	1	–1
acd	1	–1	1	1	–1
abe	1	1	–1	–1	1
abd	1	1	–1	1	–1
ace	1	–1	1	–1	1
bce	–1	1	1	–1	1
d	–1	–1	–1	1	–1
abc	1	1	1	–1	–1
ade	1	–1	–1	1	1
bde	–1	1	–1	1	1
c	–1	–1	1	–1	–1
cde	–1	–1	1	1	1
b	–1	1	–1	–1	–1
a	1	–1	–1	–1	–1
abcde	1	1	1	1	1

Remark 4.1: The method AL1 described in Section 3 can be operated in a very simple manner to generate the same design as generated by AL1. The two cases (I) k is even and (II) k is odd are dealt separately. First we describe the method for case (I).

Case I. $n = 2^k$; k is even

Step-I Generate full factorial design for $n = 2^k$ factorial experiment in standard (lexicographic) order using the steps I1 to I3 of AL1.

Step-II Retain the even numbered letters treatment combinations as such and in the same position wherever these occur.

Step-III Replace the odd numbered letters treatment combinations by the complement letters. In other words produce another treatment combination which contains those letters not present in the original treatment combination. Retain it in the same position as that of the original one. The new design generated is the same as the one produced by AL1. For application of this simplified algorithm, see Design 1 for 2^4 factorial experiments.

Step IV Steps 1 – 3 take care of steps I1 – I5 and I7 of AL1.

Case II. $n = 2^k$; k is odd

Step-I Generate full factorial design for $n = 2^{k-1}$ factorial experiment in standard (lexicographic) order using the steps I1 to I3 of AL1.

Step-II Replace the even numbered letters treatment combinations of 2^{k-1} factorial by the complement letters. In other words produce another treatment combination which contains those letters not present in the original treatment combination. Retain it in the same position as that of the original one.

Step-III For the odd numbered letters treatment combinations augment the k^{th} letter with each odd numbered letters treatment combinations in 2^{k-1} factorial experiment. Retain these in the same position as that of the original one.

Step-IV In this way first half of the treatments combinations (2^{k-1}) of 2^k factorial are obtained. The second half of the treatment combinations is obtained by writing the

complement letters of first half treatments combinations. Thus this is the desired design.

Step-V Steps I – IV take care of steps I1 – I6 of AL1.

Remark 4.2: The designs obtained above are the component wise product of the contrasts of main effects. These designs are same as generated by AL1. Thus by Lemma 2.2 these designs are not only linear trend-free for main effects but are also linear trend-free for higher orders interactions. Lemma 2.1 will be applicable on these designs.

In the next section we give an improved version of the method AL1 to generate designs for confounded two-level factorial experiments that are linear trend-free for all the main effects and also for some of the 2- and 3-factors interactions.

5. METHOD AL2 FOR CONFOUNDED FACTORIAL EXPERIMENTS

In this section we consider the problem of obtaining confounded two-level factorial designs in the presence of linear trends within blocks. The designs obtained are trend-free for main effects and some 2-factor and 3-factor interactions. Consider a $(2^k, 2^{k-p})$ factorial experiment run in $b = 2^p$ blocks of size $m = 2^{k-p}$. Suppose that p independent factorial effects are confounded in each replication. We assume the following linear additive model:

$$\boldsymbol{y} = \boldsymbol{U}\beta + \boldsymbol{B}\gamma + \boldsymbol{T}\theta + \boldsymbol{e} \qquad (4)$$

$$E(\boldsymbol{e}) = 0,\ D(\boldsymbol{e}) = \sigma^2\boldsymbol{I},$$

where $\boldsymbol{B}$ is the $n \times b$ design matrix of observations versus blocks and γ is a b-component vector of block parameters; the other symbols are same as defined in model (1). $\boldsymbol{T} = \mathbf{1}_b \otimes t$ and t is the $(m \times 1)$ linear trend vector for blocks of size m and b is the number of blocks. The condition for main effects to be linear trend-free is $\boldsymbol{X}_1' \boldsymbol{T} = \mathbf{0}$. For $s = 1, 2, ..., k$, the matrix $\boldsymbol{X}_s$ is partitioned as $\boldsymbol{X}_s = [\boldsymbol{X}_{s1}' \quad \boldsymbol{X}_{s2}' \cdots \boldsymbol{X}_{sb}']'$. To obtain the desired design that is linear trend-free for main effects and for some of the 2- and 3- factor interactions, we have the following method AL2:

I. The method AL2

L1 Generate full factorial design for $n = 2^k$ factorial experiment in standard (lexicographic) order using the steps I1 to I3 of AL1. Replace the symbols –1 and +1 in each column by 0 and 1, respectively.

L2 Fix the p factorial effects to be confounded.

L3 Solve the following 2^p equations for the p chosen contrasts (factorial effects) to be confounded

$$\boldsymbol{Ax} = \mathbf{0} \text{ and } \boldsymbol{Ax} = \mathbf{1}$$

where $\boldsymbol{x}' = (x_1, x_2, ..., x_k)$ denotes the k factors in the experiment and $\boldsymbol{A} = (a_{ij})$ is a $p \times k$ matrix of known coefficients a_{ij}'s, $i = 1, 2, ..., p$; $j = 1, 2, ..., k$. $a_{ij} = 1$ or 0 depending upon whether the j^{th} factor is present or absent in the ith factorial effect confounded. 0 is a p component vector of zeros and 1 is a p component vector of all ones.

L4 Step L3 generates $b = 2^p$ blocks of size 2^{k-p} each.

We now describe steps to convert the confounded design generated in step L2 into a linear trend-free design for main effects.

L5 The treatment combinations within the $b = 2^p$ blocks maintain the same order of sequence as in the lexicographic order of complete factorial experiment in step L1 with symbols as 0 and 1 instead of –1 and +1. Again replace the symbols 0 and 1 by –1 and +1, respectively.

L6 Let $s_1^l,\ s_2^l, ...,\ s_k^l$ denote the k columns of the l^{th} block generated in step L3, $l = 1, 2, ..., b$. Perform steps I6 and I7 on each of the b blocks separately. Let $\boldsymbol{X}_{(l)}$ denote the matrix of coefficients generated for the l^{th} block by using this step.

L7 Then $\boldsymbol{X} = [\boldsymbol{X}_{(1)}' \quad \boldsymbol{X}_{(2)}' \cdots \boldsymbol{X}_{(b)}']'$ is the required linear trend-free design for main effects.

We further describe steps to identify linear trend-free two-factor and three-factor interactions in the design generated in L6 and L7.

L8 From the linear trend-free for main effects design generated in L6 and L7 generate a new design $n \times \left[k + \binom{k}{2} + \binom{k}{3}\right]$ given by $Z = \left[X\ X^{(2)}\ X^{(3)}\right]$.

Here $\boldsymbol{X}^{(u)}$, $u = 2, 3$ contains columns corresponding to the coefficients of the contrasts of all the $\binom{k}{u}$ u-factors interactions obtained from $\boldsymbol{X}$.

L9 For $u = 2, 3$, identify the columns in $\boldsymbol{X}^{(u)}$ that are linear trend-free. Then the corresponding u-factor interactions are linear trend-free. Further, identify the columns in $\boldsymbol{X}^{(u)}$ that satisfy the condition in (3). Then the corresponding u-factor interactions are nearly linear trend-free. The remaining columns are not trend-free.

6. WORKING OF AL2

Consider again the problem of constructing a linear trend-free for all main effects design for a 2^4 confounded factorial experiment obtained by confounding the highest order interaction *ABCD*. Using step I3 of algorithm AL1, we can obtain an 16×4 array as in Table 4.1(a). Then use of step L3 of AL2 requires solving the following two equations

$$s_1 + s_2 + s_3 + s_4 = 0 \bmod(2)$$

$$s_1 + s_2 + s_3 + s_4 = 1 \bmod(2)$$

The two blocks obtained are given in Design 6.1.

Using steps L4, L5, L6 and L7 of AL2 gives the two blocks of a linear trend-free for main effects design for 2^4 factorial experiment in which *ABCD* is confounded. The design is given in Design 6.2.

Design 6.1

		Block - 1		Treatment
A	*B*	*C*	*D*	combinations
−1	−1	−1	−1	(1)
1	1	−1	−1	*ab*
1	−1	1	−1	*ac*
−1	1	1	−1	*bc*
1	−1	−1	1	*ad*
−1	1	−1	1	*bd*
−1	−1	1	1	*cd*
1	1	1	1	*abcd*

		Block - 2		Treatment
A	*B*	*C*	*D*	combinations
1	−1	−1	−1	*a*
−1	1	−1	−1	*b*
−1	−1	1	−1	*c*
1	1	1	−1	*abc*
−1	−1	−1	1	*d*
1	1	−1	1	*abd*
1	−1	1	1	*acd*
−1	1	1	1	*bcd*

Design 6.2

		Block - 1		Treatment
A	*B*	*C*	*D*	combinations
		$\boldsymbol{X}_{(1)}$		
−1	−1	−1	−1	(1)
1	1	−1	−1	*ab*
1	−1	1	−1	*ac*
−1	1	1	−1	*bc*
1	−1	−1	1	*ad*
−1	1	−1	1	*bd*
−1	−1	1	1	*cd*
1	1	1	1	*abcd*

		Block - 2		Treatment
A	*B*	*C*	*D*	combinations
		$\boldsymbol{X}_{(2)}$		
−1	1	1	1	*bcd*
1	−1	1	1	*acd*
1	1	−1	1	*abd*
−1	−1	−1	1	*d*
1	1	1	−1	*abc*
−1	−1	1	−1	*c*
−1	1	−1	−1	*b*
1	−1	−1	−1	*a*

and $\boldsymbol{X}' = \left[\boldsymbol{X}'_{(1)} \quad \boldsymbol{X}'_{(2)}\right]$.

We now identify two- and three- factor interactions that are linear trend-free in Design 6.2. Using steps L7, L8 and L9 of AL2 gives that all the two factor interactions i.e. *AB*, *AC*, *AD*, *BC*, *BD* and *CD* are linear trend-free. Similarly, among three factor interactions, *BCD* is linear trend-free, *ACD* is nearly linear trend-free and the other two *ABC* and *ABD* are neither linear trend-free nor nearly linear trend-free.

Using AL1 and AL2, one can obtain designs of complete factorial and confounded factorial experiments for any number of factors k (≥ 3) that are linear trend-free for main effects and can identify two- and three- factor interactions that are linear/nearly linear trend-free.

We give below another example of 2^5 confounded design in which some of the two factor interactions are nearly trend-free.

Example 6.1: 2^5 Linear trend-free design for Confounded Factorial Experiment (*ABCDE* confounded)

Block - 1				
A	*B*	*C*	*D*	*E*
–1	–1	–1	–1	–1
1	1	–1	–1	–1
1	–1	1	–1	–1
–1	1	1	–1	–1
1	–1	–1	1	–1
–1	1	–1	1	–1
–1	–1	1	1	–1
1	1	1	1	–1
1	–1	–1	–1	1
–1	1	–1	–1	1
–1	–1	1	–1	1
1	1	1	–1	1
–1	–1	–1	1	1
1	1	–1	1	1
1	–1	1	1	1
–1	1	1	1	1

Block - 2				
A	*B*	*C*	*D*	*E*
1	–1	–1	–1	–1
–1	1	–1	–1	–1
–1	–1	1	–1	–1
1	1	1	–1	–1
–1	–1	–1	1	–1
1	1	–1	1	–1
1	–1	1	1	–1
–1	1	1	1	–1
–1	–1	–1	–1	1
1	1	–1	–1	1
1	–1	1	–1	1
–1	1	1	–1	1
1	–1	–1	1	1
–1	1	–1	1	1
–1	–1	1	1	1
1	1	1	1	1

Note: In this design all main effects, 2-factor and 3-factor interactions are linear trend-free except the interaction BE which is nearly linear trend-free. Interactions CE and DE are not linear/ nearly linear trend-free.

Remark 6.1: For obtaining a linear trend-free for main effects design for complete factorial experiments, a simple version of AL1 is given in Remark 4.1. It needs to be further investigated whether such a simplification of AL2 is possible for obtaining a linear trend-free for main effects confounded factorial design.

7. DISCUSSION

The algorithms AL1 and AL2 have been translated in Microsoft Visual C++ program and these programmes have been used in computer aided generation of trend-free designs for the desired factorial experiment for any number of factors k (≥ 3). The catalogue of the designs obtained for 2^k factorial experiment, for $k = 3, \ldots, 7$ (for both without and with confounding factorial experiments) that are linear trend-free for main effects are available with the authors and can be obtained by sending an E-mail to klkalra@gmail.com.

ACKNOWLEDGEMENTS

The authors are grateful to the referee for his valuable suggestions that led to a considerable improvement in the presentation of results.

REFERENCES

Bradley, R.A. and Yeh, C.M. (1980). Trend-free block designs: Theory. *Ann. Statist*., **8**, 883-893.

Bradley, R.A. and Odeh, R.E. (1988). A generating algorithm for linear trend-free block designs. *Comm. Statist.-Simul. Comput*., **17**, 1259-1280.

Chai, F.S. and Majumdar D. (1993). On the Yeh-Bradley conjecture on linear trend-free block designs. *Ann. Statist*., **21**, 2087-2097.

Chai, F. S. (1995). Construction and optimality of nearly trend-free designs. *J. Statist. Plann. Inf.,* **48,** 113-129.

Cheng, C.S. and Jacroux, M. (1988). On the construction of trend-free run orders of two level factorial designs. *J. Amer. Statist. Assoc.,* **83**, 1152-1158.

Coster, D. C. and Cheng, C.S. (1988). Minimum cost trend-free run orders of fractional factorial designs. *Ann. Statist.,* **16,** 1188-1205.

Coster, D.C. (1993). Tables of minimum cost, linear trend-free run sequences for two and three-level fractional factorial design. *Compu. Statist. Data Analysis,* **16,** 325-336.

Daniel, C. and Wilcoxon, F. (1966). Factorial 2^{n-p} plans robust against linear and quadratic trends. *Technometrics*, **8**, 259-278.

Dhall, S.P. (1986). Some studies on robustness of designs, *Unpublished Ph.D. Thesis, IARI*, New Delhi.

Dwivedi, S.K. (1997). Computer aided search for optimal designs. *Unpublished Ph.D. Thesis,* IARI, New Delhi.

Gupta, V.K., Parsad, R., Bhar, L.M. and Kole, B. (2008). Computer generated supersaturated designs. *J. Ind. Soc. Agril. Statist. (In press)*

Hinkelmann, K. and Jo, J. (1998). Linear trend free Box-Behnken designs. *J. Statist. Plann. Inf.,* **72,** 347-354.

Jacroux, M., Majumdar, D. and Shah, K.R. (1995). Efficient block designs in the presence of trends. *Statistica Sinica,* **5**, 605-615.

Jacroux, M., Majumdar, D. and Shah, K.R. (1997). On the determination and construction of optimal block designs in the presence of linear trends. *J. Amer. Statist. Assoc.,* **92,** 375-382.

Kohli, P. (2006). A study on supersaturated designs. *Unpublished M. Sc. Thesis,* IARI, New Delhi.

John, P.W.M. (1990). Time trend and factorial experiments. *Technometrics,* **32,** 275-282.

Lal, K., Parsad, R. and Gupta, V.K. (2005). *A study on trend-free designs*. Project Report, IASRI, New Delhi.

Lal, K., Parsad, R. and Gupta, V.K. (2007). Trend-free nested balanced incomplete block designs and designs for diallel cross experiments. *Cal. Stat. Assoc. Bull.*, **59(237-238),** 203-221.

Majumdar, D. and Martin, R.J. (2002). Finding optimal designs in the presence of trends. *J. Statist. Plann. Inf.,* **106,** 177-190.

Nguyen, N.K. (1983). Computer aided construction of optimal designs. *Unpublished Ph.D*. Thesis, IARI, New Delhi.

Nguyen, N.K. (2004). Making experimental designs robust against trend. Submitted (Seen by the courtesy of author).

Rathore, A., Parsad, R., Gupta, V.K. (2004). Computer aided construction and analysis of augmented designs. *J. Ind. Soc. Agril. Statist.,* **57 (Special Volume)**, 320-344.

Rathore, A., Parsad, R., Gupta, V.K. (2006). Computer aided search of efficient block designs for making all possible pairwise treatment comparisons. *J. Statist. Appl.: A Publication of 'Forum for Interdisciplinary Mathematics',* **1(1),** 15-33.

Satpati, S.K, Parsad, R., Gupta, V.K. and Nigam, A.K. (2006a). Computer-aided search of efficient nested incomplete block designs for correlated observations. *J. Comb. Inf. Syst. Sci.,* **31(1-4)**, 163-186.

Satpati, S.K, Parsad, R. and Gupta, V.K. (2006b). Efficient block designs for dependent observations: A computer-aided search. *Comm. Statist.-Theory Methods,* **30(6),** 1187-1223.

Stufken, J. (1988). On the existence of linear trend-free block designs. *Comm. Statist.-Theory Methods,* **17,** 3857-3863.

URL: http://gendex.designcomputing.net/rat/

URL: http://www.iasri.res.in/design/

Yeh, C.M. and Bradley, R.A. (1983). Trend-free block designs: Existence and construction results. *Comm. Statist.—Theory Methods,* **12,** 1-21.

Available online at www.isas.org.in
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(2) 2009 175-180

A Class of Predictive Estimators in Two-Stage Sampling

L.N. Sahoo[1*], B.C. Das[1] and J. Sahoo[2]
[1]*Utkal University, Bhubaneswar 751004*
[2]*Department of Statistics, S.K.C.G. College, Parlakhemundi 761200*

(Received: May 2005, Accepted: August 2009)

SUMMARY

Under the well known prediction approach of Basu (1971), we introduce a new class of estimators for the finite population mean availing information on two auxiliary variables in a two-stage sampling.

Keywords: Asymptotic variance, Auxiliary variable, Prediction approach, Two-stage sampling.

1. INTRODUCTION

Consider a finite population U, partitioned into N first stage units (fsu) denoted by $U_1, U_2,..., U_N$ such that the number of second stage units (ssu) in U_i is M_i and $M = \sum_{i=1}^{N} M_i$. Let y_{ij} and x_{ij} denote values of the study variable y and an auxiliary variable x respectively, for the j^{th} ssu of U_i ($j = 1, 2, ..., M_i$; $i = 1, 2, ..., N$). Define $\bar{Y}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij}$, $\bar{X}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} x_{ij}$ as the means of U_i and $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} u_i\bar{Y}_i$, $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} u_i\bar{X}_i$ as the overall population means, where $u_i = NM_i/M$. To estimate $\bar{Y}$, assume that a sample s of n fsus is drawn from U and then a sample s_i of m_i ssus from the selected U_i is drawn according to the design simple random sampling without replacement. Let $\bar{y}_i = \frac{1}{m_i}\sum_{j\in s_i} y_{ij}$, $\bar{x}_i = \frac{1}{m_i}\sum_{j\in s_i} x_{ij}$, $\bar{y} = \frac{1}{n}\sum_{i\in s} u_i\bar{y}_i$, $\bar{x} = \frac{1}{n}\sum_{i\in s} u_i\bar{x}_i$ and $\bar{x}' = \frac{1}{n}\sum_{i\in s} u_i\bar{X}_i$.

When $\bar{X}$ is known accurately, Srivastava's (1980) class of estimators is defined by $t_s = \gamma(\bar{y},\bar{x})$, where $\gamma(\bar{y},\bar{x})$ is a function of $\bar{y}$ and $\bar{x}$, such that $\gamma(\bar{Y},\bar{X}) = \bar{Y}$ and satisfies certain regularity conditions in R_2, a 2-dimensional real space containing the point $(\bar{Y},\bar{X})$. But, in a two-stage sampling plan it is usually felt that efficiency of an estimator depends on how well the auxiliary information can be utilized at different stages. With this spirit, using known values of $\bar{X}_i$'s for the selected fsus, Sahoo and Panda (1997) considered a class of estimators defined by

$$t_{sp} = \mu(\frac{1}{n}\sum_{i\in s} u_i\mu_i(\bar{y}_i,\bar{x}_i),\bar{x}')$$

[*] *Corresponding author* : L.N. Sahoo
E-mail address : lnsahoostatuu@rediffmail.com

such that $\mu(\bar{Y},\bar{X})=\bar{Y}$, $\mu_i(\bar{Y}_i,\bar{X}_i)=\bar{Y}_i$, $i \in s$ and the functions $\mu(.,.)$ and $\mu_i(.,.)$ admit Srivastava's (1980) regularity conditions in R_2.

In certain practical situations we get information on another strong auxiliary variable z, taking value z_{ij} for the j^{th} ssu of U_i in such a way that the overall population mean $\bar{Z}$ is unknown, but the population means for the selected fsus *i.e.*, $\bar{Z}_i$, $i \in s$, are known. In this context, using both auxiliary variables, Sahoo and Sahoo (2005) composed a class of estimators defined by

$$t_{ss} = \alpha(\frac{1}{n}\sum_{i\in s} u_i \alpha_i(\bar{y}_i,\bar{z}_i),\bar{x}')$$

where $\bar{z}_i = \frac{1}{m_i}\sum_{j\in s_i} z_{ij}$ and the functions α and α_i satisfy regularity conditions in R_2. The basic assumption behind the construction of t_{ss} is that $\bar{X}_i, \bar{Z}_i, i \in s$ and $\bar{X}$ are known but $\bar{Z}$ is unknown. However, guided by these assumptions, here we develop a general class of estimators for $\bar{Y}$ motivated by the predictive approach of Basu (1971, p. 212, example 3).

As an example of this type of situation, we may refer to a crop survey conducted in a district with block (cluster of villages) as the fsu and village as the ssu. If y, x and z represent respectively yield, cultivated area and area under wheat, then information on the average area under cultivation per village in the i^{th} block, *i.e.*, $\bar{X}_i$ for $i \in s$, can be obtained at a low cost from the block records and average area under cultivation for the district *i.e.*, $\bar{X}$ can be known from the district records. Information on $\bar{Z}_i$ *i.e.*, average area under wheat for the i^{th} selected block can also be easily available from the block level records.

2. PREDICTION CRITERION IN TWO-STAGE SAMPLING

Let $\bar{s}$ denote the set of $(N-n)$ fsus of U which are not included in s and $\bar{s}_i$, the set of $(M_i - m_i)$ ssus of U_i which are not included in s_i, $i \in s$. Under the usual predictive set-up, it is possible to express

$$\bar{Y} = \frac{1}{M}\left[\sum_{i\in s}\left\{\sum_{j\in s_i} y_{ij} + \sum_{j\in \bar{s}_i} y_{ij}\right\} + \sum_{i\in\bar{s}} M_i \bar{Y}_i\right] \quad (1)$$

Writing $(N-n)\bar{Y}_r = \sum_{i\in\bar{s}} u_i \bar{Y}_i$

and $(M_i - m_i)\bar{Y}_{ir} = \sum_{j\in\bar{s}_i} y_{ij}$, we have

$$\bar{Y} = \frac{1}{M}\left[\sum_{i\in s}\{m_i\bar{y}_i + (M_i - m_i)\bar{Y}_{ir}\}\right] + \frac{N-n}{N}\bar{Y}_r \quad (2)$$

To estimate $\bar{Y}$, we, therefore have to predict the quantities $\bar{Y}_{ir}$ and $\bar{Y}_r$ from the sample data because the first component of the right hand side of (2) is already known. Using T_i and T as their predictors, a predictor $\hat{\bar{Y}}$ of $\bar{Y}$ of may be defined by the equation

$$\hat{\bar{Y}} = \frac{1}{M}\left[\sum_{i\in s}\{m_i\bar{y}_i + (M_i - m_i)T_i\}\right] + \frac{N-n}{N}T \quad (3)$$

Note that if $m_i = M_i$ and $n = N$; $\hat{\bar{Y}} = \bar{Y}$ the target of our prediction.

Corresponding to various suitable choices of the predictors $T_i (i \in s)$ and T, equation (3) generates a class of estimators. But, we achieve this objective by defining these predictors in terms of two auxiliary variables *i.e.*, x and z.

3. THE CLASS OF PREDICTIVE ESTIMATORS

For given s_i and s, let

$$e_i = (\bar{y}_i, \bar{x}_i, \bar{z}_i, \bar{X}_{ir}, \bar{Z}_{ir}) \in R_5$$

and $$e = (\bar{y}, \bar{x}', \bar{X}_r) \in R_3,$$

where $(M_i - m_i)\bar{X}_{ir} = \sum_{j\in\bar{s}_i} x_{ij}$, $(M_i - m_i)\bar{Z}_{ir} = \sum_{j\in\bar{s}_i} z_{ij}$, $(N-n)\bar{X}_r = \sum_{i\in\bar{s}} u_i \bar{X}_i$, R_5 and R_3 are 5- and 3-dimensional real spaces containing the points $E_i = (\bar{Y}_i, \bar{X}_i, \bar{Z}_i, \bar{X}_i, \bar{Z}_i)$, $i \in s$ and $E = (\bar{Y}, \bar{X}, \bar{X})$ respectively. Further, let $h_i(e_i)$ and $h(e)$ be some known functions of e_i and e respectively such that $h_i(E_i) = \bar{Y}_i$, $i \in s$, and $h(E) = \bar{Y}$. Let us assume that

(a) the functions h_i and h are continuous in R_5 and R_3 respectively, and

(b) the first and second order partial derivatives of these functions with respect to their arguments exist and are also continuous in their respective range spaces.

Thus, based on information available on s_i and s, $h_i(e_i)$ and $h(e)$ clearly define classes of estimators for $\bar{Y}_i, i \in s$, and $\bar{Y}$ respectively. Using $h_i(e_i)$ and $h(e)$ as predictors in places of T_i and T in our predictive equation (3), we now define a class of predictive estimators for $\bar{Y}$ by

$$t_h = \frac{1}{M}\left[\sum_{i\in s}\left\{m_i\bar{y}_i + (M_i - m_i)h_i(e_i)\right\}\right] + \frac{N-n}{N}h(e)$$

Many estimators may turn out as special cases of t_h corresponding to various selections of h_i and h. Let us consider the following simple cases:

(i) If the information on x is completely ignored, *i.e.*, if $h_i = \bar{y}_i$ and $h = \bar{y}$ then t_h becomes $\bar{y}$, the simple expansion estimator of $\bar{Y}$.

(ii) When $h_i = \dfrac{\bar{y}_i\bar{X}_{ir}\bar{Z}_{ir}}{\bar{x}_i\bar{z}_i}$ and $h = \dfrac{\bar{y}\bar{X}_r}{\bar{x}'}$, then

$$t_h \to t_R^{(h)} = \bar{y}\frac{\bar{X}}{\bar{x}'} - \frac{f}{n}\sum_{i\in s}u_i\bar{y}_i \times \left[(1-f_i) - \frac{1}{1-f_i}\left(\frac{\bar{X}_i}{\bar{x}_i} - f_i\right)\left(\frac{\bar{Z}_i}{\bar{z}_i} - f_i\right)\right]$$

a ratio-type estimator, where $f = \dfrac{n}{N}$ and $f_i = \dfrac{m_i}{M_i}$.

(iii) When $h_i = \dfrac{\bar{y}_i\bar{x}_i\bar{z}_i}{\bar{X}_{ir}\bar{Z}_{ir}}$ and $h = \dfrac{\bar{y}\,\bar{x}'}{\bar{X}_r}$, then

$$t_h \to t_P^{(h)} = \frac{f}{n}\sum_{i\in s}u_i \times \left[f_i\bar{y}_i + (1-f_i)^3\frac{\bar{y}_i\bar{x}_i\bar{z}_i}{(\bar{X}_i - f_i\bar{x}_i)(\bar{Z}_i - f_i\bar{z}_i)}\right] + (1-f)^2\frac{\bar{y}\,\bar{x}'}{\bar{X} - f\,\bar{x}'}$$

a product-type estimator.

(iv) When $h_i = \bar{y}_i - \beta_{iyx}(\bar{x}_i - \bar{X}_{ir}) - \beta_{iyz}(\bar{z}_i - \bar{Z}_{ir})$ and $h = \bar{y} - \beta_{byx}(\bar{x}' - \bar{X}_r)$

$$t_h \to t_{RG}^{(h)} = \bar{y} - \frac{f}{n}\sum_{i\in s}u_i\left\{\beta_{iyx}(\bar{x}_i - \bar{X}_i) + \beta_{iyz}(\bar{z}_i - \bar{Z}_i)\right\} - \beta_{byx}(\bar{x}' - \bar{X})$$

a regression-type estimator, where

$\beta_{\text{iyz}} = S_{iyx} / S_{ix}^2$,

$\beta_{\text{iyz}} = S_{iyz} / S_{iz}^2$, $\beta_{\text{byz}} = S_{byx} / S_{bx}^2$ such that

$$S_{iyx} = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(y_{ij} - \bar{Y}_i)(x_{ij} - \bar{X}_i)$$

$$S_{byx} = \frac{1}{N-1}\sum_{i=1}^{N}(u_i\bar{Y}_i - \bar{Y})(u_i\bar{X}_i - \bar{X})$$

$$S_{iz}^2 = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(z_{ij} - \bar{Z}_i)^2$$

$$S_{by}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(u_i\bar{Y}_i - \bar{Y})^2, \text{ etc.}$$

(v) If the estimation procedure is carried out with the involvement of x only, then $h_i = d_i(\bar{y}_i, \bar{x}_i, \bar{X}_{ir})$ so that $t_h \to t_h^{(d)}$, a class of predictive estimators defined by

$$t_h^{(d)} = \frac{1}{M}\left[\sum_{i\in s}\left\{m_i\bar{y}_i + (M_i - m_i)d_i(\bar{y}_i, \bar{x}_i, \bar{X}_{ir})\right\}\right] + \frac{N-n}{N}h(e)$$

(vi) As a specific case of t_h, we may also consider another subclass of predictive estimators defined by

$$t_h^{(k)} = \frac{1}{M}\left[\sum_{i\in s}\left\{m_i\bar{y}_i + (M_i - m_i)k_i(\bar{y}_i, \bar{z}_i, \bar{Z}_{ir})\right\}\right] + \frac{N-n}{N}h(e)$$

on considering $h_i = k_i\ (\bar{y}_i, \bar{z}_i, \bar{Z}_{ir})$.

4. ASYMPTOTIC VARIANCE OF t_h

Expanding $h_i(e_i)$ and around the points E_i and E respectively in a first order Taylor's series and then neglecting the remainder term, we get

$$h_i(e_i) = h_i(E_i) + h_{i0}(\bar{y}_i - \bar{Y}_i) + h_{i1}(\bar{x}_i - \bar{X}_i) + h_{i2}(\bar{z}_i - \bar{Z}_i) + h_{i3}(\bar{X}_{ir} - \bar{X}_i) + h_{i4}(\bar{Z}_{ir} - \bar{Z}_i) \quad (4)$$

and

$$h(e) = h(E) + h_0(\bar{y} - \bar{Y}) + h_1(\bar{x}' - \bar{X}) + h_2(\bar{X}_r - \bar{X}) \quad (5)$$

where h_{i0}, h_{i1}, h_{i2}, h_{i3}, and h_{i4} are respectively the values of first order partial derivatives of $h_i(e_i)$ with respect to $\bar{y}_i, \bar{x}_i, \bar{z}_i, \bar{X}_{ir}$ and $\bar{Z}_{ir}$ at E_i and h_0, h_1 and h_2 are respectively the values of first order partial derivatives of $h(e)$ with respect to $\bar{y}, \bar{x}'$ and $\bar{X}_r$ at E.

Noting that $h_{i0} = 1$, $h_{i1} = -h_{i3}$, $h_{i2} = -h_{i4}$, $\bar{X}_{ir} = \dfrac{M_i\bar{X}_i - m_i\bar{x}_i}{M_i - m_i}$, $\bar{Z}_{ir} = \dfrac{M_i\bar{Z}_i - m_i\bar{z}_i}{M_i - m_i}$ we have after a considerable simplification

$$t_h = \bar{y} + \frac{f}{n}\sum_{i\in s} u_i[h_{i1}(\bar{x}_i - \bar{X}_i) + h_{i2}(\bar{z}_i - \bar{Z}_i)] + h_1(\bar{x}' - \bar{X}) \quad (6)$$

Hence, after a few tedious algebraic steps (suppressed to save space), the asymptotic variance of t_h is obtained as

$$V(t_h) = \frac{1-f}{n}\left(S_{by}^2 + h_1^2 S_{bx}^2 + 2h_1 S_{byx}\right) + \frac{1}{nN}\sum_{i=1}^{N} u_i^2 \frac{1-f_i}{m_i} V \quad (7)$$

where $V_i = S_{iy}^2 + f^2 h_{i1}^2 S_{ix}^2 + f^2 h_{i2}^2 S_{iz}^2 + 2fh_{i1}S_{iyx} + 2fh_{i2}S_{iyz} + 2f^2 h_{i1}h_{i2}S_{ixz}$

Minimizing $V(t_h)$ over h_{i1}, h_{i2} and h_1 we get

$$h_{i1} = -\frac{1}{f}\frac{\beta_{iyx} - \beta_{iyz}\beta_{izx}}{1 - \beta_{izx}\beta_{ixz}} = h_{i1}^* \quad \text{(say)}$$

$$h_{i2} = -\frac{1}{f}\frac{\beta_{iyz} - \beta_{iyx}\beta_{ixz}}{1 - \beta_{izx}\beta_{ixz}} = h_{i2}^* \quad \text{(say)}$$

and $h_1 = -\beta_{byx}$

where $\beta_{izx} = S_{izx} / S_{ix}^2$, $\beta_{ixz} = S_{ixz} / S_{iz}^2$. Use of these optimum values in (7) yields the minimum asymptotic variance of the class (may be called as the asymptotic minimum variance bound (MVB) of the class) is given by

$$\min V(t_h) = \frac{1-f}{n}S_{by}^2(1-\rho_{byx}^2) + \frac{1}{nN}\sum_{i=1}^{N} u_i^2 \frac{1-f_i}{m_i} S_{iy}^2(1-\rho_i^2) \quad (8)$$

where $\rho_{byx} = S_{byx} / S_{by}S_{bx}$ and

$$\rho_i = \sqrt{\frac{\rho_{iyx}^2 + \rho_{iyz}^2 - 2\rho_{iyx}\rho_{iyz}\rho_{ixz}}{1-\rho_{ixz}^2}}$$

the multiple correlation coefficient of y on x and z in U_i such that $\rho_{iyx} = S_{iyx} / S_{iy}S_{ix}$ etc. An estimator attaining this bound is called as an MVB estimator. In the present context our MVB estimator is a regression-type estimator of the form

$$t_{RG}^0 = \bar{y} - \frac{1}{n}\sum_{i\in s} u_i\left[h_{i1}^*(\bar{x}_i - \bar{X}_i) + h_{i2}^*(\bar{z}_i - \bar{Z}_i)\right] - \beta_{byx}(\bar{x}' - \bar{X})$$

The parametric functions h_{i1}^*, h_{i2}^* and β_{byx} can be replaced by their consistent estimates computed from the sample itself. But, the asymptotic variance of the resulting estimator remains unchanged and is given by (8).

5. PRECISION OF t_h

In an effort to study the efficiency aspect of the predictive method of estimation developed in this work in relation to the classical method, our first attempt is to compare the efficiency of t_h with that of t_s. The asymptotic variance of t_s obtained through Taylor linearization is given by

$$V(t_s) = \frac{1-f}{n}(S_{by}^2 + \gamma_1^2 S_{bx}^2 + 2\gamma_1 S_{byx}) + \frac{1}{nN}\sum_{i=1}^{N} u_i^2 \frac{1-f_i}{m_i}(S_{iy}^2 + \gamma_1^2 S_{ix}^2 + 2\gamma_1 S_{iyx}) \quad (9)$$

where γ_1 is the first order partial derivative of $\gamma(\bar{y}, \bar{x})$ with respect to $\bar{x}$ when evaluated at $(\bar{Y}, \bar{X})$.

From (7) and (9), it follows that $V(t_h) \le V(t_s)$ *i.e.*, an estimator of t_h is more precise than an estimator of t_s if

$$\left|\gamma_1 + \beta_{byx}\right| \ge \left|h_1 + \beta_{byx}\right|$$

and $S_{ix}^2\left[(\gamma_1 + \beta_{iyx})^2 - (fh_{i1} + \beta_{iyx})^2\right]$

$$\ge fh_{i1}S_{iz}^2(fh_{i2} + 2\beta_{iyz} + 2fh_{i1}\beta_{ixz}) \quad \forall \; i \quad (10)$$

These sufficient conditions basically depend on the choices of different functions for composing t_h and t_s. However, they give some indication that there is enough scope for improving upon the estimators through our predictive method over classical method. But, these conditions can not lead to any straight forward conclusions if the characteristics of the functions are unknown. However, for simplicity, if we accept MVB as an intrinsic measure of precision of a class, the problem of precision comparison seems to be easier and our attention will be concentrated on the MVB estimators only.

The minimum asymptotic variance of t_s is

$$\min V(t_s) = \frac{1-f}{n}S_{by}^2(1-\rho^2) + \frac{1}{nN}\sum_{i=1}^{N} u_i^2 \frac{1-f_i}{m_i} S_{iy}^2(1-\rho^2) \quad (11)$$

and the corresponding MVB estimator is

$$t_{RG}^{(s)} = \bar{y} - \beta(\bar{x} - \bar{X})$$

where ρ is the correlation coefficient between $\bar{y}$ and $\bar{x}$ and β is the regression coefficient of $\bar{y}$ on $\bar{x}$. Hence, we see that

$$\min V(t_h) \le \min V(t_s)$$

i.e., t_{RG}^0 is more efficient than $t_{RG}^{(s)}$ if

$$\rho^2 \le \rho_{byx}^2 \text{ and } \rho_i^2 \quad \forall \; i \quad (12)$$

Turning our attention to study the precision of t_h compared to other classes of classical and predictive estimators *viz.*, t_{sp}, t_{ss}, $t_h^{(d)}$ and $t_h^{(k)}$ on the ground of MVB criterion, we see that

$$\min V(t_{sp}) = \frac{1-f}{n}S_{by}^2(1-\rho_{byx}^2) + \frac{1}{nN}\sum_{i=1}^{N} u_i^2 \frac{1-f_i}{m_i} S_{iy}^2(1-\rho_{iyx}^2) \quad (13)$$

$$\min V(t_{ss}) = \frac{1-f}{n}S_{by}^2(1-\rho_{byx}^2) + \frac{1}{nN}\sum_{i=1}^{N} u_i^2 \frac{1-f_i}{m_i} S_{iy}^2(1-\rho_{iyz}^2) \quad (14)$$

$$\min V(t_h^{(d)}) = \min V(t_{sp})$$

$$\min V(t_h^{(k)}) = \min V(t_{ss})$$

The MVB estimators of t_{sp} or $t_h^{(d)}$ and t_{ss} or $t_h^{(k)}$ are also respectively given by

$$t_{RG}^{(sp)} = \bar{y} - \frac{1}{n}\sum_{i\in s} u_i \beta_{iyx}(\bar{x}_i - \bar{X}_i) - \beta_{byx}(\bar{x}' - \bar{X})$$

$$t_{RG}^{(ss)} = \bar{y} - \frac{1}{n}\sum_{i\in s} u_i \beta_{iyz}(\bar{z}_i - \bar{Z}_i) - \beta_{byx}(\bar{x}' - \bar{X})$$

From (8), (13) and (14) we have

$$\min V(t_h) \le \min V(t_{sp}) \Rightarrow V(t_{RG}^0) \le V(t_{RG}^{(sp)})$$

and $\min V(t_h) \le \min V(t_{ss}) \Rightarrow V(t_{RG}^0) \le V(t_{RG}^{(ss)})$

Hence, we may conclude that t_h is superior to t_{sp}, t_{ss}, $t_h^{(d)}$ and $t_h^{(k)}$ on the ground of MVB criterion.

6. NUMERICAL STUDY

To study precision of the suggested methodology numerically, we consider data of two populations as described below.

Population 1. Consists of 198 blocks (ssu) divided into $N = 27$ wards of Berhampur city of Orissa. The number of blocks (M_i) of 27 wards are 6, 6, 12, 5, 6, 6, 10, 5, 6, 6, 6, 6, 6, 12, 6, 7, 7, 7, 10, 6, 6, 7, 10, 11, 9, 8 and 6. The three variables *viz*., number of educated females, female population and number of households are used as y, x and z respectively, data on which are available in Census of India (1971) document. We have taken $n = 9$ and $m_i = 2, 2, 4, 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 4, 2, 2, 2, 2, 3, 2, 2, 2, 3, 4, 3, 3$ and 2 respectively.

Population 2. MU284 population available in Sarndal *et al.* (1992, p. 660, Appendix C). It consists of 284 municipalities (ssu) divided into 50 clusters (fsu) with three variables *viz*., Revenue from the 1985 municipal taxation as y, 1975 population as x and 1985 population as z. We consider $n = 12$, and $m_i = 2$ for every i.

Relative precision of different MVB estimators compared to the simple expansion estimator $\bar{y}$, are compiled in Table 1. The estimator t_{RG}^{0} attains the maximum precision for both populations. Thus, our numerical study shows that the new methodology developed here to create predictive estimators may be useful for many practical situations.

Table 1. Relative precision of different estimators

Pop. No.	Estimators				
	$\bar{y}$	$t_{RG}^{(s)}$	$t_{RG}^{(sp)}$	$t_{RG}^{(ss)}$	t_{RG}^{0}
1	100	148	184	175	195
2	100	947	3579	3366	3725

REFERENCES

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part I. In: *Foundations of Statistical Inference*, V.P. Godambe and D.A. Sprott (eds), Holt, Rinehart and Winston, Toronto, Canada, 203-242.

Sahoo, L.N. and Panda, P. (1997). A class of estimators in two-stage sampling with varying probabilities. *South African Statist. J.*, **31**, 151-160.

Sahoo, L.N. and Sahoo, R.K. (2005). On the construction of a class of estimators in two-stage sampling. *J. Appl. Statist. Sci.*, **14**, 317-322.

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

Srivastava, S.K. (1980). A class of estimators using auxiliary information in sample surveys. *Canad. J. Stat.*, **8**, 253-254.

Available online at www.isas.org.in
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(2) 2009 181-188

Post Processing of Clusters for Pattern Discovery: Rough Set Approach

Alka Arora[1*], Shuchita Upadhyaya[2] and Rajni Jain[3]
[1]*Indian Agricultural Statistics Research Institute, New Delhi*
[2]*Kurukshetra University, Kurukshetra*
[3]*National Center for Agricultural Economics and Policy Research, New Delhi*

(Received: February 2008, Revised: May 2009, Accepted: May 2009)

SUMMARY

Most of clustering algorithms generate clustering results in the form of number of clusters and member objects in those clusters. This further requires analysis by experts in order to understand the patterns of obtained clusters. Post processing of cluster is then required in order to extract meaningful cluster pattern. In this paper a rough set based approach for pattern discovery from individual clusters is proposed. In the proposed approach, Maximum Possible Combination Reduct (MPCR) derived from rough set theory is used for generating concise cluster pattern. MPCR is defined as the set of variables which distinguishes the objects in a homogenous cluster. Therefore these variables are not considered for pattern formulation. Remaining variables are ranked for their contribution in the cluster. Cluster pattern is formed by conjunction of variables in the increasing order of their contribution in the cluster such that pattern distinctively describes the cluster with minimum error. Applicability of approach is demonstrated using soybean disease and zoo datasets from machine learning repository.

Keywords: Clustering, Data mining, Rough set theory, Reduct, Indiscernibility, MPCR, Cluster description, Pattern.

1. INTRODUCTION

Data Mining is a non trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data (Han and Kamber 2006). Clustering is an important component of data mining. The underlying assumption of clustering in data mining is to find out the hidden patterns in the data, which can be revealed by grouping the objects into clusters. According to Mirkin (2005), clustering process involves different stages, which include data pre-processing and standardization, finding clusters in data and description of clusters. Many clustering algorithms are available in literature, one can refer to Jain *et al*. 1999; Han and Kamber 2006; Mirkin 2005; for comprehensive surveys on clustering algorithms. *K*-Means and Expectation Maximization (EM) algorithms are the widely known partitional algorithms, which divide the data into k non overlapping clusters. These clustering algorithms just generate general description of the clusters like which objects are member of each cluster and lacks in generating cluster description in terms of relevant variables those define the cluster. According to Ganter and Wille (1997), cluster description is able to approximately describe the cluster in the form that "this cluster consists just of all the objects having the pattern P, where pattern is formulated using the variable and values of the given many valued context". From an intelligent data analysis perspective deriving knowledge in the form of pattern from obtained clusters is as important as grouping the objects into clusters.

[*] *Corresponding author* : Alka Arora
E-mail addresses : alkak@iasri.res.in, alka27@yahoo.com

Rough Set Theory (RST) proposed by Pawlak (1991), has been successfully applied in classification techniques for pattern/knowledge discovery (Komorowski 1999). RST has also relevance in clustering as RST divides the data into equivalence/ indiscernible classes; each indiscernible class can be considered as natural cluster. Moreover, RST performs automatic concept approximation by producing minimal subset of variables (Reduct) which can distinguish indiscernible classes in the dataset. In general, classification problems using rough sets involve computation of decision relative reduct. Clustering, an unsupervised method of data mining requires reduct computation purely on the basis of indiscernibility as there is no decision variable. Such reduct are referred as unsupervised reduct in this paper.

The proposed approach of cluster description is applied as post processing step on obtained clusters. As our aim is to generate characteristics of individual clusters, hence partition based algorithm is used to obtain non overlapping clusters. Applicability of proposed approach is studied on soybean disease and zoo datasets of agriculture domain from UCI repository (UCI). Objective of applying the proposed approach on soybean dataset is to study the relevant variables which contribute towards the occurrence of a particular disease and on zoo dataset is to characterize the animal clusters.

The paper is organized in six sections. Section 2 provides overview of rough set concepts. Section 3 gives background and related work in the area of cluster description. Section 4, provides the details of proposed approach. Section 5 details the application of proposed approach on soybean disease and zoo datasets followed by conclusions in Section 6.

2. ROUGH SET THEORY : A BRIEF OVERVIEW

RST is a mathematical approach, proposed by Pawlak (1991), further refined by Komorowski and Polkowski (1999), Yao *et al.* (1997), to cope with data analysis in the presence of imprecision, vagueness and uncertainty. In RST, dataset is represented in the form of information table; each case represents an object and columns represent variables. More formally it is an information system $X = (U, A)$ where U is non-empty, finite set of objects called the universe and A is non-empty, finite set of variables on U. With every variable $a \in A$, a set V_a is associated such that $a : U \rightarrow V_a$. The set V_a is called the domain or value set of variable a. Small table from soybean disease dataset is used for illustration (Table 1). The dataset has ten objects characterized by eight nominal variables.

2.1 Indiscernibility Relation

Indiscernibility relation is core concept of RST. Indiscernibility relation $IND(B)$, for any subset $B \subseteq A$ is defined by

$$IND(B) = \{(x, y) \mid a(x) = a(y),\ \forall\ a \in B\ ;\ x, y \in U\}$$

Two objects are considered to be indiscernible or similar by the variables in B, if and only if they have

Table 1. Small soybean dataset

id	date	precip	damage	severity	canker_lesion	fruiting_bodies	decay
X1	july	lt-norm	scattered	pot-severe	tan	absent	absent
X2	october	norm	scattered	pot-severe	tan	absent	absent
X3	september	lt-norm	whole-field	pot-severe	tan	absent	absent
X4	august	norm	whole-field	pot-severe	tan	present	absent
X5	august	lt-norm	upper-area	pot-severe	tan	absent	absent
X6	september	gt-norm	whole-field	pot-severe	dk-brown-blk	absent	absent
X7	july	gt-norm	scattered	pot-severe	dk-brown-blk	absent	firm-and-dry
X8	august	gt-norm	low-areas	pot-severe	dk-brown-blk	absent	firm-and-dry
X9	september	gt-norm	upper-area	minor	dk-brown-blk	absent	firm-and-dry
X10	october	gt-norm	whole-field	minor	dk-brown-blk	absent	firm-and-dry

the same value for every variable in B. Objects in the information system which have the same value form an equivalence relation. Equivalence relation partition set of objects (U) into set of equivalence classes. $IND(B)$ is an equivalence relation which partitions U into set of partitions denoted by $U / IND(B)$.

For example from Table 1, when B = {damage} then objects X1, X2 and X7 are indiscernible and therefore form one equivalence class; X3, X4, X6 and X10 are indiscernible and X5 is indiscernible with X9. Formally:

$$U/IND\{damage\} = \{\{X1,X2,X7\}, \{X3,X4,X6,X10\}, \{X5,X9\},\{X8\}\}$$

Similarly $U/IND\{canker\_lesion, decay\}$

$$= \{\{X1,X2,X3,X4,X5\},\{X6\}, \{X7,X8,X9,X10\}\}$$

2.2 Reduct

Concept approximation is achieved in RST through data reduction i.e. by retaining the minimum subset of variables that can differentiate all equivalence classes in the universe set. Such minimum subset is called reduct. More formally reduct R is a set of variables such that

$$R \subseteq A$$

$$IND_R(U) = IND_A(U)$$

$$IND_{R-a}(U) \neq IND_A(U) \qquad \forall\ a \in R$$

There are many methods as well as many software's available for computation of reduct, discussion on those is beyond the scope of this paper. We have considered Genetic Algorithm (GA) (Wroblewski 1995) for reduct computation, as it can produce many reducts of varying cardinality. This provides flexibility to the experimenter for selection of variables from the reduct population produced by GA. There are many approaches to consider variables from reducts generated by GA (Komorowski and Polkowski 1999). Maximum Possible Combined Reduct (MPCR) is defined as the union of variables present in the reduct sets obtained after applying GA (Jain 2004). Any variable that belongs to at least one of the reduct in the population of reducts from GA also belongs to MPCR. More formally MPCR is set of variables M, such that

$$M \subseteq A$$

$$M = \bigcup_{i=1}^{n} R_i$$ where R_i is the i^{th} reduct in the population of reducts from GA.

For Example, reduct computation on the Table 1 resulted in six reducts of cardinality three; R1 = {date, damage, canker_lesion}, R2 = {precip, damage, severity}, R3 = {date, precip, severity}, R4 = {date, precip, damage}, R5 = {date, precip, decay} and R6 = {precip, damage, decay}. MPCR set computation from these reducts is {date, precip, severity, damage, decay, canker_lesion}.

3. BACKGROUND AND RELATED WORK

Cluster description is useful in studying the object variable relationship which describes the underlying cluster. This can be applied in various areas for understanding the clusters viz. In disease diagnostic system, where there is a need to study the diseases characteristics; In Web Mining, finding pattern in the set of web users; Given a set of tourist places, finding out what features of places and tourist attract each other; In banks, customer data is available on many variables, discovery of age and salary as sufficient variables to grant loan to a customer; In characterization of animal and plant taxonomy clusters.

3.1 Review of Literature

In the literature, Mirkin (2005), Han and Kamber (2001), the problem of conceptual description of partition has received by far more attention than the problem of description of a single cluster. Decision tree is mainly used for conceptual description of partition as it provides easily understandable description. Primary goal of building a decision tree is prediction of the partition under consideration rather than its description. Limitation of this technique is that it is 'monothetic' and hence each split goes along with only one variable, and not directly applicable to cluster whose definition involve combination of variables. In clustering, the criterion is to get clusters as homogenous as possible with regard to all the variables however in decision tree; criterion is homogeneity with regard to a pre-specified decision variable.

As discussed by Mirkin (2005), the problem of producing description for a single cluster without any

relevance to other clusters has recently attracted considerable attention from the researchers. There are few references of cluster description approaches available in literature. Mirkin (1999) has proposed a method for cluster description applicable to only continuous variables. In Mirkin's approach variables are normalized first and then ordered according to their contribution weights which are proportional to the squared differences between their within group averages and grand means. A conjunctive description of cluster is then formed by consecutively adding variables according to the sorted order. Description is evaluated on precision error. Abidi *et al*. (1998, 2001) has proposed the rough set theory based method for rule creation for unsupervised data using dynamic reduct. Dynamic reduct is defined as the frequently occurring reduct set from the samples of original decision table. However, these approaches have their limitations. Mirkin's (1999) approach is applicable only to datasets having continuous variables. Abidi *et al*. (1998, 2001) in his approach has used the cluster information obtained after cluster finding and generated rules from entire data with respect to cluster/class attribute, instead of producing description for individual clusters. However, our approach is to generate user understandable cluster description for individual clusters by conjunction of significant variables which define the cluster.

3.2 Cluster Description Evaluation Criteria

As discussed by Mirkin (1999), accuracy of obtained pattern is measured in terms of Precision Error (*PE*). *PE* of pattern *P*, *PE* (*P*) is defined as

$$PE(P) = \frac{\left| false\ positive\ C(P) \right|}{\left| U - C \right|} \qquad (1)$$

where numerator, *false positive C(P)* is defined as the number of objects that lies outside cluster *C*, for which pattern *P* is true and denominator denotes the number of objects outside *C*.

4. PROPOSED APPROACH (REDUCT DRIVEN CLUSTER DESCRIPTION–RCD)

Proposed pattern discovery approach for individual clusters, called RCD is applicable as post processing step to clusters obtained using partition based clustering algorithm. RCD approach is divided into three stages.

4.1 Cluster Finding

First stage deals with obtaining clusters from dataset by applying clustering algorithm. We have used Weka implementation of EM algorithm for cluster finding (Weka). EM models the distribution of the objects probabilistically, so that an object belongs to a cluster with certain probability. The first step, calculation of the cluster probabilities, which are the expected class value, is "expectation"; the second step which deals with calculation of the distribution parameter is "maximization" of the likelihood of the distribution given the data (Mirkin 2005).

We have selected EM algorithm as it can handle both numeric and nominal variables. Weka implementation of EM algorithm has built in evaluation measure for computing the number of clusters present in the dataset. EM selects the number of clusters automatically by maximizing the logarithm of the likelihood of future data, estimated using cross-validation. Beginning with one cluster, it continues to add clusters until the estimated log-likelihood decreases (Weka).

4.2 Computation of Unsupervised Reduct

Clustering algorithm is intended to form clusters having most variable values common to their members (cohesion) and few values common to members of other clusters (distinctiveness) (Talavera 1999). Hence, variables which have similar value for majority of objects in the cluster are considered significant and rest are non significant for generating cluster pattern (Arora 2007).

Reduct accounts for discerning between the objects in a cluster, hence computation of unsupervised reduct in a cluster C provides the set of non significant variables for that cluster. Genetic Algorithm produces many reducts, hence computation of MPCR set (*RC*) in a cluster *C* provides the set of non-significant variables for that cluster. These non-significant variables (reduct) can be straight away removed from the cluster. The remaining variables (non reduct) form the set of significant variables (I) for that cluster.

4.3 Cluster Description

Cluster description approximately describes the cluster in the form of pattern. Pattern is formulated by

conjunction of significant attribute = value pairs from that cluster. There can be many possible patterns for a single cluster. Our aim is not to generate all possible patterns, but meaningful and concise pattern from the cluster. Therefore attributes in set are then ranked on Precision Error (*PE*) which is defined as

$$PE\ (a = v) = \frac{\left|\ false\ \ positive\ \ C(a=v)\ \right|}{\left|\ U-C\ \right|} \qquad (2)$$

where numerator defines the number of entities that lies outside cluster C, for which $a = v$ ($a \in A$, $v \in V_a$) is true and denominator defines the number of entities outside cluster C. An attribute value pair $a = v$ is said to be more contributing if it has less *PE*, means majority of objects satisfying this attribute value pair belongs to a single cluster.

Therefore problem of cluster description can be defined as forming a description P by combining the significant variables with less *PE* such that *PE* for P is minimum. Hence pattern P distinctively describes the cluster.

Procedure for RCD approach

1. Obtain clusters by applying partitional clustering algorithm.
2. Compute unsupervised reduct for individual clusters and then compute MPCR set (*RC*) for every cluster C.
3. Compute set of significant variables (I) for C, where $I = A - RC$.
4. Calculate *PE* for significant variables in set I for cluster C and arrange the set I in increasing order of *PE* score.
5. Combine variables from I with less *PE* to make the description such that *PE* for that description is minimum.

5. EXAMPLE OF APPLICATION

In this section, we illustrate the application of RCD approach on soybean disease dataset, followed by results of the same on Zoo dataset from UCI repository.

5.1 Soybean Dataset

In soybean disease set, Universal set (U) contains 47 objects and set of variables (A) consist of 35 multi-valued variables characterizing diaporthe-stem-canker, charcoal-rot, rhizoctonia-root-rot and phytophthora-rot diseases. All the variables are nominal in nature. Variables are broadly categorized into environmental descriptors, condition of leaves, condition of stem, condition of fruit pods and condition of root. Table 2 shows variable information of soybean dataset. It is observed that dataset is having unique value for some of the variables hence those variables are irrelevant and removed from the dataset. Reduced dataset then has 20 variables characterizing soybean

Table 2. Variable information of soybean dataset

v1	date: april=0, may=1, june=2, july=3, august=4, september=5, october=6
v2	plant-stand: normal=0, lt-normal=1
v3	precip: lt-norm=0, norm=1, gt-norm=2
v4	temp: lt-norm=0, norm=1, gt-norm=2
v5	hail: yes=0, no=1
v6	crop-hist: diff-lst-year=0, same-lst-yr=1, same-lst-two-yrs=2, same-lst-sev-yrs=3
v7	area-damaged: scattered=0, low-areas=1, upper-areas=2, whole-field=3
v8	severity: pot-severe=1, severe=2
v9	seed-tmt: none=0, fungicide=1
v10	germination: '90-100%'=0, '80-89%'=1, 'lt-80%'=2
v11	plant-growth: abnorm=1
v12	leaves: norm=0, abnorm=1
v13	leafspots-halo: absent=0
v14	leafspots-marg: dna=2
v15	leafspot-size: dna=2
v16	leaf-shread: absent=0
v17	leaf-malf: absent=0
v18	leaf-mild: absent=0
v19	stem: abnorm=1
v20	lodging: yes=0, no=1
v21	stem-cankers: absent=0, below-soil=1, above-soil=2, above-sec-nde=3
v22	canker-lesion: dna=0, brown=1, dk-brown-blk=2, tan=3
v23	fruiting-bodies: absent=0, present=1
v24	external decay: absent=0, firm-and-dry=1
v25	mycelium: absent=0, present=1
v26	int-discolor: none=0, black=2
v27	sclerotia: absent=0, present=1
v28	fruit-pods: norm=0, dna=3
v29	fruit spots: dna=4
v30	seed: norm=0
v31	mold-growth: absent=0
v32	seed-discolor: absent=0
v33	seed-size: norm=0
v34	shriveling: absent=0
v35	roots: norm=0, rotted=1

diseases. Dataset consist of instance number and class variables that are not considered while clustering.

EM clustering algorithm learnt four clusters from the dataset. Table 3 shows the dataset along with cluster information.

Table 3. Soybean dataset with clustering results

Sno	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v12	v20	v21	v22	v23	v24	v25	v26	v27	v28	v35	Cluster
0	4	0	2	1	1	1	0	1	0	2	1	0	3	1	1	1	0	0	0	0	0	cluster1
1	5	0	2	1	0	3	1	1	1	2	1	1	3	0	1	1	0	0	0	0	0	cluster1
2	3	0	2	1	0	2	0	2	1	1	1	0	3	0	1	1	0	0	0	0	0	cluster1
3	6	0	2	1	0	1	1	1	0	0	1	1	3	1	1	1	0	0	0	0	0	cluster1
4	4	0	2	1	0	3	0	2	0	2	1	0	3	1	1	1	0	0	0	0	0	cluster1
5	5	0	2	1	0	2	0	1	1	0	1	1	3	1	1	1	0	0	0	0	0	cluster1
6	3	0	2	1	0	2	1	1	0	1	1	1	3	0	1	1	0	0	0	0	0	cluster1
7	3	0	2	1	0	1	0	2	1	2	1	0	3	0	1	1	0	0	0	0	0	cluster1
8	6	0	2	1	0	3	0	1	1	1	1	0	3	1	1	1	0	0	0	0	0	cluster1
9	6	0	2	1	0	1	0	1	0	2	1	0	3	1	1	1	0	0	0	0	0	cluster1
10	6	0	0	2	1	0	2	1	0	0	1	1	0	3	0	0	0	2	1	0	0	cluster2
11	4	0	0	1	0	2	3	1	1	1	1	0	0	3	0	0	0	2	1	0	0	cluster2
12	5	0	0	2	0	3	2	1	0	2	1	0	0	3	0	0	0	2	1	0	0	cluster2
13	6	0	0	1	1	3	3	1	1	0	1	0	0	3	0	0	0	2	1	0	0	cluster2
14	3	0	0	2	1	0	2	1	0	1	1	0	0	3	0	0	0	2	1	0	0	cluster2
15	4	0	0	1	1	1	3	1	1	1	1	1	0	3	0	0	0	2	1	0	0	cluster2
16	3	0	0	1	0	1	2	1	0	0	1	0	0	3	0	0	0	2	1	0	0	cluster2
17	5	0	0	2	1	2	2	1	0	2	1	1	0	3	0	0	0	2	1	0	0	cluster2
18	6	0	0	2	0	1	3	1	1	0	1	0	0	3	0	0	0	2	1	0	0	cluster2
19	5	0	0	2	1	3	3	1	1	2	1	0	0	3	0	0	0	2	1	0	0	cluster2
20	0	1	2	0	0	1	1	1	1	1	0	0	1	1	0	1	1	0	0	3	0	cluster3
21	2	1	2	0	0	3	1	2	0	1	0	0	1	1	0	1	0	0	0	3	0	cluster3
22	2	1	2	0	0	2	1	1	0	2	0	0	1	1	0	1	1	0	0	3	0	cluster3
23	0	1	2	0	0	0	1	1	1	2	0	0	1	1	0	1	0	0	0	3	0	cluster3
24	0	1	2	0	0	2	1	1	1	1	0	0	1	1	0	1	0	0	0	3	0	cluster3
25	4	0	2	0	1	0	1	2	0	2	1	1	1	1	0	1	1	0	0	3	0	cluster3
26	2	1	2	0	0	3	1	2	0	2	0	0	1	1	0	1	1	0	0	3	0	cluster3
27	0	1	2	0	0	0	1	1	0	1	0	0	1	1	0	1	0	0	0	3	1	cluster3
28	3	0	2	0	1	3	1	2	0	1	0	1	1	1	0	1	1	0	0	3	0	cluster3
29	0	1	2	0	0	1	1	2	1	2	0	0	1	1	0	1	0	0	0	3	0	cluster3
30	2	1	2	1	1	3	1	2	1	2	1	0	2	2	0	1	0	0	0	3	1	cluster4
31	0	1	1	1	0	1	1	1	0	0	1	0	1	2	0	0	0	0	0	3	1	cluster4
32	3	1	2	0	0	1	1	2	1	0	1	0	2	2	0	0	0	0	0	3	1	cluster4
33	2	1	2	1	1	1	1	2	0	2	1	0	1	2	0	1	0	0	0	3	1	cluster4
34	1	1	2	0	0	3	1	1	1	2	1	0	2	2	0	0	0	0	0	3	1	cluster4
35	1	1	2	1	0	0	1	2	1	1	1	0	2	2	0	0	0	0	0	3	1	cluster4
36	0	1	2	1	0	3	1	1	0	0	1	0	1	2	0	0	0	0	0	3	1	cluster4
37	2	1	2	0	0	1	1	2	0	0	1	0	1	2	0	0	0	0	0	3	1	cluster4
38	3	1	2	0	0	2	1	2	1	1	1	0	2	2	0	0	0	0	0	3	1	cluster4
39	3	1	1	0	0	2	1	2	1	2	1	0	2	2	0	0	0	0	0	3	1	cluster4
40	0	1	2	1	1	1	1	1	0	0	1	0	1	2	0	1	0	0	0	3	1	cluster4
41	1	1	2	1	1	3	1	2	0	1	1	1	1	2	0	1	0	0	0	3	1	cluster4
42	1	1	2	0	0	0	1	2	1	0	1	0	2	2	0	0	0	0	0	3	1	cluster4
43	1	1	2	1	1	2	3	1	1	1	1	0	2	2	0	1	0	0	0	3	1	cluster4
44	2	1	1	0	0	3	1	2	0	2	1	0	1	2	0	0	0	0	0	3	1	cluster4
45	0	1	1	1	1	2	1	2	1	0	1	1	2	2	0	1	0	0	0	3	1	cluster4
46	0	1	2	1	0	3	1	1	0	2	1	0	1	2	0	0	0	0	0	3	1	cluster4

In order to study the disease characteristics, reduct analysis is carried out on individual four disease clusters. Table 4 shows the MPCR variables in different clusters.

Table 4. MPCR variables in different clusters

	MPCR variables
Cluster1	v1, v5, v6, v7, v8, v9, v10, v20, v22
Cluster2	v1, v4, v5, v6, v7, v9, v10, v20
Cluster3	v1, v5, v6, v8, v9, v10, v12, v20, v25, v35
Cluster4	v1, v3, v4, v5, v6, v8, v9, v10, v21, v24

Reduct analysis on different clusters shows that it has different MPCR variables, as variables are having different values in different clusters. Variables are not common across clusters and as such some variables are playing role in one cluster and not in other cluster.

Let us consider cluster4 for illustration (Table 3). Cluster4 has 17 entities of phytophthora-rot disease. To study the disease characteristic, reduct analysis is carried out on this cluster. Reduct computation on this cluster resulted in 22 reducts of varying cardinality. MPCR set is then computed from these reducts. Removal of MPCR variables (v1, v3, v4, v5, v6, v8, v9, v10, v21, v24) (Table 4) resulted in cluster having same value for all of its instances. These remaining variables (v7 = 1, v12 = 1, v20 = 1, v22 = 2, v23 = 0, v25 = 0, v26 = 0, v27 = 0, v28 = 3, v35 = 1) are playing major role in characterizing this specific cluster. PE is calculated for these remaining variables. In Cluster4, PE for variable v7 = 1 is 13/30 (Equ. 2), as 3 entities from Cluster1 and 10 entities from Cluster3 are satisfying this condition (Table 3). Similarly PE for other variables in this cluster are v12 = 1(21/30), v20 = 1(21/30), v22 = 2(0), v23 = 0(20), v25 = 0(25), v26 = 0(20), v27 = 0(20), v28 = 3(10) and v35 = 1(1). PE for variable v22 = 2 is zero, hence variable v22 = 2 describes this cluster with no error.

Table 5. Patterns obtained for soybean disease clusters

Cluster	Pattern	PE
Cluster 1 (diaporthe-stem-canker)	stem-cankers = above-sec-nde or fruiting-bodies = present	0
Cluster 2 charcoal-rot	precip = lt-norm or stem-cankers = absent or canker-lesion = tan or int-discolor = black or sclerotia = present	0
Cluster 3 (rhizoctonia-root-rot)	canker-lesion = brown ^ temp = lt-norm	0
Cluster 4 (phytophthora-rot)	canker-lesion = dk-brown-blk	0

Let us consider another example of Cluster3 (Table 3) which has ten entities corresponding to disease rhizoctonia-root-rot. After the removal of MPCR variables (Table 4) (v1, v5, v6, v8, v9, v10, v12, v20, v25, v35) from this cluster, remaining variables (v3, v4, v7, v21, v22, v23, v24, v26, v27 and v28) are having same value for all of its instances. PE for these remaining variables are v3 = 2(23/37), v4 = 0(7/37), v7 = 1(19/37), v21 = 1(8/37), v22 = 1(6/37), v23 = 0(27/37), v24 = 1(16/37), v26 = 0(27/37), v27 = 0(27/37) and v28 = 3(17/37). There is no variable with PE zero, therefore as per proposed approach combine together the variables with less PE, v22 with PE 6/37and v4 with PE 7/37. Description P: v22 = 1 and v4 = 0 describes this cluster with zero error. Similarly for Cluster2 variables v3 = 0, v21 = 0, v22 = 3, v26 = 2 and v27 = 1 have zero PE, hence any of these variables can describe the cluster completely. Cluster1 have variables v21 = 3 and v23 = 1 with zero PE, hence either of these variables can describe the cluster without error. Results of cluster description on soybean disease clusters are summarizes in Table 5 (combining together name of the variables from Table 2):

5.2 Zoo Dataset

Zoo dataset consist of 101 instances of animals with 17 variables and 7 output classes (UCI). There are 15 boolean attributes, with value one and zero corresponding to the presence and absence of hair, feathers, eggs, milk, backbone, fins, tail, airborne, aquatic, predator, toothed, breathes, venomous, domestic and catsize. The attribute number of legs {0, 2, 4, 5, 6, 8}correspond to character variable. Variables animal name and class are not considered for clustering.

EM clustering algorithm learnt four clusters from the data instead of seven classes that is known in the dataset. Table 6 shows EM clustering results on Zoo dataset. Previous studies on clustering for zoo dataset and cluster validity indices also indicated better partitioning at two, four and seven clusters (Mitra *et al.* (2002)).

Table 6. EM clustering results on zoo dataset

Cluster Name	Cluster 0	Cluster 1	Cluster 2	Cluster 3
No. of objects	21	40	20	20

Unsupervised reduct is computed for individual clusters and then MPCR is computed from them. Table 7 shows MPCR variables in different clusters. Table 8 shows the results of cluster description for animal clusters.

Table 7. MPCR variables in individual clusters

Cluster	Reduct
Cluster 0	hair, airborne, predator, toothed, venomous, legs, domestic, backbone, breathes
Cluster 1	eggs, airborne, aquatic, predator, toothed, legs, tail, domestic, catsize
Cluster 2	airborne, aquatic, predator, domestic, catsize
Cluster 3	eggs, milk, aquatic, predator, breathes, venomous, legs, domestic, catsize

Table 8. Cluster description for animal clusters

Cluster	Number of elements in Cluster	Pattern	PE
Cluster 0	20	tail = 0 ^ milk = 0	0
Cluster 1	40	milk = 1 ^ hair = 1	0
Cluster 2	20	feathers = 1	0
Cluster 3	20	fins = 1	0.024

6. CONCLUSION

Clustering provides unsupervised grouping of objects in the form of clusters which needs to be analyzed and understood. In this paper, we presented reduct driven approach for selection of significant variables from individual clusters. Ranking of significant variables on precision error resulted in formulation of meaningful and concise cluster pattern. With the application of proposed approach on soybean and zoo datasets, it is observed that obtained patterns distinctively described the clusters with no or minimum errors. In future, RCD approach will be experimented with other datasets from different domains to study the effectiveness of this approach in generating cluster pattern.

ACKNOWLEDGEMENT

Authors are grateful to the anonymous referee for giving valuable suggestions that helped in significantly improving the quality of the paper.

REFERENCES

Arora, A., Upadhyaya, S. and Jain, R. (2007). Rough set approach for generating cluster description. *Proc. of the Information Systems, Technology and Management (ICISTM-2007)*, IMT Ghaziabad, ISBN: 81-8424-182-8, Allied Publishers Pvt. Ltd, 304-310.

Abidi, S.S.R., Hoe, K.M. and Goh, A. (2001). Analyzing data clusters: A rough set approach to extract cluster defining symbolic rules. *Proc. Fisher, Hand, Hoffman, Adams (eds.) Lecture Notes in Computer Science: Advances in Intelligent Data Analysis*, 4th Intl. Symposium, IDA-01. Springer Verlag, Berlin.

Abidi, S.S.R. and Goh, A. (1998). Applying knowledge discovery to predict infectious disease epidemics. *Proc. H. Lee and H. Motoda (eds.) Lecture notes in Artificial Intelligence 1531-PRICAI'98: Topics in Artificial Intelligence*, A Springer Verlag, Berlin.

Ganter, B. and Wille, R. (1997). *Formal Concept Analysis: Mathematical Foundations*. Springer- Verlag, New York Inc., Secaucus, New York.

Han, J. and Kamber, M. (2001). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Jain, A.K., Murty, M.N. and Flynn, P.J. (1999). Data clustering: A review. *ACM Computing Surveys*, **31(3)**, 264-323.

Jain, R. (2004). Rough Set based Decision Tree Induction for Data Mining. *Ph.D. Thesis*, JNU, New Delhi.

Komorowski, J., Pawlak, Z. and Polkowski, S. (1999). Rough sets: A tutorial. *In: S. K. Pal, A. Skowron (ed.). Rough Fuzzy Hybridization: A New Trend in Decision-Making*, Springer-Verlag, Berlin, 3–98.

Mirkin, B. (1999). Concept learning and feature selection based on square-error clustering. *Machine Learning*, **35**, 25-40.

Mirkin, B. (2005). *Clustering for Data Mining: Data Recovery Approach*. Chapman and Hall.

Mitra S., Pal, S.K., Mitra, P. (2002). Data Mining in Soft Computing Framework : A Survey, *IEEE Transactions on Neural Networks*, **13(1)**, 3-14.

Pawlak, Z. (1991). *Rough Sets: Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers.

RSES: Rough Set Exploring System, available at: http://logic.mimuw.edu.pl/~rses.

Talavera, L. (1999). Feature selection as retrospective pruning in hierarchical clustering. *Proc. Third International Symposium on Intelligent Data Analysis*, IDA99 Amsterdam, Springer Verlag, The Netherlands.

UCI: Repository of Databases for Machine Learning and Data Mining, Urvine, UCI.

WEKA: A machine learning software available at: http://www.cs.waikato.ac.nz/~ml.

Available online at www.isas.org.in
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(2) 2009 189-197

Discretization based Support Vector Machines for Classification

Anshu Bharadwaj[1*] and Sonajharia Minz[2]
[1]*Indian Agricultural Statistics Research Institute, New Delhi*
[2]*Jawaharlal Nehru University, New Delhi*

(Received: December 2008, Revised: August 2009, Accepted: August 2009)

SUMMARY

Discrete values have important roles in data mining and knowledge discovery. They are about intervals of numbers which are more concise to represent and specify, easier to use and comprehend as they are closer to the knowledge level representation than continuous ones. Discretization is the process of quantizing continuous attributes. It has been used for decision tree classifier. The success of discretization can significantly extend the borders of many learning algorithms. Support Vector Machines (SVM) are the new generation learning system based on the latest advances in statistical learning theory. SVM is the recent addition to the toolbox of data mining practitioners and are gaining popularity due to many attractive features, and promising empirical performance. In this paper, a new approach to classify data using SVM classifier, after discretization is looked into. The classification results achieved after discretization based SVM are much better than the classification results using simple SVM in terms of accuracy. To acquire the better accuracy, discretization has been instrumental This is an attempt to extend the boundaries of discretization and to evaluate its effect on other machine learning techniques for classification namely, support vector machines.

Keywords: Support vector machines, Discretization, Radial basis function, Confusion matrix, Boolean reasoning based method, Entropy based method.

1. INTRODUCTION

Support vector machine (SVM) is a novel learning method based on statistical learning theory. SVM is a powerful tool for solving classification problems with small samples, nonlinearities and local minima, and is of excellent performance. To address the discretization process of continuous-valued features in an efficient and proper manner has always been an important issue for any machine learning technique. SVM is a widely used method for classification in variety of applications. The results of the experiment conducted in this study clearly show that the classification results using SVM are better when discretization process is undertaken before the classification. However, various methods of discretization affect the classification accuracy. Therefore, it is important to decide a method to improve the performance of the SVM model. The points in the dataset that fall on the bounding planes of the hyperplane in a SVM are called support vectors. They play an important role in the theory as well as in the classification task at the prediction stage. Vapnik (1974, 1979, 1998) has shown that if the training vectors are separated without errors by an optimal hyperplane, the expected error rate on a test sample is bounded by the ratio of the expectation of the support vectors to the number of training vectors. Since this ratio is independent of the dimension of the problem, and, if one can find a good set of support vectors, good generalization is guaranteed. We aim at a good generalization from the classification task that we have carried out using SVM after discretization. Even though

[*] *Corresponding author* : Anshu Bharadwaj
E-mail addresses : anshu@iasri.res.in, ans_dix@yahoo.com

SVMs can handle continuous attributes, its performance can be significantly improved by replacing a continuous attribute with its discretized values. Data discretization is defined as a process of concerting continuous data attribute values into a finite set of intervals and associating with each interval some specific data value. There are no restrictions on discrete values associated with a given data interval except that these values must induce some ordering on the discretized attribute domain. Discretization significantly improves the quality of discovered knowledge (Catlett 1991), (Pfahringer 1995) and also reduces the running time of various data mining tasks such as association rule discovery, classification, and prediction. In this study, we have also used two spatial datasets. These datasets have been used to examine the performance of the classification technique used for classical data mining task on it. Spatial datasets differ from non-spatial datasets as they have spatial aspects involved in them. Here the spatial datasets used are in the vector format. The spatial attributes in the spatial datasets used, are latitudes and longitudes. The datasets have been considered just to experiment with it using discretization based SVM classifier. In this paper, we describe discretization methods and compare them according to accuracy of the classification results. We focus our work to find out the significance of discretization before classification using SVM.

Section 2 of this paper gives the overview about the data preprocessing step of data mining along with the need of discretization and detailed description of the applied discretization methods. Section 3 deals with the basic concepts of support vector machines and its parameters in detail. Section 4 describes the confusion matrix as the performance evaluation measure for the classifier. Section 5 gives the detail about the experimental setup, summary of the data used and its analysis. Section 6 contains the results and Section 7 draws the conclusions.

2. DATA PREPROCESSING

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user. It is the most critical step in data mining process that includes the preparation and transformation of the initial dataset. Raw data are seldom used for data mining. Many transformations may be needed to produce more useful features for selected data mining methods such as prediction or classification. Discretization of numerical attributes is one of the important data preprocessing techniques. In this paper we have discretized the data before classifying it using SVM, as the preprocessing step.

2.1 Why Discretization?

There are many advantages of using discrete values over continuous one. Discrete features are closer to knowledge level representation (Simon 1981) than continuous ones. Data is reduced and simplified using discretization. For both users and experts, discrete features are easier to understand, use and explain. As reported by Doughterty *et al.* (1995), discretization makes learning more accurate and faster. In general, obtained results using discrete features are usually more compact, shorter and more accurate than using continuous ones; hence the results can be more closely examined, compared, used and reused. In addition to the many advantages of having discrete data over continuous one, a suite of classification learning algorithms can only deal with discrete data.

2.2 Discretization Methods

A large number of machine learning and statistical techniques can only be applied to datasets composed entirely of nominal variables. However, a very large proportion of real datasets include continuous variables, that is variables measured at intervals or ratio level. One solution to this problem is to partition numeric variables into sub-ranges and treat each such sub-range as a category. This process of partitioning continuous variables into categories is usually termed as discretization. Transformation of a continuous attribute to a categorical attribute involves two subtasks, deciding how many categories to have and determining how to map the values of the continuous attribute. They are then divided into n intervals specifying n – 1 split points. In the second, rather trivial step, all the values in one interval are mapped to the same categorical value. Therefore, the problem of discretization is one of deciding how many split points to choose and where to place them. The result can be represented either as a set of intervals $\{(x_0, x_1), (x_1, x_2),..., (x_{n-1}, x_n)\}$, where

x_0 and x_n may be + ∞ or – ∞ respectively, or equivalently, as a series of inequalities $x_0 < x \leq x_1,..., x_{n-1} < x < x_n$. A variety of discretization methods have been developed along different lines due to different needs: supervised vs. unsupervised; dynamic vs. static; global vs. local; splitting (top-down) vs. merging (bottom-up), and direct vs. incremental.

2.2.1 Supervised and unsupervised discretization methods

Data can be supervised or unsupervised depending on whether it has class information. Likewise, supervised discretization considers class information while unsupervised discretization does not; unsupervised discretization is seen in earlier methods like equal-width and equal-frequency. In unsupervised methods, continuous ranges are divided into sub ranges by the user specified width (range of values) or frequency (number of instances in each interval). This may not give good results in cases where the distribution of the continuous values is not uniform. Furthermore, it is vulnerable to outliers as they affect the ranges significantly (Catlett 1991). To overcome this shortcoming, supervised discretization methods were introduced and class information is used to find the proper intervals by cut-points. In this study, we have used both the unsupervised and supervised methods of discretization to discretize the datasets before applying SVM. We have selected two of the most popular and widely used methods of supervised discretization and similarly one unsupervised method of discretization is also selected. The supervised discretization methods used are described briefly for a better understanding of the methods.

The methods used are

1. Unsupervised : Equal-frequency
2. Supervised: Entropy based and Boolean reasoning based methods

2.2.1.1 *Entropy-based discretization method*

Entropy based discretization method uses a minimal entropy heuristic for discretization of continuous attributes. This method tries to find a binary cut for each attribute. Following a method introduced by Fayyad and Irani (1993), the minimal entropy criteria can also be used to find multi-level cuts for each attribute. The algorithm uses the class information entropy of candidate partitions to select binary boundaries for discretization.

2.2.1.2 *Boolean reasoning/rough set based discretization method*

The method that we have discussed (entropy based) discretize only one attribute at a time. It may therefore introduce more cuts than is absolutely necessary for discerning between the decision classes. Nguyen and Nguyen (1996), and Nguyen and Skowron (1995, 1997) have introduced a supervised method that considers all of the attributes simultaneously and creates consequently fewer cuts. Their method is developed with basis in rough sets methods and Boolean reasoning.

3. SUPPORT VECTOR MACHINE

The foundations of SVM based on statistical learning theory were developed by Vapnik (1998) and Burges (1998) to solve the classification problem. The SVM is the recent addition to the toolbox of data mining practitioners and are gaining popularity due to many attractive features, and promising empirical performance. They are a new generation learning system based on the latest advances in statistical learning theory. The formulation embodies the Structural Risk Minimization (SRM) principle, which has been shown to be superior (Gunn *et al.* 1997), to traditional Empirical Risk Minimization (ERM) principle, employed by conventional neural networks. SRM minimizes an upper bound on the expected risk, as opposed to ERM that minimizes the error on the training data. It is this difference which equips SVM with a greater ability to generalize, which is the goal in statistical learning. SVM belongs to the class of supervised learning algorithms in which the learning machine is given a set of examples (or inputs) with the associated labels (or output values). Like in decision trees, the examples are in the form of attribute vectors, so that the input space is a subset of R^n. SVMs create a hyperplane that separates two classes (this can be extended to multi class problems). While doing so, SVM algorithm tries to achieve maximum separation between the classes. Separating the classes with a large margin minimizes a bound on the expected generalization error. By "minimum generalization error", it means that when new examples (data points with unknown class values) arrive for classification, the

chance of making error in the prediction (of the class to which it belongs) based on the learned classifier (hyperplane) should be minimum. Intuitively, such a classifier is one which achieves maximum separation-margin between the classes. The two planes parallel to the plane are called bounding planes. The distance between these bounding planes is called margin and by SVM "learning", i.e. finding hyperplane which maximizes this margin. The points (in the dataset) falling on the bounding planes are called the support vectors. "Machine" in Support Vector Machine is nothing but the algorithm (Soman *et al.* 2006). SVM was designed initially as binary classifier i.e. it classifies the data into two classes but researchers have extended its boundaries to be a multi-class classifier. SVM was first introduced as a training algorithm (Boser *et al.* 1992) that automatically tunes the capacity of the classification function maximizing the margin between the training patterns and the decision boundary (Cristianini and Shaw-Taylor 2000). This algorithm operates with large class of decision functions that are linear in their parameters but not restricted to linear dependences in the input components. For the computational considerations, SVM works well on two important practical considerations of classification algorithms i.e. speed and convergence.

3.1 SVM and its Parameter

To construct an optimal hyperplane, SVM employees an iterative training algorithm, which is used to minimize an error function. According to the form of the error function, SVM models can be classified into two distinct groups

1. SVM for classification
2. SVM for regression

In this study we are dealing with classification problem, so the SVM for classification is described here.

For SVM, training involves the minimization of the error function

$$\frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i$$

subject to the constraints

$$y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0, \quad i = 1, \ldots, N$$

where C is the capacity constant or the model complexity, w is the vector of coefficients, b a constant and ξ_i are parameters for handling non-separable data (inputs). The index i labels the N training cases. Note that $y \in \pm 1$ is the class label and x_i is the independent variable. The kernel ϕ is used to transform data from the input (independent) to the feature space. It should be noted that larger the C, the more the error is penalized. Thus, C should be chosen with care to avoid over fitting.

3.2 Radial Basis Function

There are a number of kernels that can be used for support vector machine models. These include Linear, Polynomial, Radial Basis and Sigmoid.

A radial basis function (RBF) is a real-valued function whose value depends only on the distance from the origin, so that $\phi(x) = \phi(\|x\|)$; or alternatively on the distance from some other point c, called a *center*, so that $\phi(x, c) = \phi(\|x - c\|)$. Any function ϕ that satisfies the property $\phi(x) = \phi(\|x\|)$ is a radial function. The norm is usually to use RBF, although other distance functions are also possible. The following expression describes the RBF kernel for SVM

$$\phi = \exp\{-\gamma |x - c|^2\}, \text{ where } \gamma > 0$$

where γ is called the RBF kernel parameter. The RBF kernel is the most popular kernel type due to its localized and finite response across the entire range of real x-axis.

4. PERFORMANCE EVALUATION MEASURE: CONFUSION MATRIX

Evaluation of the performance of the classification model is based on the counts of the test records correctly or incorrectly predicted by the model. These counts are tabulated in a table called Confusion Matrix. Table 1 depicts the confusion matrix for a binary classification model. Each entry f_{ij} in this table denotes the number of records from class i predicted to be of class j. For instance f_{01} is the number of records from

Table 1. Confusion Matrix

		Predicted Class	
		Class = 1	Class = 0
Actual Class	Class = 1	f_{11}	f_{10}
	Class = 0	f_{01}	f_{00}

class 0 incorrectly predicted as of class 1. Based on the entries in the table the total number of correct prediction made by the model is (f_{11}+ f_{00}) and the total number of incorrect predictions is (f_{10} + f_{01}).

5. EXPERIMENT AND ANALYSIS

Using the discretization methods before applying SVM, we clearly see that discretization simplifies data (continuous values are quantized into intervals) without sacrificing data consistency much (only a few inconsistencies occur after discretization). We have to evaluate the ultimate objective of discretization of the datasets before applying SVM—whether discretization helps improve the performance of learning and understanding of learning results. The kernel used for training is RBF. The improvement is measured in terms of the classification accuracy. The evaluation of the performance of the classification model is done using Confusion Matrix. As a general approach of solving classification problems, each dataset is split into two datasets training sample dataset and test sample dataset. Training dataset consists of the records having class labels and is used to build the classification model whereas the test dataset contains records without class labels and is used to validate the model, built by training dataset. Though discretization is usually a needless preprocess step for SVM, which can deal with continuous and hybrid attributes directly, it has been still attractive to use discretized datasets because it has improved the classification performance and reduced the training time.

5.1 Data Description

Four datasets are selected from various sources, with all numeric features and varying data sizes. The datasets used in the study are Boston2, CIMMYT and Hurricane. Boston2 and Hurricane datasets are from public domain i.e. UCI repository available online and the CIMMYT dataset is a live dataset. The live dataset used for this comparative study is Rice dataset. This dataset is in vector data format of spatial databases. Spatial attributes in the datasets are latitudes and longitudes. The data is obtained from Resource Conservation Technologies from Rice-Wheat Consortium, CIMMYT, India. Here only a small part of data with 50 observations has been used for illustration purpose. There are 4 classes in which the data has to be classified. Number of attributes in the dataset is 10 that includes the latitudes and longitudes being spatial attributes of the dataset. The CIMMYT dataset is modified as two different datasets, first by considering all the variables (latitudes and longitudes) as CIMMYT1, and secondly by ignoring the spatial variables, i.e. dropping the variables containing the spatial information, as CIMMYT2. The results may be different and the conclusions drawn here may change with the full set of data. The sample dataset is from different districts of Western Uttar Pradesh and contains different treatments (i.e. different types of seed cultivation), the spatial aspect of the location (longitudes and latitudes) with various biometrical characters of the rice plant. The task is to classify the varieties in different classes.

The second dataset is Boston2. This example illustrates an analysis of the Boston house price data (Harrison and Rubinfeld 1978) that was reported by (Lim *et al.* 1997). Median prices of housing tracts were classified as Low, Medium, or High on the dependent variable price. There was one categorical predictor, Cat1, and 12 ordered predictors, Ord1 through Ord12. The complete data set contains a total of 1012 cases.

The third data used in this study is the Hurricane data. This data was originally obtained from Atlantic tropical cyclone “best” track and intensity records managed by the Tropical Prediction Center (Formerly the National Hurricane Center, Jarvinen *et al.* 1984), where “best” refers to an accurate assessment of storm location based on a post analysis of available data. The dataset extends back to 1886 and includes all tropical cyclones, that reached tropical storm strength. A storm has latitude and longitude coordinates and maximum sustained winds every 6 hours during the storm’s existence. Data are most reliable after 1944 when the US Air force began aircraft reconnaissance missions to investigate individual storm. The dataset has six independent and one dependent variable. Each storm contains the Julian day, the latitude and the longitude for initial depression and initial hurricane stages, that is the day and position for which the storm was first reported as a tropical depression and a hurricane, respectively. Day D and day H are the Julian days on which the storm first reached depression and hurricane strengths respectively, long D and long H are the initial depression and hurricane longitudes, respectively, lat D and long H are the initial depression and hurricane

latitudes, respectively; TROP and BARO are tropical only and baroclinically influenced hurricanes, respectively. Summary of datasets can be found in Table 2.

Table 2. Summary of datasets

S.No.	Dataset	Total number of instances	Number of features	Number of classes	source of data
1.	CIMMYT	50	10	4	CIMMYT, INDIA
2.	CIMMYT1	50	8	4	CIMMYT, INDIA
3.	BOSTON2	1012	13	3	STATISTICA
4.	HURRICANE	209	6	2	STATISTICA

5.2 Experimental Set-up

A total of three discretization methods (equal-frequency (unsupervised), entropy and Boolean reasoning (supervised)), have been used to study the effect of discretization on classification results. Experimental design is given in Table 3.

The datasets are split into train and test datasets, then the discretization algorithms (entropy based, Boolean reasoning and equal-frequency) are used to discretize the train dataset one by one. Once the train dataset is discretized using any of the algorithms, the same cuts points (Liu *et al.* 2002) or intervals generated for the train dataset using the particular discretization algorithm are saved in a file and the same cuts points are then used to discretize the test dataset, for test dataset the class labels are not used during discretization. Once the data has been split (into train and test datasets) and discretized, the original dataset (i.e. the undiscretized data) has not been used anywhere in the study. The experiment was conducted with 8 runs each for each dataset. Each run means, to classify the data at split of different seed value. Seed values used for splits are 1000, 900, 800, 750, 600, 500, 350, 100. The seed values were randomly selected. Classification using SVM was carried out on the discretized datasets so that the results can be compared and the effect of the discretization on SVM can be studied. CIMMYT and Hurricane datasets are spatial datasets in vector format with latitude and longitude as spatial attributes.

Table 3. Experimental design

S.No.	Experimental Steps
1.	Split each dataset into Training (70%) and Test Sample (30%) datasets of each complete data. Split of the dataset is carried out using simple random sampling
2.	Discretize the train and test datasets separately. (Use all the three methods, i.e. equal- frequency, entropy and boolean reasoning for discretizing the train dataset and then use the same cuts to discretize the test dataset without using class labels, for all datasets)
3.	Apply SVM for classification on the datasets (both train and test separately) using 10x10 fold cross-validation.
4.	Compare the classification results with the SVM classification results without discretization. Classification accuracy of train and test datasets is compared separately.

Table 4. Hurricane training data sample after discretization

Attribute	Continuous values	Intervals after discretization
DAYDEPR	224, 239, 285, 231, 266, 257, 237, 243, 245, 364	[*, 270), [279, 287), [288, *)
LONDEPR	45.7, 25.6, 78.2, 19, 62.2, 61.7, 74.9, 67.7, 56.4, 50.9	[*, 58.2), [78.1, 80.6), [62.1, 62.7), [60, 61.8), [70.4, 77.3), [66.8, 67.8)
LATDEPR	12.2, 12.3, 14.3, 14.6, 14.4, 16, 24, 19.3, 11.2, 22.1	[*, 14.2), [14.4, 14.9), [15.4, 16.5), [22.7, 24.1), [19.2, 19.8), [21.5, 22.3)
DAYHUR	228, 245, 287, 240, 269, 258, 239, 244, 250, 365	[*, 270), [282, 290), (292, *)
LONHURR	62.5, 58.8, 82.1, 64.7, 73.8, 67, 76.2, 69.5, 70.8, 55.2	[62.3, 62.9), [58.5, 59.3), [80.5, 82.2), [62.9, 65.0), [73.7, 74.0), [66.7, 67.5), [75.5, 76.4), [69.2, 69.7), [70.4, 70.9), [54.6, 55.3)
LATHURR	15.4, 14.1, 18.5, 21.9, 24.5, 20.9, 28.9, 24.8, 22.2, 20.6	[*, 24.7), [28.9, 29.0),[24.7, 25.2)

A sample (10 data points) of Hurricane dataset after discretization is given in Table 4 for better understanding. The table shows the discretized training data sample using Entropy method, it includes original continuous values and the intervals into which the data has been divided after discretization.

6. RESULTS AND DISCUSSION

The results are shown in Table 5. Each result consists of the classification accuracy of the SVM learning technique with and without discretization of the datasets.

SVM classification using discretization shows that the results obtained are improved and better classification accuracy is attained. The parameter of SVM decision function i.e. capacity or model complexity does not get affected by discretization as discretization process works on the dataset rather than the model. Similarly, the parameter of the RBF kernel i.e. γ also remain unaffected by the discretization of the datasets before applying SVM.

It is also observed from the results given in the above table that the supervised discretization algorithms are better than the unsupervised discretization algorithm as the classification accuracies using the supervised discretization algorithms are better than the unsupervised discretization algorithm. Out of the supervised discretization algorithms, Boolean reasoning based algorithm is performing better in attaining better classification accuracy. It is known that supervised discretization is better than the unsupervised discretization but we have used one method of unsupervised discretization to compare the difference it brings to the classification accuracy if the data is classified after unsupervised discretization as compared to the supervised discretization. It is observed that for one of the datasets, Boston2, the classification accuracy attained after discretization using unsupervised method (equal frequency) is higher than the classification accuracy attained after supervised classification using entropy based method. Although this accuracy is less than the accuracy attained using the other supervised discretization method i.e. Boolean Reasoning.

Discretization yields the reduction in unique tuples by assigning the discretized value of the attribute to the objects whose numeric value lies in the corresponding discrete interval. Thus, we could observe that there had been a reduction in the number of support vectors per class during classification of the discretized dataset. The number of support vectors was reduced to give better classification accuracy.

Table 5. Results of classification using SVM

Dataset	SVM with discretization						Without discretization	
	Entropy		Boolean reasoning		Equal frequency		Original	
	Train	Test	Train	Test	Train	Test	Train	Test
Boston2	90.25	89.27	98.16	98.94	94.18	94.86	79.85	79.01
Hurricane	91.37	88.88	97.26	85.18	88.12	88.88	89.72	87.93
CIMMYT	84.57	80.00	92.14	53.15	76.57	69.00	61.85	76.85
CIMMYT1	78.23	56.89	85.71	60.00	68.73	63.00	57.33	74.00

Table 6. Comparison of classifiers in terms of classification accuracy

Dataset	ANN		SVM		Discretization based SVM					
					Entropy		Boolean reasoning		Equal frequency	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Boston2	75.43	78.78	79.85	79.01	90.25	89.27	98.16	98.94	94.18	94.86
Hurricane	80.91	83.89	89.72	87.93	91.37	88.88	97.26	85.18	88.12	88.88
CIMMYT	36.13	55.44	61.85	76.85	84.57	80.00	92.14	53.15	76.57	69.00
CIMMYT 1	35.65	42.06	57.33	74.00	78.23	56.89	85.71	60.00	68.73	63.00

Hurricane dataset has earlier been classified using a method explained in (Elsner *et al.* 1996). The method used is Partially Adaptive Classification Trees (PACT) algorithm (Shih 1993) based on linear discriminant analysis (Mardia and Bibby 1979) and tree structured classification method (Brieman *et al.* 1984). This algorithm gives a classification accuracy of around 90% which is less than the accuracy attained by discretization based SVM classification. The classification accuracy attained by supervised discretization method based SVM for hurricane dataset are 91.37 and 97.14 respectively for entropy based method and Boolean reasoning based method. Similarly in Minz and Dixit (2007), these four datasets have been classified using Artificial Neural Network and SVM and it is seen that result obtained by using discretization based SVM are much better than the results obtained by the earlier two methods. Comparative results are shown in Table 6.

7. CONCLUSION

The study was undertaken with an aim to explore the effects of discretization on support vector machines. Although data discretization has been a step for applying machine learning technique of classification such as decision tree but it has not been tried for support vector machines classifier, the reason being its ability to handle continuous and hybrid data unlike the decision tree algorithm ID3, which can handle only discrete datasets for classification. Therefore, we tried to explore the effect of discretization of the datasets before applying SVM classifier. This was done with the aim of attaining better classification accuracy without disturbing or distorting the parameters (C and Gamma) of SVM. The results clearly indicate that the accuracies of discretization based SVM are better as compared to the classification accuracy without SVM of the same datasets when they were classified without getting discretized. We have also observed that the supervised discretization algorithm works better than the unsupervised discretization algorithm. Among supervised discretization algorithm also, Boolean Reasoning based method performs best. This study establishes that discretization can be used for SVM classifiers also.

ACKNOWLEDGEMENT

The authors thank the referee for the critical evaluation and suggestions that led to commendable improvement in this paper. The authors are also thankful to CIMMYT, India, for providing the rice data for this study.

REFERENCES

Boser, B.E., Guyon, I.M. and Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. *Proc. of 5th Annual Workshop on Computer Learning Theory,* Pittsburgh, PA: ACM, 144-152.

Brieman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees.* Wadsworth, 358.

Burges, J.C. (1998). *A Tutorial on Support Vector Machines for Pattern Recognition.* Data Mining and Knowledge Discovery, **2**, 121-167.

Catlett, J. (1991). On changing continuous attributes into ordered discrete attributes. *Proc. Fifth European Working Session on Learning,* Springer-Verlag, Berlin, 164-177.

Cristianini, N. and Shaw-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press, Cambridge.

Dougherty, J., Kohavi, R. and Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. *Proc. Twelfth International Conference on Machine Learning,* Morgan Kaufmann, Los Altos, CA, 194-202.

Elsner, J.B., Lehmiller, G.S. and Kimberlain, T.B. (1996). Objective classification of atlantic hurricanes. *J. Climate,* **9**, 2880-2889.

Fayyad, U.M. and Irani, K.B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In: *Machine Learning. 13th IJCAI,* **2**, Morgan Kaufmann, Chambery, France, 1022-1027.

Gunn, S.R., Brown, M. and Bossely, K.M. (1997). Network Performance Assessment for Neurofuzzy Data Modelling. In: *Intelligent Data Analysis,* **1208**, Lecture Notes in Computer Science (X. Liu, P. Cohen and M. Berthold (ed.), 313-323.

Harrison, D. and Rubinfeld, D.L. (1978). Hedonic prices and the demand for clean air. *J. Environ. Econ. Manage.,* **5**, 81-102.

Jarvinen, B.R., Neumann, C.J. and Davis, M.A.S. (1984). A Tropical cyclone data tape for the North Atlantic Basin.

1886-1983: Contents, Limitations and Uses. *NOAA Tech. Report. NWH NHC 22*, 21.

Lim, T.S., Loh, W.Y. and Shih, Y.S. (1997). An empirical comparison of decision trees and other classification methods. *Technical Report 979, Department of Statistics, University of Wisconsin, Madison,* Wisconsin, http://www.stat.wisc.edu/~limt/compare.ps

Mardia, K.V. and Bibby, J.T.M. (1979). *Multivariate Analysis.* Hartcourt Brace and Company, 521.

Minz, S. and Dixit, A. (2007). Neural networks and support vector machines: A comparative study for spatial data classification. *Proc. of the National Conference on Methods and Models in Computing 2007, New Delhi,* December 13-14, 2007.

Nguyen, S.H. and Skowron, A. (1995). Quantization of real value attributes: Rough set and boolean reasoning approach. *Proc. Int.Workshop Rough Sets and Soft Computing* at *2nd Joint Conf. Information Sciences (JCIS'95)*, Durham, NC, 34-37.

Nguyen, S.H. and Nguyen, S.H. (1996). Some efficient methods for rough set. *Proc. of the Sixth International Conference, Information Processing and Management of Uncertainity in Knowledge-Based Systems (IPMU-96),* 2, Granada, Spain, July 1-5.

Nguyen, S.H. and Skowron, A. (1997). Quantization of real value attributes: Rough set and boolean reasoning approach. *Bulletin of International Rough Set Society,* **1(1),** 5-16.

Pfahringer, B. (1995). Compression-based discretization of continuous attributes. *Proc. Twelfth International Conference on Machine Learning,* Morgan Kaufmann, *San Francisco,* CA, 456-563.

Shih, Y.S. (1993). Tree structured classification. *Ph.D. Dissertation, University of Wisconsin-Madison.*

Simon. H.A. (1981). *The Sciences of the Artificial.* **2**, Cambridge, MA, MIT.

Soman, K.P., Diwakar, S. and Ajay, V. (2006). *Insight into Data Mining: Theory and Practice.* Prentice Hall of India Pvt. Ltd.

Vapnik, V. (1974). *Theory of Pattern Recognition.* Nauka, Moscow.

Vapnik, V. and Chervonenkis, A. (1979). *Theory of Pattern Recognition.* Nauka, Moscow.

Vapnik, V. (1998). *Statistical Learning Theory.* Wiley, New York.