Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 4459 - 4474
Received 20 January 2021; Accepted 08 February 2021.

Small Area Estimation in Health Sector by Support Vector Machine (SVM)

[1]Basavarajaiah DM, [3]Gopal Krishna Mithra CA, [1,2] Narasimhamurthy B, Veeregowda BM, [1]Mahadevappa D Gouri
[1]Karnataka Veterinary Animal and Fisheries Sciences University, Bidar [2]National Institute of Epidemiology, NIE, ICMR, Chennai, Tamilnadu [3]Departments of Paediatrics, KIMS, Hubbli, Karnataka
Corresponding author: Dr Gopal Krishna Mithra CA, KIMS, Hubbli

Abstract

TheGovernment of India has initiated many public health policies in rural and urbansetup.An impact assessmentof the programme is a prime factor to explore status of the programme. Different analytical stepswere used forassessment of various policiesat national level.Small area estimation by support vector machine isone of the advanced statisticalmethods to magnify theprogramme at national level presenting in 3D dimensionality. The method of survey conducted in small area has differing based on the objective of our interest. If the objective willfocus on narrowpoint, certain limited number of health indicators willbeselected for surveypurpose in smaller area. But, our objective of interest is ofbroader perspective lack of limitation as we will be selecting larger area for survey. Usually, the health and social survey mainly dependson rationality, economy and time duration. Byminimizing the economy andtime duration;hence we will achieve goodsuccess in researchprogramme. Without this thumb rule of research hypothesis, ourtested hypothesismay not yield good resultsfrom the experimentalarea. Keeping this in view, initiallysmall area will be selected with limited sample size, later on the research is magnified at populationlevel in different dimensions. A very limited research hypothesis was testedbased on small area estimationapproach mathematically in health sector. In this ample ofresearch gap.The present study we attempted to fill the research gap to explore new statistical methods applicable to small area estimation and we have demonstrated robustness of the model by using sensitivity analysis vetted in real health data sets.The study was conducted in selected states and Union territoriesof India. The selected area constitutes of418 districts which are administered bytheir respective State /Union territories. As per the official record, district level health planning involves 660rates with various relative comparative attributes. In each selected sites, various health domains secondary data sets were collected based on SOP.The collected data was modelled by using Mat labsoftware, the support vector machine learning (SVM) was used to estimate various health indicators and magnify the results at population level, In this study intervention,we have adopted two- stage sampling procedure for selection of 20 sub-centres in selected districts of south India, we visited all households and collected relevant information (datasets) regarding recent births and socioeconomic data from selected respondents of every married women aged between 15–49 years. SVM was adopted to optimize thecensoredsurvey data on every married woman to know the general marital fertility (GMFR) rate. As per the analysis, the median general marital fertility (GMF) rate was285.22(63.25%) [Likelihood10.52, R^2(%) =0.86, Akiake information 125.22].Our model results revealed that,the medianGeneral Marital Fertility Rate (GMFR) is simple estimation of number of rates of births (simulated selected socio-economic factors) in small area. Our demonstratedmodel willhelpus topredict various health outcomes and also supportthepolicy makers for implementation of public health policy at nationallevel to demystify the art of valid decisions. The SVMsummary willmeasure various generalized attributes and also reports small area variation inclusion with fertility rates and necessary intervention atsample level.

Key words: SFR, SVM, Small area estimation, fertility

Introduction

The support vector machines (SVM) was presented for the first time in 1992, it is an advanced technique to maximize the marginal errors that can separate two classless from different rates of research survey (small area) [1,2,4]. There are two main models used for building the SVM classifier, namely linear and nonlinear SVM models. Based on the SVM analytics, we have

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 4459 - 4474
Received 20 January 2021; Accepted 08 February 2021.

developed a new branch of non parametrical regression simulation techniques called Support Vector Regression (SVR) model [5]. Here, we have separated the hyper plane of different slopes of selected parameters and were substituted by the regression hyperplane and separations of margins of slopes, then were placed by the regression method [6]. The SVR was described with known definition, every three points on the plane aline can be drawn [7, 10, 11]. A novel type of learning machine called support vector machine (SVM), which has been receiving increasing interest in the areas ranging from its original application with different patterns recognition to other applications such as regression estimation due to its remarkable generalization performance[12,13].

Small area depends on many contexts, as it relates to the number of cases of disease/vital events/counts of a particular screening method and results of system that are observed at a particular point of time[14,18]. As a rough guide, we define any region containing fewer than 20 cases of diseases present in the selected region are known as small area [1-5]. In case of leprosy, a small area is nothing but district because of the low number of leprosy cases of l recorded, approximately20 cases in a low endemic district containing a population of around three millions[11,12,19]. Another realistic example is number of cases of cancer in the selected sites denoted in small area (incidence was calculated based on the entire general population - at least 100,000 individuals) [18, 20]. Based on geographical area, the endemicity of a disease depends upon the overall incidence of the disease [19]. The concept of "small area estimation" developed in accordance with standard operating protocol (SOP); it is very useful tool for public health surveys, policy and research intervention [20, 22, and 23]. The method will allow us to compare one or more attributes oftenly *viz*., fertility, mortality, risks of cancer, surgical operation, health behaviour patterns, incidence of insulin dependent diabetes and low birth weight recorded in small areas [21-23]. In this regard, Primary Health Centre (PHC) / Sub centre (SC) / village or ward within a districts would be considered as "Small area [24, 25, 28]. The small area compilation and its relative comparison involve laborious and it will make more complexity for data handling. A simple measure of variation is necessary and explores the rates on small area estimation by advanced statistical techniques [22]. More complicated measures were estimated based on the hierarchical models kind of methods that are statistically robust in nature besides difficulty in interpreting the results by public health officials because of its complexity [29,30]. A simple measures such as external Quotient (EQ) *i.e.* the ratio of lowest to highest rate, many literature suggested that, it is a satisfying measure for small area estimation, but it cannot be recommended at all time because of lack of good statistical algebraic properties [28,30,32]. Though, the standard deviation (SD) and coefficient of variation (CV) are simple discrete measures of small area estimation, they are neither statistically feasible nor robust [30, 32]. The sophisticated techniques are very important for estimation of good source of variation *viz*., systematic variation (SV) and Support vector machine tool, these two methods will accurately estimate the parameters of small area [18,19,22].Further, it is assumed that, the marginal error differences was kept at minimal stage in selected sites, between the observed and unexpected variations aroused due to Poisson fluctuationsand lack of normalcy $X_{ij} \neq N(\mu, a^2)$with mean $\mu$and common variancea^2. Certain limitations include;(i) it has no proper methods for estimation of standard error and (ii) the estimates of SV can sometimes be negative [23,24]. To overcome all these limitations, we demonstrated and formulated SVMempirical linear model [26, 28].

A brief account of these methods and their limitations are given elsewhere [14]. Being a major scientific issue and public health policy, for example, there are20small areas producing birth and fertility rates (SFR) (mortality/fertility/morbidity). Interpreting all the 20 rates simultaneously is cumbersome [30, 31]. Therefore, health administrators need varied measures of variation to describe the rates in small areas, enable to comprehend and effective interpretation of results[14, 15, and 16]. These measures will help the public health administrator

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 4459 - 4474
Received 20 January 2021; Accepted 08 February 2021.

to identify and monitor the areas with high, medium and low rates and according to plans the accessing health care services.Recently, centiles approachare being used to derive simple measures to derive small area estimation [17, 18]. Due to paucity of literature in connection with advanced methodof small area, the present article reports new analytical model (address issues and challenges suitable public health policy in terms of practicable measures to health professionals to understand and interpret them meaningfully) [10,12,17].These measures demonstrate fertility data sets, surveyed in health units in selected districts of Tamil Nadu, south India.Public health professionals can easily understand the concept of systematic variance in small area estimation. The methodwas employed based on various componentsor attributes of small area, suggested random variation (within each area variation) of the observed rates are rallying around the true rate and systematic variation (variation among the observed rates across areas) [10, 18, 21].

Studies on systematic variation areimportant for public health[3]. We proposed a popular method for computing and identifying the factors affecting systematic variance [22, 23]. It was assumed that,observed births O_i (standardized general marital fertility rate) in selected area wouldfollow binomial distribution (BD) with mean e_i, where e_i is the age-sex standardized number of births in i^{th}selected area. In our previous analytical methods we have reviewed at greater accuracy to know the importance of systematic variations in real data sets , as per the literature projected,models are more complicated for public health care professionals demonstrated models were found to be lack of precision [25,26]. For example, it was reported that, the systematic variationof births in 20 sub centres of selected sites, in Tamil Nadu was (0.023). This value has no public health importance andpublic health professionals rather ignore the size of small area estimation and concentrate on statistical significance. Usually, this driven approach is relatively compared to the test statistic χ^2distribution,the resultshowedinsignificance of estimation of small area. Other suggested method SV depends on significance testing[16, 17].However, in recent years and coinciding with increasing availability of computers, statistical software packages and programmable pocket calculators, there has been an upsurge of significance testing, sometimes bordering on the indiscriminate [18,19]. This unfortunate development led the authorities to lament regarding the excessive use of hypothesis testing at the expense of other ways of assessing results [12,13].A serious limitation of basing interpretation entirely on the basis of findings of a significance test arises from the naïve, but often held view by many non-statisticians that, a significant results proved the existence of a real and important variation and that conversely, we obtain significant results demonstrated(no real variation was found) [14,15]. To overcome this*emphasis* on significance testing of small area estimation,many scientists employed machine learning methods to derive the estimation of small area [16, 18, 22].In this paper simple measures are suggested using, first, rates across small areas and second, SV to derive three statistics: i) the median rate across small areas; ii) standardised fertility ratio (SFR); and iii) the number and percentage of births that could be averted, if all small areas had the same rate as areas with median SFR[23,24,25]. Median rate has good statistical properties: first, doing hypothesis testing or constructing confidence intervals, is possible; second, the median cannot be zero unless more than half the areas have 'nil' rates. The drawback is that it does not take care of small area variations [25, 26,28]. Therefore, wesuggested SVM to derive SFRs. These SFRs are similar to standardized mortality ratios (SMRs) having all the above statistical properties [21, 24,29]. More importantly, no assumption was made on the distribution of fertility events to derive the above statistics [31,32,33]. These measures have good statistical properties [18, 19, 22]. The point estimates and their 95% confidence intervals can easily be computed even if rates in some areas are zero, and between and within variation among the rates is considered [21, 24, 29]. The suggested methodology was applied to the fertility data collected in a well-planned survey conducted in a health unit of south India [25,27,30].SVM model assessed the quality of conceptual cost estimates slightly more accurately than the discriminant analysis model[33].

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 4459 - 4474
Received 20 January 2021; Accepted 08 February 2021.

Methods

The fertility data sets of married women was extracted through a well-planned sample survey conducted in one health unit selected district (HUD) of Tamil Nadu State. A total of twenty sub centres were randomly chosen for the study based on two-stage sampling design. In the first stage, ten public health care centres (PHCs) were selected by probability proportional to size (PPS) technique, later two sub centres were randomly chosen from each selected public health care centres (PHCs) .A well-trained graduate field investigators conducted the survey, interview was done after obtaining written consent, all participants were interviewed separately (married women aged 15-49 years). From each of the 20 selected sub centres we collected data with respect to fertility based on structured pre-tested questionnaires. The questionnaire tools included information pertaining to the outcome of recent pregnancies, socio-economic factors of the households and the individual woman (Kuppuswamy scale 2020). Similar survey was conducted in the same 20 sub centres of the health unit district (HUD), simultaneously we collected relevant information on fertility profile of women for the year 1994 (extracted from census data). Quality control (QC) was cross checked to know the clarity of collected data.

The fertility data on two successive years 1993 and 1994 was pooled in order to eliminate the spillover effect and comprehend to increase precision of survey. We incept the simple measures for reporting small area variations in selected area with different traits *viz*., socio-economic parameters, General Marital Fertility Rates (GMFRs) and different biological parameters. The GMFR each of the 20 sub centres were indirectly standardized with socio-economic factors to explore age-specific fertility rates of SOP (5-year age-groups were used).The population data were extracted from census data of Tamil Nadu state, the observed number of currently married women and survey data was relatively compared with SVM mixed effect model. An observed births and standardized GMFR of sub centres were collected from sample survey. The analysis was limited to standardized GMFR alone because i) the selected socio-economic factors may not affect fertility before 15 years, ii) the standardized GMFR could be regarded as 'reproductive index' in the population. Our aim of small area variation analysis is to develop index for generating various hypotheses that may lead to future studies to identify different pathways for preventing unwanted pregnancies; the concept which has adjudicated in the public health programme and simplify the art of decision making for public health administrators. The median and its 95% confidence interval of the standardized GMFR in each selected socio- economic factors were computed. The number of births averted was computed under the hypothesis that all the sub centres had the median standardized GMFR. For example, the standardized GMFRs for illiterate women vary from general population, and the median standardized GMFR was 99.3 per 1000. Ten subcenter had higher standardized GMFR than the median. The formula for number of births averted in an i^{th} sub center, if all the areas had the median standardized GMFR is b_i = (Observed standardized $GMFR_i$- median standardized $GMFR_i$) number of women in i^{th} subcenter. The total number of births that could be averted by all sub centers having higher standardized GMFR than median = Σb_i. Similar methodology was adopted for the number of births that could be averted among illiterate women, if all the areas had the median standardized SFR. This was done for each of the selected socio- economic factors and the percentage of births that could be adjusted in each factor. Suppose o_i and e_i are observed and expected number of births, respectively in the sub-centre for illiterate women then $\sum_{i}^{K} \frac{(o_i - e_i)^2}{}$ follows a $\chi^2_{(k-1)\,df}$ [21]. The expected number of births in each sub-

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 4459 - 4474
Received 20 January 2021; Accepted 08 February 2021.

$_{i=1}$ e_i

centre was obtained on the basis of age-specific marital fertility rates applied to the age- distribution of the married females in that area. Observed variance was calculated by dividing the χ^2 value by the number of births studied in each factors. Under the Null Hypothesis (H_0) all the variation occurs due to random variance and the expected value of χ^2(*i.e.,* k-1). The random component of the variance = (k-1)/n.SV = χ^2/n - (k-1)/n. We estimated SV in each selected factors or variables, using nonparametric bootstrap, Poisson regression model was carried out based on observed births as the dependent variable and the expected number of births as the independent variable. The deviance was provided from χ^2 value to test of goodness of fit of the above model. This χ^2 value was used for computing the systematic variance. Non-parametric random sample of 'k' pairs from the observed sample with replacement was drawn. Using Poisson regression we have obtained the deviance of χ^2 and estimated the SV model outputs. We have repeated the procedure 1000 times to find the distribution of SV and its 95% CI chance values.

SVM Model formulation

There are many approaches used for small area estimation, the formulation of the model depends on the parameters of selection and objective of our interest. Suppose weconsider the data sample of $\{y_i\}, \text{ics}\bar{y} =$
$n\ i\text{cc}w_iy_i\ \sum$
$\text{Where} w_i = n_i^{-1}$
design weight n_i is the probability of selecting the unit ith
w_i
sample 'S' (The sample were selected based on the stratified random sample with cluster approach; Strata 1:District;Strata 2: Tahasil; Strata 3: Hobli and Villages) weights are assigned independently form $'y_i'$
$_u$
$\{x_iy_i\}$ ics $X = (x_1, x_2, x_3 \ldots . x_q)$ Population totals for 'q' auxiliary variables generated regression estimators .In case of vector support machine,the eqn becomes

$$y_n = w_ix_i + b_j \quad (1.1)$$

$$y_i - w_ix_i - b_j \leq s_k \text{ (Negative response)}$$

$$w_ix_i + b_j - y_i \leq s_k \text{ (Positive response)}$$

$$\text{Manimization} = \sum^n \quad \text{Nin}\ ^1 \|w\|^2 \quad -$$

$$(y_i - \hat{y}_i)^2, (x_i, x) + b_{ij} \quad 2$$

(1.2)

$$\text{Manimization} = \sum_{i=1}^{n} (y_i - y_i)^2, k(@(x_1 x_{2\ldots} x_i)\ @(x) + b_{ij} \quad (1.3)$$

The eqn (1.3) kernel density or function obtained from independent variables; Gaussian radial basis, the SVM will be

$$k(x_i, x_j) = \exp\left[-\frac{\|s_i - s_j\|^2}{2a^2}\right] \quad (1.4)$$

Multiple linear regression (MLR) is a unique method used to forecast the linear relationship between a dependent variable considering with more than two or more independent variables.

Observed data $y_i = b_0 + b_1x_1 + b_2x_2 + \cdots \ldots . b_pX_p$ (1.5)

Predicated data $\hat{y}_t = b_0 + b_1x_1 + b_2x_2 + \cdots \ldots . b_pX_p$

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 4459 - 4474
Received 20 January 2021; Accepted 08 February 2021.

$\text{Minimization} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$; $b = (x^u x)^{-1} x^u yb = C^u_{XX}\, Cyx$

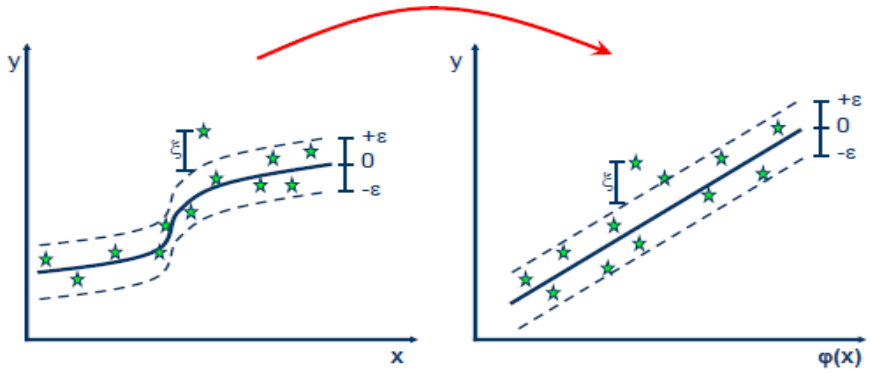
$y^u = bx$; $y = y^u + s$

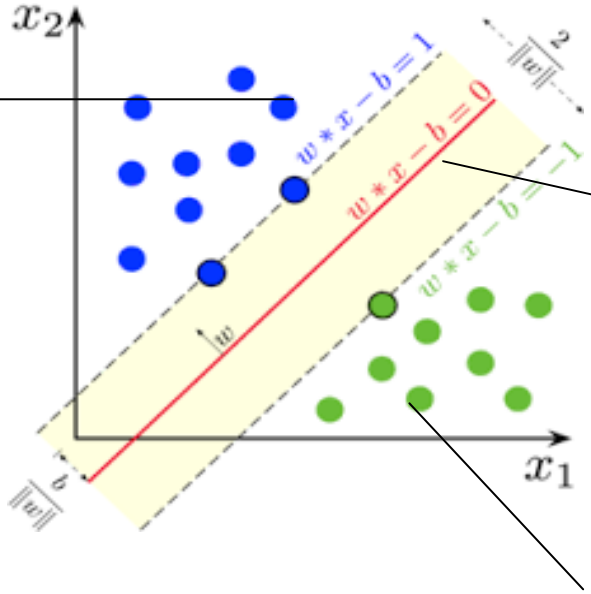
Co efficient of determination R^2 (%) $= \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

$SST = \sum_{i=1}^{N} (y_i - \bar{y}_i)^2$; $SST = \sum_{i=1}^{N} (y_i^u - \bar{y}_i^u)^2$; $\sum^{N} (y_i - y^u)^2$

$F = \frac{MSR}{MSE} = \frac{SSR/dfMSR}{SSE/dfMSR}$; df MSE= (n-p-1)

Significance of testt $\frac{\frac{b_i}{(n-p-1)}}{SE(b_j)} = \frac{b_j}{fSE(s^F s)} \frac{SE}{n-p-1} = J^{\frac{SSE}{}}$





Response variable Y_i or X_2

Hyperplane

Independent variable X_1or X_2

Area level approach

The area level model includes random areas specific effects and area covariates Xi s

$\theta_i = X_i b_j + z_i u_i$ (1.7)

θ_i = Parameter of the interestz_i = known positive constant$\mu = 0$; $a_u^2 (u_i \sim N(0, a^2_u)$

$\hat{\theta} = \theta + \varepsilon$; $\hat{\theta}$ = Direct design unbiased estimation s = are independent sampling errors

$\hat{\theta}_i = X_i b_j + Z_i v_i + \varepsilon_i$ Where i=1,2…m (1.8)

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 4459 - 4474
Received 20 January 2021; Accepted 08 February 2021.

This is the special case of the general linear mixed effect model with diagonal co variance structure

$\hat{\theta} = N(X b_{i}, z^2 \sigma^2_{i} + \sigma^2_{u} + \sigma^2_{e})$

$\hat{\theta} \sim N(Xb, ZGZ^{u} + R)\ v = ZGZ^{u}$

Best linear unbiased BLUP for $\theta b = (X^{u}V^{-1}X)^{-1}X^{u}V^{-1}Y$ (1.9)

Y is the SVM vector of the observation, then eqn (1.9) becomes

$\hat{u} = GZ^{u}V^{-1}(Y - X\hat{b})$

b and u can be obtained by penalized MLE's 'u' considered as fixed

$\hat{b} = (X^{u}\hat{V}^{-1}X)^{-1}X^{u}\hat{V}^{-1}Y$ (1.10)

$\hat{u} = \hat{G}Z^{u}\hat{V}^{-1}(Y - X\hat{b})$ (1.11)

Expected best linear unbiased (EBLUPS) eqn(1.11) model constructed based on theoretical approach, we estimated the parameters from average of the auxiliary variables. In many medical and life science applications, the EBLUPS is a standardized technique to derive small area estimation from SVM approach. In the advent of spatial, time, spatiotemporal robust estimation process, we estimated the binary counts of dependent variables ($Y_1, Y_2 \ldots Y_n$) with auxiliary variables of ($X_1, X_2 .. X_n$)

$\theta_p = \alpha_p + X^{u}_{i} b_p + \varepsilon_i$

M-quintile of SVM 8_p in the order of e with $ec(0,1)$ of the conditional distribution $(\overset{Y}{s} = x)$ is defined as

$\int \varphi P\ (y - \theta p(x) dF(y/x) = 0$ (1.12)

$\varphi P(r) = \begin{cases} (1-p)\varphi(r) r < 0 \\ p\varphi(r)\ r \geq 0 \end{cases}$ (1.13)

M- quintile regression is a unified model that includes quintile regression line associated with SVM approach of selected study sites to know the effect of residuals support vectors machine (SVM). Small area SVM model estimates borrow strength from all samples to capture random area effects, given the hierarchical structure of the data sets collected from different study sites . We characterized the conditional variability across the population of interest to magnify 3D level. Linear missed effect model captures random area effects as differences in the conditional distribution of 'Y' given 'X' between small areas. The M-quintile determine the area effect of individual unit belonging to the larger proportion .

$Y_{ij} = X^{T}_{ij} b_{™} 8_i + s_{ij}$ (1.14)

Y_{ij} = Observed value of i^{th} patient and j^{th} location

8_i = Unknown parameter estimated from support vectors machine (SVM)

$b_{™}$ = Regression co efficient auxiliary variables

X^{T}_{ij} = Transformed observation of i^{th} patient and j^{th} locationto extraction of support vectors machine (SVM)

(ii) Multiple Cluster support Vector machine (MCSVM)

MCSVMis one of the new emerging MLapproaches suitable for various classification and emerging regression problems. The important goal of MCSVM is to construct two parallel planes in each clusters by optimizing a pair of QPPs, in such a manner each of the hyper plane is nearer to the data samples of one or more classes while distant from the data points of neighbor cluster. As shown in figure there are two (cluster) classes

- class 1 and class 2 which are divided by using two non-parallel planes in such a way that, each plane is nearer to the data points, samples of one class while farther from the other classrespectively.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 4459 - 4474
Received 20 January 2021; Accepted 08 February 2021.

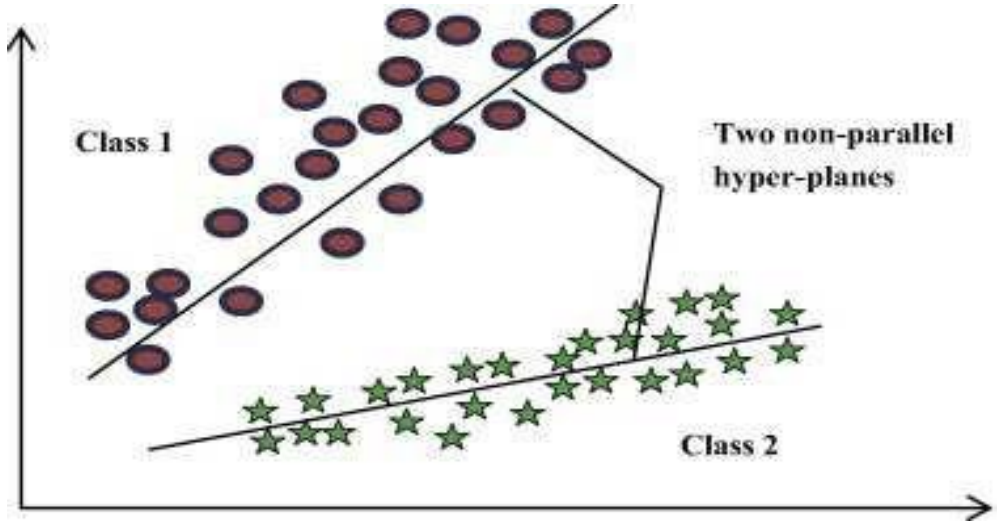


Fig1 classification of clusters or classes using MCSVM

SVM was applied on real data sets of COVID 19 pandemic , samples of diagnostic tests of COVID19 obtained from the RAPID test kit coded as positive and negative test results are denoted by 'm' and 'n' respectively.The positive and negative cluster data sets are represented by matrix form $X_1 c \mathbb{R}^{Nsk}$ and $X_2 c \mathbb{R}^{Nsk}$…..$X_i c \mathbb{R}^{nsk}$correspondingly where$\mathbb{R}$ indicates the k- dimensional real space.Eqn of two or more parallel hyper planes in k-dimensional real space $\mathbb{R}^k$ are given below.

$X w_1^T + b_1, X w_2 + b_2 \ldots X w_i + b_j = 0$.Here $w_1, w_2 . . w_i$ indicates normal vectors of two or more hyper plane , $b_1, b_2 \ldots b_j$are bias term or slope.The formulation of cluster SVM for linear cases are obtained

$$\min\left(\frac{1}{w_1, b_1, \xi}\right) \frac{1}{2}\|x_1 w_1 + \varepsilon_1 b_1\|^2 + C_1 \varepsilon^T{}_2 \xi$$

$$s.t\ (x_1 w_1 + s_2 b_1) + s \geq s_2 £ \geq 0$$

$$\text{Nin}\ \frac{1}{(w_2, b_2, n)} \frac{1}{2}\|x_2 w_2 + s_2 b_2\|^2 + C_2 s^T{}_1 n$$

s.t $(x_1 w_2 + \varepsilon_2 b_2) + n \geq \varepsilon_1, n \geq 0$; for n^{th} cause the eqn becomes

$$\min_{i}(w , b , n) \frac{1}{j} \|x w + \varepsilon b\|^2_2 + C \varepsilon^T_i \quad (1.15)$$

Leg rangier multiplier, the eqn will be equated in the form of the followings

$$\frac{6L}{6w_i} = X^T_n (X_i w_n + \varepsilon_i b_n) \quad (1.16)$$

$$\frac{\partial L}{\partial b_1} = \varepsilon^T_1 (X_i w_P + \varepsilon_i b_P) + \varepsilon_2 \alpha = 0$$

$$\frac{\partial L}{\partial \xi} = C \varepsilon^T_1 - b^T_2 - \alpha^T = 0$$

$$\alpha^T(X_2 w_1 + \varepsilon_2 b_1) - \varepsilon_1 + c_2 = 0 \quad (1.17)$$

If $\alpha = 0$; $b = 0$ from eqn (1.17)

$$\left[\begin{matrix} X_1^T \\ T^1_S \end{matrix}\right][X_1^u \; s_1][\begin{matrix} w_1 \\ b_1 \end{matrix}] + [\begin{matrix} X_2^T \\ s_2^T \end{matrix}]\alpha = 0$$

$A^T A u_1 + B^T \alpha = 0$; $u_1 = -(A^T A)^{-1} B^T \alpha$

$u_1 = -(A^T A + \delta I)^{-1} B^T \alpha$, similarly $u_2 = -(B^T B + \delta I)^{-1} A^T v$ (1.18)

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 4459 - 4474
Received 20 January 2021; Accepted 08 February 2021.

Results

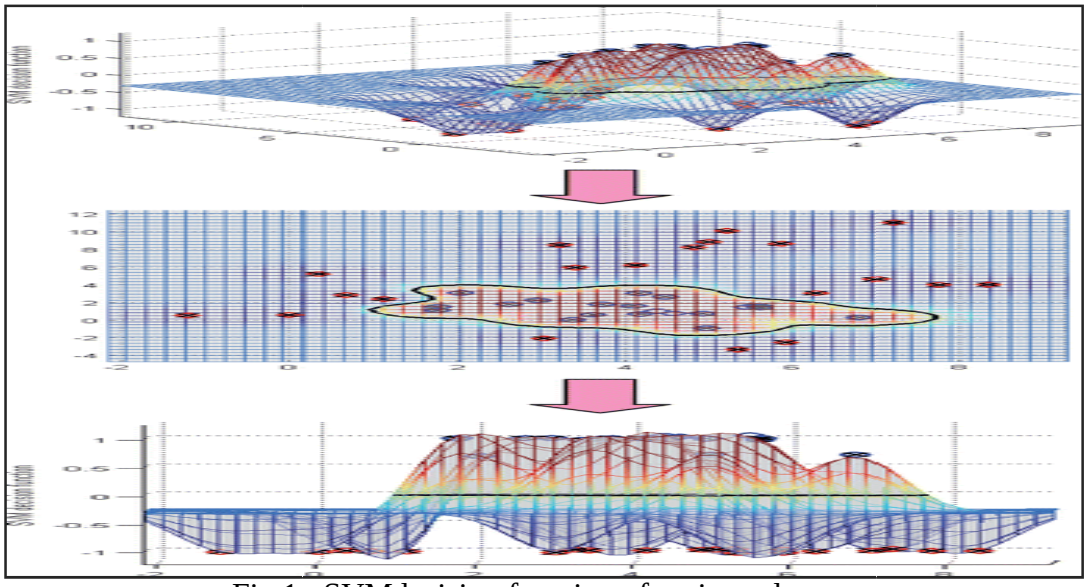


Fig 1 : SVMdecision function of various clusters

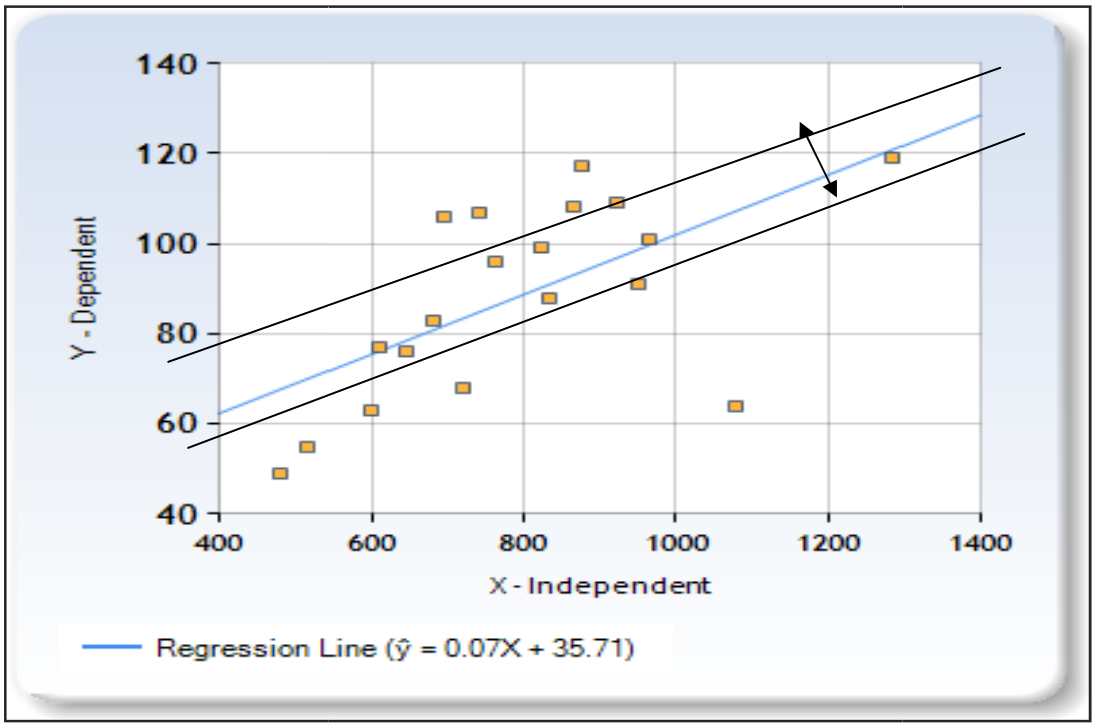


Fig 2: Support vector regression model

The performance of SVM on demographic profile in selected sites depends on the combination of several parameters,itshows very good sensitive gain to propagate the SVM out puts, the results shows good optimization techniques and minimizing the training errors (e_1,e_2,e_3..e_n) whose effect on the RMSE presented in (Fig 1) and generalization of output presented in the (Fig 2), the simulation of SVR regression line has fitted based on demographic profile of the subjects (X and Y independent variables), all demographic

profiles were kept constant during the process of optimization ($\hat{y} = 0.07X + 35.71$; predicted results of SVM $\hat{y} = 0.11X + 79.81$) .Simulation depicted that, tradeoff between generalization of various demographic parameters that controls the

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 4459 - 4474
Received 20 January 2021; Accepted 08 February 2021.

maximizing the margin and minimizing the training errors of SVM model. In practice the model construction shows a wide range of parameters, it was substituted to extract the optimal performance of the model ,assessed by using ordinary least square analysis and LASSO regression method (Fig3&4), predicted model results shows a full descriptor of data sets and trade off between the simulation techniques.The optimal value of s_l depends on the type of noise present in the data, which is usually ignored during the process of model building of SVM even if enough knowledge of the noise is available to select an optimal value of $s_{l,}$ we carefully consider the number of resulting support vectors of individual traits, if insensitivity will happen in any traits we ignored the training sets until values (Table 1) can reach maximum epoch and allow for the possibility of sparsely for the formulation of new segment of prediction with support vector regression methods like LASSO and OLS.

Table 1 : Support vector regression model outputs-trainer data

$X - M_x$	$X - M_y$	$(X - M_x)^2$	$(X - M_x)(X - M_y)$
-29.8947	7.7895	893.6953	-232.8643
-180.8947	-11.2105	32722.9058	2027.9252
-70.8947	-20.2105	5026.0637	1432.8199
-309.8947	-39.2105	96034.7479	12151.1357
-193.8947	-25.2105	37595.169	4888.1884
30.1053	10.7895	906.3269	324.8199
171.1053	12.7895	29277.0111	2188.3463
494.1053	30.7895	244140.0111	15213.241
40.1053	-0.2105	1608.4321	-8.4432
-274.8947	-33.2105	75567.1163	9129.3989
-49.8947	18.7895	2489.4848	-937.4958
85.1053	28.7895	7242.9058	2450.1357
158.1053	2.7895	24997.2742	441.0305
72.1053	19.7895	5199.169	1426.9252
-146.8947	-12.2105	21578.0637	1793.662
129.1053	20.7895	16668.169	2684.0305
-110.8947	-5.2105	12297.6427	577.8199
287.1053	-24.2105	82429.4321	-6950.9695
-98.8947	17.7895	9780.169	-1759.2853

OLS regression produces regression co efficient that are unbiased estimators of corresponding population co efficient with least variance .However, there may be a model with less variance (*ie* smaller SSE).This occurs in the following situations viz there are many independent variables , especially when there are more variables than observations, data is close to multicollinearity in which (XX^T) is not invertible or close to bring non invertible. In OLS, the goal is to minimize the SSE

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \hat{y}_i = \sum_{i=1}^{n} b_j x_{ij}$$

For ridge regression

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + h \sum_{i=1}^{n} b_j^2 \qquad (1.19)$$

$$B = (X^T X)^{-1} X^T Y$$

$$Cov(B) = a^2 (X^T X)^{-1}$$

Variance inflation rate $VF = (n-1)(X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1}$

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 4459 - 4474
Received 20 January 2021; Accepted 08 February 2021.

λis ridge term added to the diagonal X^TX

While ridge regression poses multicollinearity issues, it is not so easy to determine which variables should be retained in the model. These variables will converge zero more slowly yielded lambda value which can shows increasing trend but it will never equated zero. LASSO which stands for least absolute selection and shrinkage tool, addresses this issue since with this type of regression, some of the regression co efficient will be zero,indicating that the corresponding variables are not contributing to the model.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + h \sum_{i=1}^{n} |b_j| \qquad (1.20)$$

$$b_j = \frac{c_j}{z_j} \quad z_j = \sum_{i=1}^{n} x_{ij}^2 \quad c_j = \sum_{i=1}^{n} x_{ij} \left(y_i - \sum_{h*j} b_h x_{ih}\right)$$

In the cyclic co ordinate descent algorithms, at inception level fix all the b_j to some assumptions (eg., zero) and there by determine co efficient as described in the eqn(1.20) continue to do this until convergence (*ie* value don't change more than a predefined state)

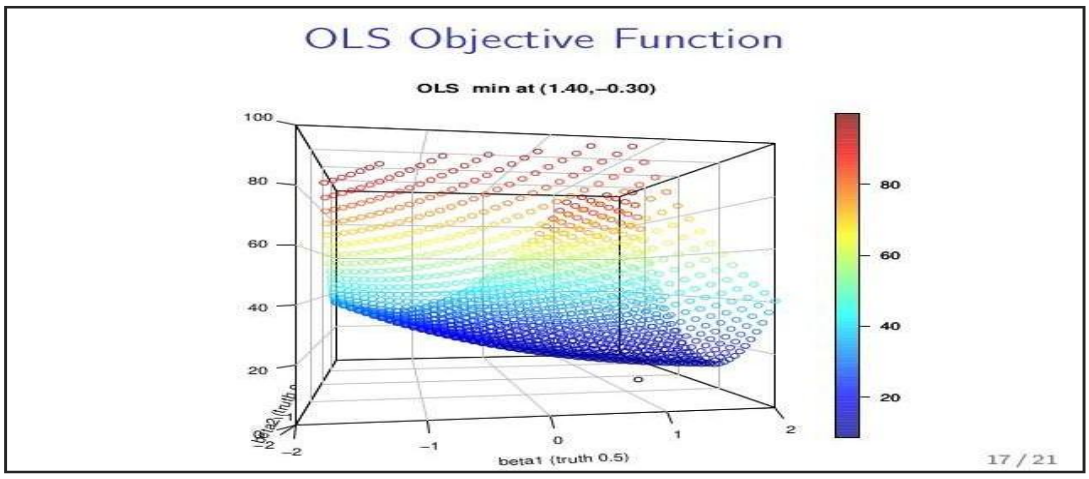


Fig 3 : Optimization of SVM in ordinary least square analysis

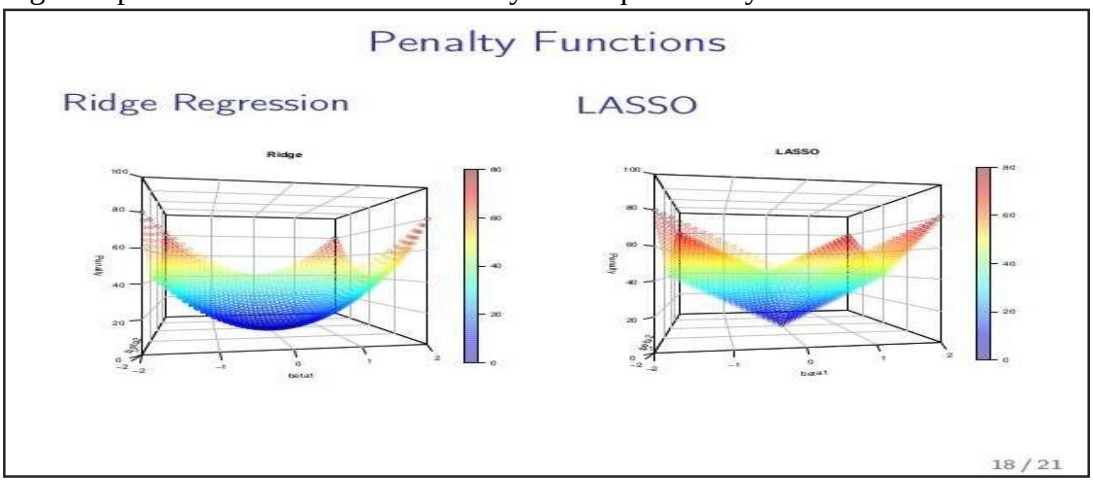


Fig 4: Optimization of SVM in LASSO regression method

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 4459 - 4474
Received 20 January 2021; Accepted 08 February 2021.

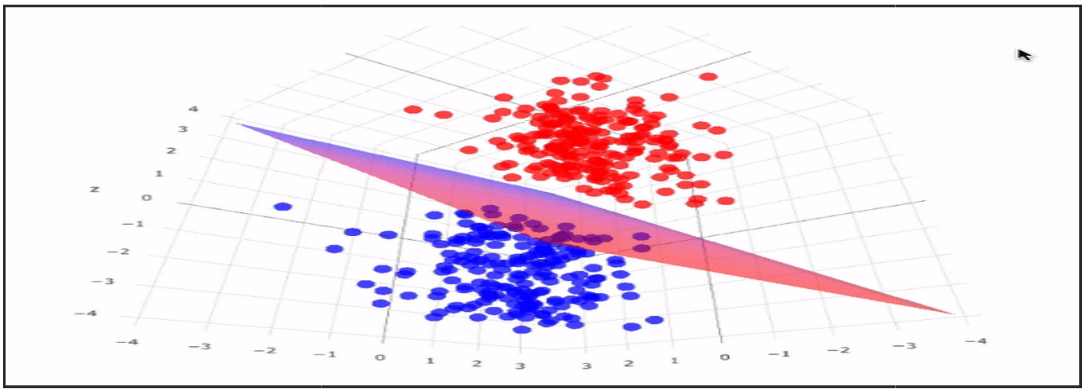


Fig 4: Grouping of K^{th} nearest value of SVM Eigen vectors

GMFRs available in eachSCof allsocio-economic factors, parameters were standardized with age specific comparison. Thus, there are 20 standardized GMFRs included instudy area. Discussing 20 GMFRs on each socio-economic factor is difficult to understand by a public health officials (Figure 1). A better approach is to use a summary measure viz., Median

GMFR for each socio-economic factor. Results shows the median standardized GMFR of eight selected socio-economic factors of the currently married women was 285.22 (63.24%) (Table 1). In case of illiterate women, the median GMFR was 99.3% per 1000 (95%CI: 95.0 – 103.5); the number of births that could be averted was (4.0%). If median is used as the measure for SAV then the percentage of births that could be averted in the selected socio-economic factors varied between (2.4%) and (5.0%) respectively. The number of births that could be averted under the assumption underlying that, (10-20%)sub centres had shows the lowest GMFR (107)in illiterate women, which is accounted (9.7%). The percentage of births that could be averted under the above assumption varies from (9.7%) to (17%).Pooled TFR was determined from traditional and Support vector machine, the results shows TFR in traditional methods was 4.41 and SVM was 5.36 (marginal differences 0.95),SVM method found to be more epoch rather than traditional method of calculation of TFR ($p<0.001$, sensitivity 98.56% AUC 0.92). Fertility rate was calculated from adjusted data sets , the highest fertility rate was recorded in 30-34 years (1.44%) followed by 35.-39 years (1.11).Fertility rate has seen lower value in younger age group (<1.0) because due to life style, food habits and biological risk factors early puberty,hormonal changes and environmental variations ($p>0.01$). ASFR was determined to know the relative comparison between formulated model , the ASFR in 20-24 years age group was (927.12) in traditional method, in case of SVM simulation , results were seen (988.13) in association with (992.65) and GMFR was (285.22; $p<0.01$;$R^2=0.95$,akiake likelihood ratio was 136.28). SVM measures expressed small area variation in tangible way that can help public health officials as well as programme managers to make judgement about whether to proceed with more detailed studies. More sophisticated methods use variations across areas underlying rates, called optimization techniques. (Table 2) shows the SVMon eight selected socio-economic factors. Although it provides a way of ranking of socio-economic factors according to the size of the small area variation, it is very easy to interpret. For example, the SVM onthe family planning measures among women (0.044) was not statistically significant from 0 (P=0.2) because point estimatation was marginal differences is larger than other methods which was statistically insignificant (P=0.252).

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 4459 - 4474
Received 20 January 2021; Accepted 08 February 2021.

Table 1: Descriptive statistics of TFR, ASFR, GFR and MFR in selected sites

Age group	No. of births	No. of women	Fertility rate
15-19	11269	97862	0.115152
20-24	43330	68731	0.630429
25-29	17695	70361	0.25147
30-34	77236	53526	1.442962
35-39	53524	47812	1.119468
40-44	24325	37360	0.651097
45-49	6375	27631	0.230719
Traditional method		SVM model	
Pooled TFR	4.41	5.36	
ASFR (20-24)	43330/46736=927.12	988.13	
GFR	974.53	992.65	
GMFR	276.43	285.22	
GRR	1.45	1.75	
Child women ratio		345	
Sex ratio at birth		785 per 1000	
Pregnancy prevalence rate		83.55% per 1000MWRA	

A better approach used to formulate the SVM to describe the TFR, GMFR, GRR and PRR.We haveestimated distribution of SFR corresponding with socioeconomic factors. In this instance, among family planning users, the median SFR was 0.9940;<1% (95% CI 0.6159 to 1.3886); the percentage of births has averted under the static assumptionsin small areas estimation.As per the model results median SFR was (4.7%) (95% CI 3.4 to 6.0). These measureshave relatively expressed good variation and meaningful way to describe the demographic and reproductive indicators , that can help the public health professionals to identify important reproductive traits from the trainers data sets of SVM model . From the model inputs we will be able to manipulate various transgenic factors of selected age group of the population, simultaneously we will draw necessary interventions among reproductive population. The resulted findings combineall information pertaining to systematicvariation and the average number of births attributed in particular socio-economic factor*etc*. Even if the median SFR shows more value of marginal errors we can alter the model in association with adjusted reproductive parameters. Small areas leadwith percentage of births that might be averted is large(number of births that could be potentially averted with small number and the ranking of socio-economic factors on fertility), the traits were reconstructed from SVM simulation techniques and we can take other suggested measures simultaneously.

Discussion

The first point of interest from the study was that three simple measures were used in this paper to describe small area variation among fertility rates [3,4,11]. These simple measures need not be limited to use with fertility data [2,8,10]. For example, these could also be used to report on small area variation in surgical or hospital admission rates. The main advantage of the two simple measures inculcate direct attention to the size of small area variations and its precision [5,6,18]. Public health authorities have been advocating this type of statistical reasoning rather than mechanically determining statistical significance for many years [18]. These simple measures correlatedare based on the assumption that all areas could achieve rates similar to those in the areas with median rate if only we exploredthat all that might be reasonably known about the socio-economic factors of the fertility in question, such as GMFR, and could apply this in practical programs in the community [15, 17, 19]. The concept has been applied to comparisons among large areas where the effects of random variation are small and are often ignored. The present approach extends the idea to small areas by accounting for the effects of

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 4459 - 4474
Received 20 January 2021; Accepted 08 February 2021.

random variation in fertility rates [2,3].For some socio- economic factors (the family planning users, those living in kuchcha households, labourers, illiterate husbands) unimportant small area variation was found [5]. This does not mean that the overall GMFR for these factors cannot be improved. However, in setting priorities for future research, it will be useful for public health managers to concentrate initially on those factors of fertility with substantial small area variation. In other words, if small area variation is substantial; the opportunity for improving the overall GMFR by identifying the modifiable factors for small area variation may be greater or else being equal.Definition of small areas varies depending on (i) the geography/location of the area, (ii) the fertility or the parameter of interest and (iii) administration. The analysis in this paper was based on 20 sub centre level data which are under the control of district health authorities. However, these comparisons were of particular interest to public health administrators. Any future action that might be taken to reduce regional variation will require action by the concerned district health authorities. The method used in this paper could also be used byother geographical units such as census blocks, Tahasil or PHCs. The limitations of routinely collected data should also be considered. For example, this study was carried out on the basis of the data collected through a representative sample survey in a district of south India. It may be possible that some of the small area variations might be due to inconsistencies reported in the birth certificates. Studies in India such as National Family Health Surveys (NFHS I & II) had shown that 'marginalized' population are inconsistently reported as the significant socio-economic factor for fertility [22,23].

This might explain the large variations in fertility rates due to various socio-economic factors across sub centres and warrant further investigation.Also, changes in the prevalence of socio-economic indicators among women/households for fertility may take several years to be reflected in reductions in fertility rates. For example, increase in age at marriage had resulted reductions in fertility among women after 10 to 15 years. Hence, for some factors substantial variation in fertility rates does not mean that there is currently substantial variation in modifiable socio-economic factors[5,11,12]. Although some might argue that a weakness of the method is that the choice of median, as a simple measure, is arbitrary, the situation is similar to converting standard error to confidence interval [2,5,8].The most commonly used confidence level in practice is 95% and hence the same is employed in this analysis. Similarly, instead of using the routine high *versus* low rates across small areas, median GMFR is suggested, because it is simple, easy to compute and more importantly understandable by the public health managers [15,18,22].The number and percentage of births that can be averted' suggested that at least 5% of the births could be averted under the given situation. However, this can be increased further with necessary intervention. Simple measures of small area variations described in this paper rely on the nonparametric bootstrap technique and thereby no assumptions were made on the distribution of births and this is an advantage to study the small area variation among fertility rates over the standard way of parametric modelling that used to be employed in practice. For example, assuming that the births followed binomial distribution [24,25,32]. If the variation in observed births is larger or smaller than the mean, assumption of binomial distribution theory will underestimate or overestimate the within area variation leading to overestimation/under estimation of the systematic variance [2, 10,18]. Most of the analysis on this issue has concentrated on the fertility events, for example, GMFR [3,4]. Similar work for events, such as hospital admission rates is an important area for future study. As we discussed earlier, 'the number needed to treat' (the reciprocal of absolute risk) is a better way of expressing the results of clinical studies than the traditional measures such as relative risk. In other words,

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 4459 - 4474
Received 20 January 2021; Accepted 08 February 2021.

the interpretation (the number of women who can be educated to prevent one birth) is easily understood by public health managers as understood by the clinicians.

Conclusion

Present study will help public health managers as well as policy makers to decide future course of action for more detailed studies based on the SVM's simple measures for reporting small area variation. The derivation of the model is very easy to construct multiple reproductive traits encompass with relative measures of fertility.

Acknowledgements

We grateful to all field investigators for awesome collection of research data during the process of survey, eagerly we acknowledge the participants and state health care professionals and bureaucrats for permitting us to conduct this research study.

Reference

[1] Bogart's model to Forager fertility.' 1986; In:W. Penn Handworker(ed), Culture and Reproduction, West View Press , London,59-82.
[2] Breslow NE.Extra- Poisson variation in log-linear models.' Applied Statistics 1984; **33**,38-44
[3] Bongaarts JA. A frame work for analyzing the proximate determinants of fertility.' Population Development Review 1978; **4**: 105-132.
[4] Crosse EA, Alder RJ, Ostbye T, Campbell MK. ' Small area variation in low birth weight: looking beyond socio-economic factors.', Canadian Journal of public health 1997; **88(1):**57-6
[5] Diehr P.Small area statistics: Large statistical problems", American Journal of Public Health 1984; 7(4): 313-314.
[6] Diehr P,Cain K, Connel F,Volin E. hat is too much variation The null hypothesis in small area analysis. Health Services Research 1990; **24**:741-771.
[7] Easterlin RA.Economic framework for fertility analysis; Studies in Family Planning, 1975; **6:** 54-63.
[8] Easrerlin RA, Crimmins EM. Fertility Revolution: A supply demand analysis. 1985; the university of Chicago Press, Chicago
[9] EcobR, MacintyreS. Small area variations in health related behaviours; do these depend on the behaviour itself, it's measurement, or on personal factors?', Health place 2000; **6(4):** 261-274.
[10] Jain AK. Fecundity and its relation to age in a sample of Taiwanese Women. Population Studies 1969; **23**: 69-86.
[11] James WH. The fecundability of U.S women. Population Studies 1975; **25**, 69-86.
[12] Mason KO. Status of women: A review of it's relationship to Fertility and Mortality.' 1984; The Rockfeller Foundation, Newyork.
[13] Kerdprasop, Nittaya & Poomka, Pumrapee & Chuaybamroong, Paradee & Kerdprasop, Kittisak. Forest Fire Area Estimation using Support Vector Machine as an Approximator 2018; 269-273. doi: 10.5220/0007224802690273.
[14] Kingsley Davis K, Judith Blake. ' Social structure and fertility: an analytic frame work.', Economic Development and Social Change 1956; **6(3):** 211-235.
[15] Ko BC, Kim HH, Nam JY. Classification of Potential Water Bodies Using Landsat 8OLI and a Combination of Two Boosted Random Forest Classifiers. 2015; 15: 13763–13777. doi:10.3390/s150613763.
[16] Michael Coory, Robert Gibberd. New measures for reporting the magnitude of small – area variation in rares,Statistics in medicine 1998; **17:** 2625-2634.
[17] McPherson K, Wennberg JE, Hovind OB, Clifford P. ' Small – area variations in the use of common surgical procedures: an international comparison of New England, and Norway", New England Journal of Medicine 1982; **307:**1310-1314.
[18] Ministry of Health and Family Welfare. Draft National Health Policy", Government Of India, New

Delhi (2001).
[19] Mountrakis G, Im J, Ogole C . Support vector machines in remote sensing: a review. ISPRS J Photogramm Remote Sens 2011; 66:247–259
[20] Martins S, BernardoN,Ogashawara et al. Support Vector Machine algorithm optimal parameterization for change detection mapping in Funil Hydroelectric Reservoir (Rio de Janeiro State, Brazil). Model. Earth Syst. Enviro 2016; **2:** 138
[21] Pal M, Mather PM Support vector machines for classification in remote sensing. Int J Remote Sens 2005; 26:1007–1011
[22] Sumi Kala1 , Magan Singh , Sujay Dutta , Narendra Singh , Shashank Dwivedi .Application of support vector machines for fodder crop assessment ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume IV-5, 2018
[23] Shen, Yu-ting. A Study of Fitting Local Geoid Model by Least Squares Support Vector Machine-A Case Study of Taichung Area. Master thesis, Department of Civil Engineering, National Chung H sing University, 2011(in Chinese).
[24] Srinivasan K, Murthy BN. Finding from two large scale surveys in Bihar and Rajasthan. Dynamics of Population and Family Welfare, Himalays Publishing House, Bombay.1993;9:223-245.
[25] Staines A, Bodansky HJ, McKinney PA et al. Small area variation in the incidence of childhood insulin-dependent diabetes mellitus in Yorkshire, UK: links with overcrowding and population density.", International Journal of Epidemiology 1997; 26(6),307-313.
[26] Sung Hoon et al. Application of Support Vector Machines in Assessing Conceptual Cost Estimates J. Computing in Civil Engineering 2007; 4(7) :215
[27] Twigger JP, Jessop EG, Small area variation in hospital admission random or systematic?' , Journal of Public Health 2000;114: 328-329.
[28] Ustuner M, Sanli FB, Dixon B. Application of support vector machines for land use classification using high-resolution rapid eye images: A sensitivity analysis. Eur. J. Remote Sens 2015; 48: 403–422. doi:10.5721/EuJRS20154823
[29] Vapnik VN. The Nature of Statistical Learning Theory, New York: Springer- Verlag,1995a.
[29] Osnes K Comparing methods for estimating the variation of risks of cancer between small areas Journal of Epidemiology and Biostatistics 2000; **5(3):**193-201
[30] Westerling R. ' Components of small area variation in death rates: a method applied to data from Sweden Journal of Epidemiology and Community Health 1995; **49**:214-221.
[31] Yuan-yuan dong ren. A Study of Support Vector Machine (SVM) Kemel Function and Parameter Selection, *Technology Innovation Herald 2010;* 9: 6-7
[32] Illesinghe A, &Palaniappan S. towards predicting credit risk in Sri lanka's banking sector.
[33] BHATIA, K. Library & Information Officer, National Institute of Occupational Health, Meghaninagar, Ahmadabad, India.
[34] Babu D R, Rao K N, Kolati S. (2019). The design of refrigeration, thermal insulation and an equipment for healthy ripening of mango and banana without using harmful chemicals. International Journal of Mechanical and Production Engineering Research and Development (IJMPERD), ISSN (P), 2249-6890.
[35] Murty A, Satyanarayana M, Devi I. (2019). Compressor Health Monitoring using IOT. International Journal of Mechanical and Production Engineering Research and Development, 8(3), 117-124.
[36] Thakur G, Lahari D K. A study of psychological impact on physical health and fitness among adolescents.
[37] Ai I, Ra A, Ao F. low dose gonadotropin protocol for ovulation induction in low resource centre.