Indian J Med Res 123, June 2006, pp 788-798

Disease mapping using mixture distribution

K. Chandrasekaran & G. Arivarignan*

*Tuberculosis Research Centre (Madurai Unit), ICMR, Government Rajaji Hospital & *Department of Applied Mathematics & Statistics, Madurai Kamaraj University Madurai, India*

Received December 15, 2004

***Background & objectives*: Data on infectious diseases like tuberculosis (TB) have been analyzed in the past without giving adequate attention to spatial variations. Earlier studies also attempted to display disease status of sub regions, usually census tracts, by categorizing them into quartiles, that helps the authorities to identify high- or low-risk areas. This approach is based mainly on binomial and Poisson models for disease data, and the recent attempts focus on using mixture models of Poisson distribution. We carried out this study to find wards of Madurai Corporation having high risks for TB disease, to develop a model of mixture of Poisson distributions for the number of cases and to classify each ward to one of many risk groups for TB disease, and to represent spatial distribution of TB incidence in Madurai city.**

***Methods*: Mixture models were used in finding the number of risk groups which might have produced the observed counts of TB patients in 72 wards of Madurai Corporation. The number of risk groups and the Poisson parameters of each group were found by maximum likelihood approach using the computer package C.A.MAN (Computer Assisted Mixture ANalysis). Bayesian methods were used to associate each ward to a particular risk group. The results were geographically presented in maps by using ArcView mapping software.**

***Results*: Using binomial model, 26 wards were categorized as high risk wards, and with mixture model approach 15 wards showed standardized morbility ratio (SMR) >1. The wards along river Vaigai and densely populated wards had high risk.**

***Interpretation & conclusion*: Our findings demonstrate the usefulness of the mixture models for disease data with geographical variations.**

Key words Disease mapping - mixture models - spatial statistics

Data collected in biological and medical research, usually contain variations. Identifying and separating these variations due to external and known factors is done by statistical methods. The usual procedure is to assume a probability density for the variable of interest (*e.g.*, the number of people having the disease), which is used to answer various questions on the (observed) values of the variable and to compute the measure of the variations of the variable.

But the variations found in the observed values might also have been caused by unobserved covariates or may be due to clustered observations. In such cases, the observed values are said to have over-dispersion or extra-heterogeneity. When data such as the number of people having a disease, are collected on small regions such as city wards, census tracts, *etc*., the variations that are found in these data may be due to spatial proximity between regions or may be due to inter-differences in variables of regions. Thus, over-dispersion is always present in the spatially referenced data.

The analysis of extra-heterogeneity and the representation of the geographic variations of the disease on a map of the study region has become an important topic in epidemiological research. Such maps are more useful to policymakers and implementers as they can target regions with high risks. Identification of high risk groups (or regions) also provides valuable hints for possible exploration and gives additional directives to control the disease or to obtain more health funding.

Howe[1], in his overview of disease mapping pointed to John Snow's famous map which traced cholera to the pump containing the contaminated water. Dole[2], and Hutt and Burkitt[3] have described variations in cancer incidence through maps. A common approach in map construction is the Choropleth method[4].

We undertook this study to analyze the data concerning smear positive TB patients in Madurai city, Tamil Nadu, and present the geographical distribution of TB incidence by mapping the chosen epidemiological measure (incidence rate) on the study region.

The Choropleth map of incidence rates of TB cases was obtained. This map was shown to be inadequate in representing the geographical distribution as the data had over-dispersion. We proposed a mixture distribution[5] (more explicitly a mixture of Poisson distributions) for the disease incidence and the non parametric maximum likelihood estimator (NPMLE) of mixture distribution was obtained by Computer Assisted Mixture Analysis (CAMAN) package[6]. Choropleth maps were prepared using ArcView software (ESRI, USA) for the estimated mixture distribution and to interpret the result. All statistical computations were carried out using R-public domain software.

Material & Methods

Data on patients with smear-positive TB were obtained from the records of Madurai Unit of Tuberculosis Research Centre (TRC) of the Indian Council of Medical Research (ICMR) for the period 1999 to 2003. These records included the cases referred from other hospitals. According to 2001 census, the population of Madurai was 9,28,675. The size of the population of wards varied from 4794 to 27972 in 2001 (Fig. 1). The number of observed cases and the size of the population of 72 wards of Madurai Corporation, Madurai, Tamil Nadu, are given in Table I.

Disease mapping: we first considered traditional method of presenting the spatial distribution of TB incidence by mapping the incidence rate of the disease, and introduced the notations, for ward number i ranging from 1 to 72, o_i denoted the observed number of cases in the i-th ward, and n_i denoted the size of the population in the i-th ward.

The incidence rate (IR) of the disease was defined as the ratio of the number of observed cases to the population at risk. More explicitly, $IR_i = o_i / n_i$. Fig. 2 shows the Choropleth map of the incidence rate of wards of Madurai city assigning shades based on a classification using quantiles (the four values that divide the distribution into five groups with equal number of values). The map presents the geographical distribution of the disease in the study region. However, it should be noted that the incidence rate was influenced by the ward population and that the incidence rate was small for the wards (ward numbers 7, 10, 17, 31, 58, 59 and 61) with large population. To estimate

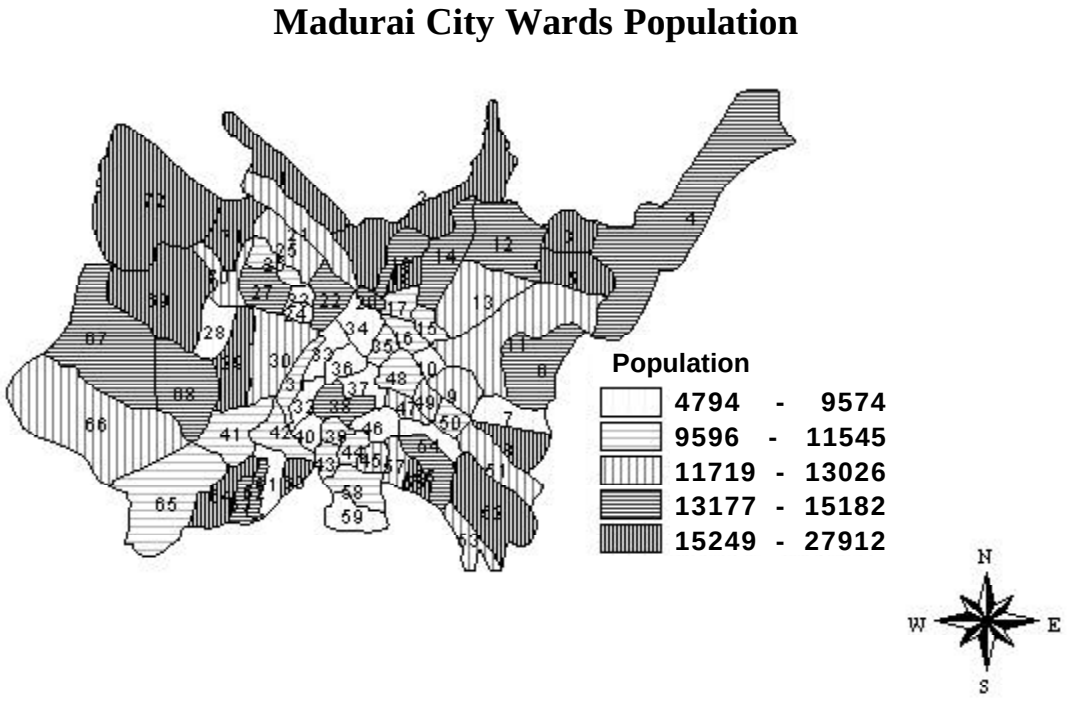

Fig.1. Choropleth map of population of wards.

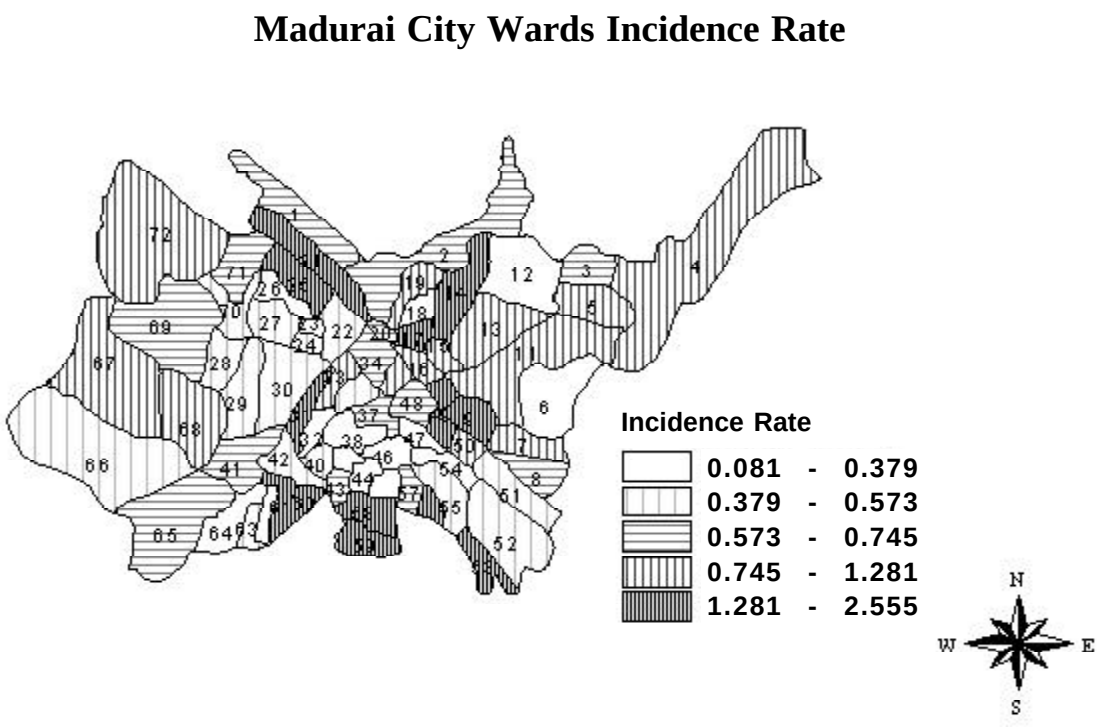

Fig.2. Choropleth map of incidence rate.

Table I. Number of observed cases, population and incidence rate for wards of Madurai city

Ward Number	o_i	n_i	o_i/n_i	Ward Number	o_i	n_i	o_i/n_i
1	17	23019	0.739	37	5	7589	0.659
2	17	27912	0.609	38	4	13177	0.304
3	10	15249	0.656	39	4	11119	0.36
4	12	13318	0.901	40	4	8695	0.46
5	14	18579	0.754	41	7	9799	0.714
6	5	13673	0.366	42	4	10415	0.384
7	5	4794	1.043	43	7	10581	0.662
8	10	16594	0.603	44	3	11545	0.26
9	23	13026	1.766	45	2	11788	0.17
10	19	8659	2.194	46	2	7971	0.251
11	10	11823	0.846	47	1	12289	0.081
12	5	13197	0.379	48	6	10217	0.587
13	12	12658	0.948	49	16	12117	1.32
14	28	13783	2.031	50	12	11195	1.072
15	9	10321	0.872	51	7	12791	0.547
16	14	10928	1.281	52	9	19776	0.455
17	15	10140	1.479	53	19	12988	1.463
18	6	16230	0.37	54	2	9574	0.209
19	13	13227	0.983	55	7	14689	0.477
20	10	13724	0.729	56	20	15418	1.297
21	16	12168	1.315	57	8	12797	0.625
22	7	15182	0.461	58	16	10481	1.527
23	3	8886	0.338	59	23	9002	2.555
24	4	8556	0.468	60	28	18513	1.512
25	19	11719	1.621	61	19	9197	2.066
26	4	11037	0.362	62	4	13868	0.288
27	6	13936	0.431	63	6	14018	0.428
28	4	9053	0.442	64	4	17494	0.229
29	10	24889	0.402	65	7	11134	0.629
30	7	12215	0.573	66	6	12954	0.463
31	22	9596	2.293	67	13	13551	0.959
32	2	8226	0.243	68	14	14945	0.937
33	8	8420	0.95	69	9	15684	0.574
34	7	9554	0.733	70	6	12179	0.493
35	10	9931	1.007	71	19	25496	0.745
36	4	8445	0.474	72	16	21236	0.753

the incidence rate by pooling the information from all the wards, binomial model was used for the number of TB cases. If we assume a constant incidence rate q for the TB disease in the whole study region, then q can be treated as "velocity" at which new cases occur homogeneously in all the wards. Thus, the probability for an individual to get the disease was q. An estimate of the incidence rate q is given by

$\hat{q}$ = The total number of observed TB patients / the total number of population

$$= \sum o_i / \sum n_i$$

With this estimate of q, the number of TB cases for the *i*-th ward can be modeled as binomial with parameters n_i and q. That is,

$$o_i \sim binomial(n_i, q)$$

The expected number of cases (e_i) in the *i*-th ward was given by the mean of the binomial distribution, namely $e_i = n_i\theta$, $i = 1, 2, \ldots, 72$. A measure of the difference between the observed number of cases o_i and the expected number of cases e_i could serve as a rate. In the literature the widely used rate has been defined by o_i / e_i, which is called standardized morgbidity ratio (SMR)[7] for the i-th ward (SMR_i). As the incidence rate IR_i was proportional to SMR_i, Choropleth map of SMR was not shown. As such mapping SMR does not help to identify high risk groups.

For the purpose of computing probabilities (and for extending the model), the binomial model for the *i*-th ward may be approximated by a Poisson model as n_i is large and q_i is very small. Thus we proposed that the number of cases for the *i*-th ward had

$$Poisson(m_i)$$

where the parameter m_i was a function of the expected number of cases e_i in an area and a relative risk λ_i for the *i*-th ward. That is

$$o_i \sim Poisson(\lambda_i e_i)$$

so that the probability of getting an observed count *y* in the *i*-th ward was given by

$$e^{-\lambda_i e_i} (\lambda_i e_i)^y / y!$$

where *y* was a non negative integer. We computed the standard error of SMR_i ($=o_i/e_i$) for the *i*-th ward as proportional to e_i. Hence the fluctuation in the observed count was indirectly proportional to the expected count. It is to be noted that when e_i is small, the SMR can change a lot by small changes in o_i, which may be simply due to chance. Thus our assumption of having a fixed relative risk λ_i failed to capture the over-dispersion in the data. This has to be kept in mind when interpreting the SMR values.

Moreover by having a separate relative risk (λ_i) for each ward we had as many parameters as the number of wards and the estimate of λ_i by o_i/e_i was not consistent. It is proposed in the literature that λ_i may be assumed to stem from a population of parameters with suitable distribution. Recently several models have been presented for describing spatial variation of rates[7-9].

Poisson mixture: Schlattmann and Bohning[5] showed that discrete mixtures were useful for modeling the population heterogeneity, which was common in disease mapping problems. In this approach, the relative risk of the ward is assumed to be realizations of a random variable, which is a mixture of Poisson distribution, *i.e.*

$$o_i \sim Poisson(\lambda e_i)$$

where, the relative risk λ, is treated as a random variable and is assumed to have a discrete probability distribution taking *k* values $\lambda_1, \lambda_2, \ldots, \lambda_k$ with probabilities $p_1, p_2, \ldots, p_k$ respectively, for some fixed k. Thus, we write,

$$o_i \sim \sum_{j=1}^{k} p_j Poisson(e_i \lambda_j).$$

The above distribution is called mixture distribution with $Poisson(e_i\lambda_j)$ as the component density and with the mixing distribution $\Pr[\lambda = \lambda_j] = p_j, j = 1,2...k$, which is represented by the following notation:

$$P = \begin{pmatrix} \lambda_1 & \lambda_2 & \cdots & \lambda_k \\ p_1 & p_2 & \cdots & p_k \end{pmatrix}$$

It may be noted that the mixing distribution P does not have any specific form for the density function depending on any parameter and hence it is said to be in non parametric form. The maximum likelihood estimator of P (denoted by $\hat{p}$) is called non parametric maximum likelihood estimator (NPMLE)[10]. We distinguished between flexible support size and fixed support size; in the first case the number of mixing components k was unknown, and in the later case, k was assumed to be known. In either case the estimation was done using maximum likelihood approach, which could be implemented by EM (expectation maximization) algorithm. The algorithm is detailed below (The theoretical basis is given in the Appendix). However, we have accomplished this by using the computer package C. A. MAN[6]. The software description is given by Schlattman and Bohning[11].

Appendix

We present here the derivations for the EM algorithm.

We have

$$o_i \sim \sum_{j=1}^{k} p_j Poisson(x \mid e_i\lambda_j), \quad i = 1, 2, \ldots, n,$$

where, $Poisson(x \mid e_i\lambda_j) = e^{e_i\lambda_j}(e_i\lambda_j)^x/x!$

The likelihood function for the sample o_i, $i = 1, 2, \ldots, n$ is,

$$L = \prod_{i=1}^{n}\sum_{j=1}^{k} p_j e^{e_i\lambda_j}(e_i\lambda_j)^{o_i}/o_i! \qquad (1)$$

Since this involves summation, it is not possible to take logarithm and it is difficult to use MLE method. We define a binary random variable g_{ij} which takes 1 if i-th observation has come from a Poisson distribution with parameter $e_i\lambda_j$ and 0 otherwise. Here g_{ij} is the missing data. The data $\{o_i\}$ are called incomplete data and $\{o_i, g_{ij}\}$ are called complete data. Then L is rewritten as

$$L = \prod_{i=1}^{n}\prod_{j=1}^{k}\{p_j e^{e_i\lambda_j}(e_i\lambda_j)^{o_i}/o_i!\}^{g_{ij}} \ldots(2)$$

The expression (2) is called likelihood function of complete data. The log-likelihood function of the complete data is

$$l = \sum_{i=1}^{n}\sum_{j=1}^{k} g_{ij}\, log\{p_j e^{e_i\lambda_j}(e_i\lambda_j)^{o_i}/o_i!\}$$

The E step of the EM algorithm is to compute the expected value of log-likelihood of complete data

$$E(l) = \sum_{i=1}^{n}\sum_{j=1}^{k} E_{ij}\, log\{p_j e^{e_i\lambda_j}(e_i\lambda_j)^{o_i}/o_i!\} \ldots(3)$$

where $E_{ij} = E[g_{ij} \mid o_i, \lambda_j, p_j]$

$= \Pr[g_{ij} = 1 \mid o_i, \lambda_j\ p_j]$

as g_{ij} is a binary random variable taking values 0 and 1. This probability can be computed by Bayes formula

$$\frac{p_j e^{e_i\lambda_j}(e_i\lambda_j)^{o_i}/o_i!}{\sum_r p_r e^{e_i\lambda_r}(e_i\lambda_r)^{o_i}/o_i!}$$

The M-step of the EM algorithm is to maximize the expected log-likelihood. We differentiate (3) with respect to parameters and equate them to zero to get the MLE's.

This gives the estimates:

$$p_j = \frac{1}{n}\sum E_{ij}$$

$$\lambda_j = \sum_i E_{ij}\, o_i / \sum_i E_{ij} e_i,$$

The algorithm for estimating the weights p_j and parameters λ_j of Poisson mixing distribution:

1. Start with some initial values for p_j and λ_j, $j = 1, 2, \ldots, k$

2. E-step: Compute $E_{ij,}$ expected log-likelihood, by Bayes formula:

$$E_{ij} = \frac{p_j f(o_i \mid e_i\lambda_j)}{\sum_r p_r f(o_i \mid e_i\lambda_r)} \quad i = 1, 2, \ldots, n$$

where $f(x|m) = m^x e^{-m}/x!$

3. M-step: Maximization of the expected log-likelihood function. This gives new values for the parameters:

$$p_j = \sum_i E_{ij}/n$$

$$\lambda_j = \sum_i E_{ij} o_i / \sum_i E_{ij} e_i$$

4. Repeat the steps 2 and 3 with new approximates until the desired accuracy is achieved.

Table II. Maximum likelihood estimates provided by C. A. MAN

Components	Weight p_j	Parameter λ_j	Log-likelihood
	0.3418	0.420	
k=3	0.4410	0.686	-224.2514
	0.2172	1.632	
	0.4631	0.449	
	0.3173	0.752	
k=4	0.1052	1.379	-220.1599
	0.1144	1.842	
	0.0031	0.415	
	0.3151	0.416	
k=5	0.4470	0.665	-220.7683
	0.1431	1.398	
	0.0916	1.905	

Using C. A. MAN package, the NPMLE of the mixture distribution was obtained for each value of *k* ranging from *3* to *5*. The results are given in Table II.

Since the values of log-likelihood were almost equal for the models *k=4* and *k=5*, any one of these models could have been chosen. But we selected the model *k=4* so as to have a parsimonious model. The observed values o_i were found to be arising from four subpopulations where j-th subpopulation represented fraction p_j of the whole population and j-th subpopulation had Poisson distribution with parameters $e_i\, l_{j,} j$ = 1, 2, 3, 4. The estimated values of l_j and p_j were

$$\hat{P} = \begin{pmatrix} 0449 & 0.752 & 1.379 & 1.842 \\ 0.463 & 0.317 & 0.105 & 0114 \end{pmatrix}$$

and the mixture density was given by

$$f(o_i \mid \hat{P}) = 0.463\, f(o_i \mid 0.449) + 0.317\, f(o_i \mid 0.752) + 0.105 f(o_i \mid 1.379) + 0.114\, f(o_i \mid 1.842)$$

where $f(o_i \mid l_j) = e^{-e_i l_j} (e_i l_j)^{o_i} / o_i!$

We next assigned each ward to one of the four subpopulations having relative risk l_j (j = 1, 2, 3, 4). For this, we computed the posterior probability for the membership of *i*-th ward in *j*-th subpopulation, (j = 1, 2, 3, 4) by Bayes formula:

$$P(l_j \mid o_i) = \hat{p}_j f(o_i \mid \hat{l}_j) / \sum_{r=1}^{4} \hat{p}_r f(o_i \mid \hat{l}_j),$$

i=1, 2, 3, 4, j=1, 2, 3, 4.

The *i*-th area was assigned to that subpopulation j for which it has the highest posterior probability of belonging. Table III presents the posterior probability and the assigned group membership for all wards, and Choropleth map (Fig. 3) presents these results. To identify low and high risk areas, we combined the two smaller relative risks (l_1

Table III. Posterior probabilities for each component of the mixture distribution

Ward number	Posterior probability for components				Associated components
	1	2	3	4	
1	0.135646538	0.853255325	0.011023513	7.46248E-05	2
2	0.389934008	0.609623861	0.000441775	3.56373E-07	2
3	0387464838	0.601444615	0.010807865	0.000282683	2
4	0.098848103	0.745550526	0.14042867	0.015172701	2
5	0.170247897	0.80623646	0.023064149	0.000451494	2
6	0.843992963	0.155661708	0.000341165	4.16357E-06	1
7	0.208579372	0.481105726	0.196478206	0.113836696	2
8	0.483691045	0.512079738	0.004168462	6.0755E-05	2
9	5.99289E-07	0.001428606	0.251978799	0.746591996	4
10	6.46946E-07	0.000678989	0.138527581	0.860792783	4
11	0.174384961	0.717469089	0.096905595	0.011240355	2
12	0.825215259	0.174271656	0.000505498	7.58744E-06	1
13	0.076967728	0.700438088	0.194583179	0.028011005	2
14	1.67458E-09	4.24132E-05	0.099340258	0.900617327	4
15	0.180178299	0.678607291	0.121022859	0.020191552	2
16	0.004769684	0.199186413	0.515270848	0.280773055	3
17	0.0007014	0.061386902	0.463109238	0.47480246	4
18	0.870045291	0.129838283	0.000115805	6.21098E-07	1
19	0.051554477	0.668330021	0.24354773	0.036567771	2
20	0.285182427	0.683158464	0.030129802	0.001529307	2
21	0.001602396	0.131905661	0.5529347	0.313557243	3
22	0.747497046	0.251731246	0.000763086	8.62198E-06	1
23	0.793876729	0.203776777	0.002224556	0.000121938	1
24	0.673547555	0.318066225	0.007732697	0.000653523	1
25	2.52884E-05	0.011112431	0.374166188	0.614696093	4
26	0.810059708	0.188844192	0.001065479	3.06218E-05	1
27	0.776621774	0.222600766	0.000766318	1.11424E-05	1
28	0.705267671	0.289129149	0.005245976	0.000357204	1
29	0.909122954	0.090871445	5.59885E-06	2.21567E-09	1
30	0.556326289	0.435784821	0.00757787	0.00031102	1
31	3.3658E-08	0.000127121	0.092118579	0.907754266	4
32	0.842721408	0.155835533	0.00136825	7.48088E-06	1
33	0.157706522	0.609101498	0.181410726	0.051781255	2
34	0.353681243	0.590688862	0.049207434	0.006422462	2
35	0.08536901	0.60169695	0.247567181	0.065366859	2
36	0.666156893	0.32466946	0.008426296	0.00074735	1

Contd...

Ward number	Posterior probability for components				Associated components
	1	2	3	4	
37	0.467832766	0.487189558	0.038380638	0.006597038	2
38	0.887298097	0.112519772	0.00018009	2.04134E-06	1
39	0.813665512	0.185310594	0.000996264	2.76298E-05	1
40	0.682644404	0.309862062	0.006941297	0.000552237	1
41	0.372692889	0.580529964	0.041865087	0.004912061	2
42	0.780880987	0.217284306	0.001768113	6.65943E-05	1
43	0.433195684	0.540170024	0.024581396	0.002052896	2
44	0.892267595	0.107482962	0.000245213	4.23045E-06	1
45	0.937038126	0.06289327	6.78153E-05	7.88069E-07	1
46	0.832648616	0.165559113	0.001689093	0.000103178	1
47	0.966401582	0.033583612	1.47029E-05	1.02877E-07	1
48	0.539663668	0.445631	0.01370178	0.001003552	1
49	0.001527923	0.12761358	0.551248733	0.319609763	3
50	0.037970676	0.523942695	0.344426841	0.093659788	2
51	0.597578481	0.397342031	0.004922218	0.00015727	1
52	0.796300228	0.203560598	0.000138795	3.7971E-07	1
53	0.000101889	0.031204142	0.497730611	0.470963357	3
54	0.887682194	0.111860172	0.000444121	1.35135E-05	1
55	0.719907114	0.278946709	0.001130352	1.58246E-05	1
56	0.000351511	0.090309699	0.631729165	0.277609625	3
57	0.467733895	0.51999422	0.011770831	0.000501054	2
58	0.000292619	0.038926809	0.440566958	0.520213614	4
59	3.92524E-09	2.94011E-05	0.055426945	0.94454365	4
60	4.84048E-07	0.003191721	0.46155723	0.535250565	4
61	1.26098E-06	0.001135601	0.168786267	0.830076871	4
62	0.905559191	0.094339441	0.000100524	8.43776E-07	1
63	0.780671381	0.218601476	0.000717082	1.00613E-05	1
64	0.964194299	0.035801182	4.51164E-06	7.82811E-09	1
65	0.475575462	0.506681229	0.016649957	0.001093352	2
66	0.723906666	0.274371928	0.001683882	3.75227E-05	1
67	0.059360734	0.701760717	0.211318687	0.027559862	2
68	0.055987147	0.745591913	0.181201408	0.017219532	2
69	0.547875459	0.448666125	0.003403259	5.51563E-05	1
70	0.676938713	0.319866441	0.003098148	9.66974E-05	1
71	0.101946532	0.889031841	0.00898466	3.69681E-05	2
72	0.135787028	0.846977708	0.017047697	0.000187567	2

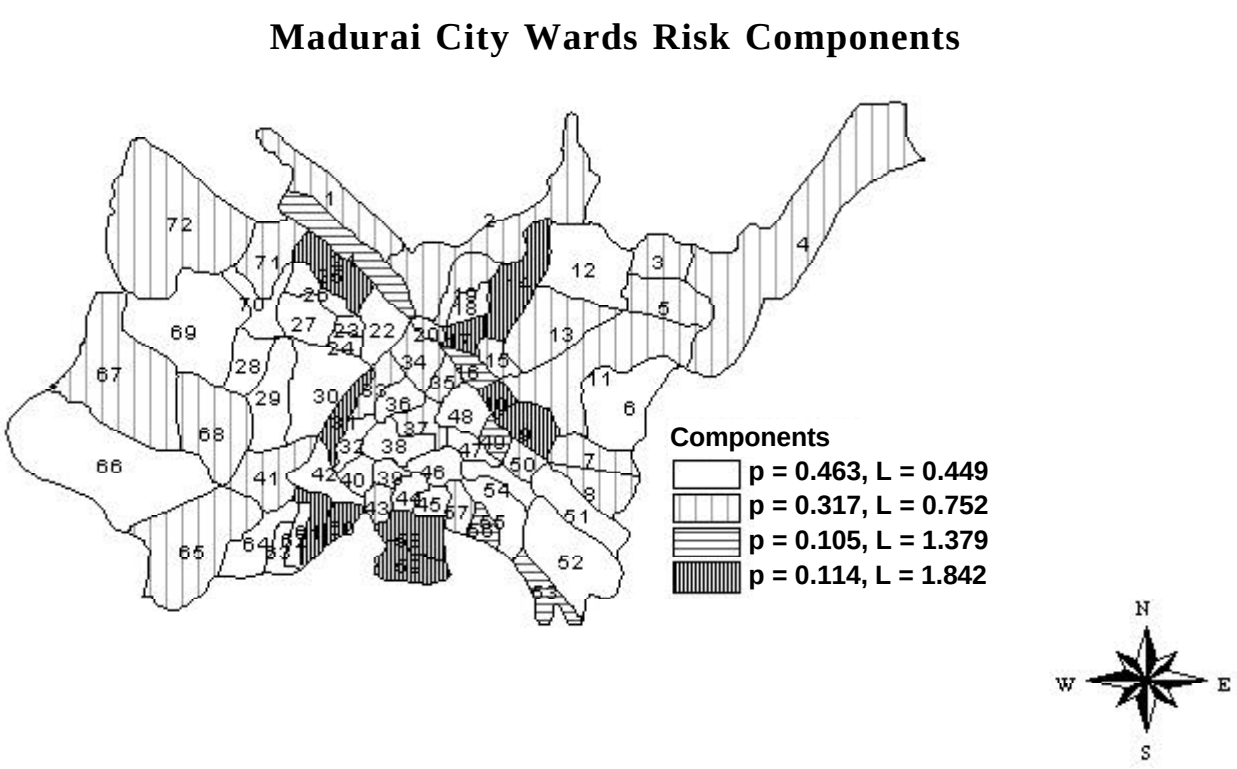

Fig. 3. Choropleth map of predicted components of mixture distribution.
(P - Population proportion & L - relative risk)

and l_2 which are less than 1) and the two higher relative risks (l_3 and l_4 which are more than 1). The proportion of low risk in the population was 0.78 (= $p_1 + p_2$) and that of high risk was 0.22 (= $p_3 + p_4$) .

Results

Under the binomial model, 26 wards (namely 4, 11, 7, 9, 10, 13, 14, 15, 16, 17, 19, 21, 25, 31, 33, 35, 49, 50, 53, 56, 58, 59, 60, 61, 67, 68) were categorized as high risk region (SMR > 1). Under the mixture model approach 15 wards were found to have SMR more than 1. The five wards namely, 16, 21, 49, 53 and 56 were classified as belonging to third subpopulation and 10 wards (namely 9, 10, 14, 17, 25, 31, 58, 59, 60 and 61) to fourth subpopulation. Of these 15 wards, seven wards (namely 9, 10, 16, 17, 21, 25 and 49) were situated on the bank of river Vaigai. The remaining wards were densely populated. Thus, the densely populated wards, and those along the river Vaigai, had high risk for TB incidence.

Discussion

In this study, we outlined the traditional method of computing SMR according to some basic models, namely, binomial and Poisson. For the spatially referenced data, due to the presence of over-dispersion, the traditional methods fail. Several efforts have been undertaken to provide valid estimates of SMR[7-9]. We used the mixture distribution approach[5,6] which are known as empirical Bayesian methods[12].

One of the pitfalls connected with this approach was in determining the number of components of the mixture of distributions. We solved the problem by computing (from C. A. MAN) the log-likelihood ratio statistic and by comparing the log-likelihood ratio statistic among subpopulations. Another

weakness of this non parametric empirical Bayes approach lies in the fact that it does not take autocorrelation into account. However, from a mathematical point of view, mixture models within an empirical Bayes framework provide a satisfactory method of clustering regions[13].

Acknowledgment

Shri K. Chandrasekaran thanks the Indian Council of Medical Research (ICMR) for granting study leave to work as a research fellow at Madurai Kamaraj University.

References

1. Howe, GM. Historical evaluation of disease mapping in general and specifically of cancer mapping. In: Boyle P, Muir CS, Grundmann E, editors. *Cancer mapping*. Berlin: Springer; 1990 p. 1-21.

2. Doll R. The epidemiology of cancer. 1980; *45* : 2475-85.

3. Hutt MSR, Burkitt DP. *The geography of non-infectious disease*. Oxford: Oxford University Press; 1986.

4. Walter SD, Birnie SE. Mapping mortality and morbidity patterns: an international comparison. *Int J Epidemiol* 1991; *20* : 678-89.

5. Schlattman P, Bohning D. Mixture models and disease mapping. *Stat Med* 1993; *12* : 1943-50.

6. Bohning D, Schlattman P, Lindsay BG. C.A.MAN-computer assisted analysis of mixtures: statistical algorithms. *Biometrics* 1992; *48* : 283-303.

7. Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 1987 ; *43* : 671-81.

8. Bernadinelli L, Montomoli C. Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Stat Med* 1992; *11* : 983-1007.

9. Besag J, York J, Mollie A. Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Statist Math* 1991; *43* : 1-59.

10. Laird NM. Non parametric maximum likelihood estimation of a mixing distribution. *J Am Stat Assoc* 1978; *73* : 803-11.

11. Schlattman P, Bohning D. Computer packages C.A.MAN (Computer Assisted Mixtures analysis) and DISMAP. *Stat Med* 1993; *12* : 65.

12. Maritz JS, L Win, T. *Empirical bayes methods*, 2nd ed. London: Chapman and Hall; 1989.

13. Mariott FHC. *The Interpretation of multiple observations*. London: Academic Press; 1974.

Reprint requests: Shri K. Chandrasekaran, Tuberculosis Research Centre, Madurai Unit (ICMR)
Ward No. 62, Government Rajaji Hospital, Madurai 625020, India
e-mail: chandru_pmpi@yahoo.com