Chapter 6

Poisson Mixtures Distribution

6.1 Introduction

In the previous chapters we presented Disease mapping of SMRs and incidence rates based on quantiles. However, this approach has been criticized by many researchers. One criticism is that classification based on percentiles is rather arbitrary, because there is no guarantee that such a classification can validly detect high or low risk areas (Schlattmann et al [178]). Another problem involves the instability of the crude SMR, especially when rare diseases are investigated in an area with a small population. In such case, both the observed and expected values are low. As a result, an area with a small population tends to present an extreme SMR, yielding a map which is dominated by the least reliable information (Bernardinelli et al [18]; Heisterkamp et al [109]).

Another method, used traditionally, is based on a classification using the probability value. However, a disease map which is based on this approach often faces the problem of misclassification as well, because an area with a small population size has a greater chance of showing a significant result (Bohning [34]). Additionally, the significant method approach faces the problem of multiple testing, and even adjusting for the number of comparisons does not lead to a consistent estimate of heterogeneity (Schlattmann et al [180]).

As both of these traditional approaches have some deficiencies and disadvantages in representing the geographical distribution of disease, many researches have sought alternative solutions for mapping disease. Empirical Bayes (EB) estimation provides a more stable relative risk estimate, and thereby overcomes some deficiencies of traditional maps which are based on the SMR. It was found that the EB approach removes the random variability which is present in data from small population counts (Bohning [34]; Bohning and Schlattmann [33]), leading to smooth map with fewer extremes in the relative risk estimates (Clayton and Kaldor [46] ; Marshall [141]; Mollie and Richardson [152]; Devine and Louis [65]).

As a solution to the foregoing problems, mixed modelling has been proposed for disease mapping (Bohning and Schlattmann [33]). The mixture model approach reduces the random variation in the disease map than do the percentiles method, the significant method, or the EB estimation. A disease map based on the mixed model approach not only provides a shrinkage estimator in the form of the mean of the posterior distribution, but also provides an estimate of the underlying risk structure (Schlattmann et al [180]). Another methodological advantage of using mixture model for disease mapping is that an estimate of the number of components (each with its respective coloring pattern) is provided (Schlattmann [180]; Bohning and Schlattmann [33]).

In this chapter we investigate the geographical distribution of Cancer in Chennai wards for the year 2001 by applying the mixture model method.

6.2 Methods

6.2.1 Data Sources

The basic geographical aggregation unit for disease mapping in this study was the ward, as before. The standardized Incidence Ratio(SIR) as well as SMR were used as indicators of the occurrence of Cancer in each ward.

As the sex of the population affects the incidence of Cancer, the sex standardized ratio was considered to be the proper epidemiological measure for this study. We used the indirect standardized method for calculating the sex standardized ratio to determine the rate of occurrence of Cancer for each ward.

6.2.2 Poisson Mixture

Schlattmann and Bohning [178] showed that the discrete mixtures are useful for modelling the population heterogeneity, which is 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 values λ_1, λ_2, ..., λ_k with probabilities p_1, p_2, ..., 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 a mixture distribution with $Poisson(e_i\lambda_j)$ as the component density and with the mixing distribution $\Pr[\lambda = \lambda_j] =$

$p_j, j = 1, 2, \cdots, k,$ which is represented by the following notation:

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

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) (Laird [123]). The constant k is called the support size and we distinguish between flexible support size and fixed support size; in the first case the number of mixing components k is unknown, and in the later case, k is assumed to be known. In either case the estimation is done using the maximum likelihood approach, which can be implemented by **EM (Expectation Maximization) algorithm.** The algorithm as implemented by us for the current problem is detailed below (The theoretical basis is given in the appendix). However, we have also accomplished this by using the computer package C. A. MAN (Schlattmann and Bohning [178]).

6.3 EM Procedure

We present here the derivation for the EM algorithm. We have

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

where $Poisson(x|e_i\lambda_j) = e^{-e_i\lambda_j}(e_i\lambda_j)^x/x!$ The likelihood function for the sample $o_i, i = 1, 2, \ldots, n$ is,

$$L = \prod_{i=1}^{n} \sum_{j=1}^{k} P_j e^{-e_i\lambda_j}(e_i\lambda_j)^{o_i}/o_i! \tag{6.1}$$

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

$$L = \prod_{i=1}^{n}\prod_{j=1}^{k}(P_j e^{-e_i\lambda_j}(e_i\lambda_j)^{o_i}/o_i!)^{z_{ij}} \tag{6.2}$$

above expression is called likelihood function of complete data. The log-likelihood function of the complete data is

$$l = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}log(P_j e^{-e_i\lambda_j}(e_i\lambda_j)^{o_i}/o_i!)$$

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

$$E(l) = \sum_{i=1}^{n}\sum_{j=1}^{k} E_{ij}log(P_j e^{-e_i\lambda_j}(e_i\lambda_j)^{o_i}/o_i!) \tag{6.3}$$

where

$$E_{ij} = E[z_{ij}|o_i, \lambda_j, p_j]$$

$$= Pr[z_{ij} = 1|o_i, \lambda_j, p_j]$$

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

$$\frac{P_j e^{-e_i\lambda_j}(e_i\lambda_j)^{o_i}/o_i!}{\sum_r P_r e^{-e_i\lambda_r}(e_i\lambda_r)^{o_i}/o_i!}$$

M-step of the EM algorithm is to maximize the expected log-likelihood. We differentiate (6.3) with respect to parameters and equate them to zero to get the MLE's. This gives the estimates:

$$\begin{aligned} p_j &= \frac{1}{n}\sum E_{ij} \\ \lambda_j &= \sum_i E_{ij}o_j / \sum_i E_{ij}e_i \end{aligned}$$

6.3.1 Algorithm

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

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

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

   $$E_{ij} = \frac{p_j f(o_j|e_i\lambda j)}{\sum_r p_r f(o_i|e_i\lambda_r)}$$

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

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

$$\begin{aligned} p_j &= \sum_i E_{ij}/n \\ \lambda_j &= \sum_i E_{ij}o_i / \sum_i E_{ij}e_i \end{aligned}$$

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

6.4 The Posterior Probability of Mixing Distribution

The final step in the mixture model approach for disease mapping was to classify the SIR and SMR values for each area into one of the components of the mixing distribution. This was accomplished by applying Bayes theorem and by using the estimated mixing distribution as the prior distribution. Classification was done by computing the probability of each area belonging to a certain component. When z_{ij} is the unobserved variable which describe area i in sub population j (z_{ij}), the posterior probability is defined as:

$$Pr(z_{ij} = 1|o_i, \hat{P}, e_i) = \frac{\hat{p}_j f(o_i \hat{\lambda_j}, e_i)}{\sum_{j=1}^{k} \hat{p}_j f(o_i, \hat{\lambda_j}, e_i)}, \text{for} j = 1, \ldots, \ k \text{and} i = 1, \ldots, n.$$

The i^{th} area is then assigned to that sub population j for which it has the highest posterior probability of belonging (Bohning and Schlattmann [33]).

6.5 Results

The table 6.1 gives the mixing probabilities and the associated parametric values for k (number of components) $=$ 2,3, and 4. It may be noted that the Log likelihood value is smaller for $k = 4$ (the results for $k = 5$, 6 etc. are not better than that for $k = 4$ and hence are not given here). The four components Poisson Mixture model is given in table 6.2. It may be noted that 58% of wards may have higher incidence/relative risk and the remaining wards have lesser/lower incidence for the Cancer disease. We computed the posterior probability for each component for each ward (see table 6.3). Each ward is assigned to a particular component so that the posterior probability is larger. These results are also given in table 6.3 Finally we present Choropleth maps based on those results.

Table 6.1: NPMLE, MLE for Lower Components Models, Corresponding Log-Likelihoods

(1)	(2)	(3)	(4)
k=2	0.6740 0.3260	0.8070 1.372	-570.695
k=3	0.0489 0.6735 0.27751	0.2504 0.8565 1.4158	-554.969
k=4	0.0489 0.6735 0.27751	0.2504 0.8565 1.4158	-554.969
k=5	0.0405 0.3795 0.3851 0.1949	0.2243 0.7521 1.0263 1.5039	-552.369

Column number (1): Number of Components, (2): Weights, (3): Parameters, (4): Log-likelihood values.

Table 6.2: Results of Fitting Mixed Model to Disease Mapping of Cancer.

Parameter	Values			
Mean SMR	$\lambda_1 = 0.2243$	$\lambda_2 = 0.7521$	$\lambda_3 = 1.0263$	$\lambda_4 = 1.5039$
Weights	$P_1 = 0.0405$	$P_2 = 0.3795$	$P_3 = 0.3851$	$P_4 = 0.1949$

Table 6.3: Assigned Component for Each Ward

ward	component1	component2	component3	component4	Assigned Component
1	5.96852E-08	0.974314887	0.025684969	8.39137E-08	2
2	4.26867E-10	0.813849331	0.186128326	2.23428E-05	2
3	9.84979E-11	0.565572825	0.433836972	0.000590203	2
4	0.003923612	0.859551793	0.136057592	0.000467002	2
5	0	0.099436499	0.858713812	0.041849689	3
6	2.7948E-09	0.199158861	0.706605441	0.094235695	3
7	3.43587E-12	0.03906703	0.544780647	0.416152323	3
8	3.4775E-06	0.802044474	0.197655183	0.000296865	2
9	2.97071E-08	0.230266186	0.668296924	0.101436861	3
10	2.959E-13	0.176723251	0.805524859	0.01775189	3
11	1.17262E-10	0.199149111	0.750785403	0.050065486	3
12	0.211897019	0.723558639	0.064364381	0.00017996	2
13	9.81608E-05	0.722626049	0.274592473	0.002683317	2
14	7.79011E-09	0.510589668	0.485954311	0.003456013	2
15	0.01850261	0.892960813	0.088379514	0.000157063	2
16	0.000122868	0.79392271	0.205082432	0.00087199	2
17	1.03557E-05	0.683778056	0.313578783	0.002632805	3
18	8.46945E-07	0.29527668	0.608391571	0.096330902	3
19	0.220704448	0.716984099	0.062145379	0.000166074	2
20	0.005823329	0.803518004	0.188881422	0.001777245	2
21	0.770432961	0.222066838	0.007495655	4.54557E-06	1
22	3.72489E-10	0.234595439	0.725245947	0.040158614	3
23	1.23746E-10	0.041546141	0.454074183	0.504379676	4
24	1.35E-09	0.0911653	0.594181376	0.314653322	3
25	2.12169E-12	0.01527584	0.317221875	0.667502286	4

26	0.00023535	0.796747124	0.202041588	0.000975938	2
27	0.002915848	0.79979193	0.195611219	0.001681002	2
28	1.91002E-12	0.055193182	0.662112263	0.282694555	3
29	0.000109163	0.609248494	0.378641968	0.012000375	2
30	0.183619896	0.743821459	0.07232254	0.000236106	2
31	8.07192E-09	0.331692692	0.642824806	0.025482494	3
32	0	0.000210724	0.105615763	0.894173513	4
33	0	0.069257292	0.831661534	0.099081174	3
34	5.65259E-09	0.667276151	0.332250549	0.000473294	2
35	6.338E-10	0.227841432	0.723002425	0.049156143	3
36	1.55584E-06	0.899718702	0.100259801	1.99414E-05	2
37	0	0.013811468	0.383600116	0.602588416	4
38	3.51105E-05	0.926144133	0.073805247	1.55096E-05	2
39	2.00032E-06	0.757675806	0.24175276	0.000569434	2
40	0	0.000611477	0.233870565	0.765517958	4
41	0.00114724	0.775876527	0.220868737	0.002107496	2
42	2.53085E-09	0.236789509	0.704321484	0.058889004	3
43	0.858404231	0.133121547	0.00844936	2.48613E-05	1
44	0.012693639	0.763286634	0.219957447	0.004062281	2
45	2.22199E-07	0.349117131	0.606673289	0.044209359	3
46	0.00135141	0.611395662	0.366574754	0.020678174	2
47	1.1301E-08	0.068779641	0.457007298	0.47421305	4
48	0.009118738	0.897123717	0.093602113	0.000155431	2
49	0.000146475	0.573458459	0.407286425	0.019108641	2
50	1.88539E-12	0.354235626	0.643332107	0.002432267	3

51	1.15271E-07	0.892419913	0.107567025	1.29463E-05	2
52	3.00278E-10	0.112201375	0.686310782	0.201487843	3
53	0	8.42678E-07	0.002460484	0.997538673	4
54	3.94707E-09	0.567790586	0.430710581	0.001498829	2
55	6.65516E-11	0.202966522	0.755095885	0.041937593	3
56	2.13037E-09	0.440241244	0.55428826	0.005470494	3
57	6.79598E-05	0.687588113	0.308323699	0.004020229	2
58	1.66647E-11	0.279935038	0.709411655	0.010653307	3
59	0.001307991	0.831272644	0.166683148	0.000736218	2
60	4.38734E-05	0.849979125	0.149774483	0.000202519	2
61	1.13072E-10	0.039748604	0.444953897	0.515297499	4
62	0	4.59413E-07	0.032312873	0.967686667	4
63	0	0.047877614	0.949486949	0.002635437	3
64	0	5.29026E-08	0.006937094	0.993062853	4
65	0	0.001730924	0.864336056	0.133933021	3
66	0	0.044135929	0.914443242	0.041420828	3
67	0	2.5159E-06	0.007841425	0.992156059	4
68	5.98053E-05	0.923390902	0.07652909	2.02023E-05	2
69	0	0.00341221	0.230495173	0.766092617	4
70	0	1.01008E-05	0.010578411	0.989411488	4
71	0.009506294	0.740574837	0.244214556	0.005704313	2
72	0	0.020688492	0.574148084	0.405163424	3
73	1.64149E-07	0.443519746	0.541018027	0.015462063	3
74	0	6.22773E-05	0.092900886	0.907036837	4
75	0	0.353958981	0.645543494	0.000497525	3

76	0.000282025	0.810283035	0.188649695	0.000785246	2
77	3.88701E-06	0.582955491	0.409799528	0.007241094	2
78	0	0.000327156	0.070394795	0.929278049	4
79	1.21468E-09	0.229331136	0.71529851	0.055370354	3
80	5.77805E-07	0.700808518	0.298202363	0.000988541	2
81	0.001996647	0.857391669	0.140177811	0.000433873	2
82	0.036066232	0.879861495	0.083908283	0.00016399	2
83	5.12304E-09	0.196714583	0.69523493	0.108050482	3
84	0.000250524	0.63171062	0.356806848	0.011232008	2
85	0.000822974	0.625705829	0.35764904	0.015822158	2
86	2.11742E-10	0.038398668	0.418578753	0.543022579	4
87	0.02756185	0.810076907	0.160809698	0.001551545	2
88	0.001094965	0.819173418	0.178815879	0.000915738	2
89	2.96835E-09	0.095609156	0.582541946	0.321848894	3
90	7.87435E-08	0.20610475	0.642108813	0.151786358	3
91	0.96595645	0.033526799	0.000516657	9.37927E-08	1
92	0.001017199	0.7659642	0.230593084	0.002425517	2
93	0	0.000205336	0.035300505	0.964494159	4
94	1.87065E-07	0.393630873	0.579416869	0.026952072	3
95	0.022879568	0.753991807	0.218356337	0.004772288	2
96	0.175409746	0.749418698	0.074914888	0.000256669	2
97	1.7288E-11	0.102694941	0.743042223	0.154262836	3
98	3.85193E-08	0.201684021	0.655900527	0.142415413	3
99	2.05968E-09	0.141777629	0.679031684	0.179190685	3
100	0	0.001342852	0.076160952	0.922496196	4

101	1.19154E-06	0.327547494	0.596337604	0.07611371	3
102	0.001333187	0.609903144	0.367805519	0.020958151	2
103	0	0.000885074	0.073858838	0.925256088	4
104	0	0.011447514	0.38319398	0.605358505	4
105	2.24952E-12	0.022656783	0.407684644	0.569658573	4
106	1.35422E-05	0.649181843	0.346363374	0.004441241	2
107	2.85502E-11	0.12357832	0.75828336	0.11813832	3
108	0.009697704	0.9520807	0.038214156	7.43982E-06	2
109	2.83284E-08	0.332471814	0.633907497	0.03362066	3
110	0	0.002703372	0.130478455	0.866818174	4
111	1.38653E-07	0.364162872	0.60161988	0.034217109	3
112	1.37663E-05	0.518802712	0.461355038	0.019828484	2
113	2.3708E-06	0.595803216	0.398684095	0.005510319	2
114	2.13364E-09	0.503334164	0.493973555	0.002692279	2
115	7.54733E-05	0.803456395	0.195826778	0.000641353	2
116	2.27901E-11	0.113840725	0.752630454	0.133528821	3
117	0	8.75601E-05	0.043673265	0.956239175	4
118	5.54996E-05	0.863947232	0.135846642	0.000150626	2
119	8.37428E-13	0.038306611	0.583687766	0.378005623	3
120	3.92362E-10	0.285666061	0.692902659	0.02143128	3
121	4.09629E-13	0.050688312	0.686384793	0.262926896	3
122	0	1.72724E-07	0.001706879	0.998292949	4
123	1.4657E-05	0.657227029	0.338670982	0.004087332	2
124	6.94716E-05	0.689740768	0.306264174	0.003925586	2
125	1.32957E-11	0.147680035	0.783492413	0.068827552	3

126	3.26497E-08	0.289461916	0.655497189	0.055040862	3
127	7.41851E-08	0.425564593	0.55894814	0.015487193	3
128	0	3.9523E-05	0.066347989	0.933612488	4
129	3.88619E-07	0.918017289	0.081975459	6.86416E-06	2
130	3.87443E-12	0.421895419	0.576775002	0.001329579	3
131	0	0.067839703	0.91541593	0.016744366	3
132	2.61398E-13	0.170081153	0.810611422	0.019307425	3
133	0	0.003168346	0.198651484	0.79818017	4
134	0	0.021761049	0.520464841	0.45777411	3
135	0	0.141175795	0.840556376	0.01826783	3
136	9.86191E-07	0.568009172	0.42587342	0.006116421	2
137	1.33287E-06	0.534362235	0.45600915	0.009627282	2
138	7.1186E-09	0.628980643	0.370191402	0.000827948	2
139	2.08721E-08	0.549761112	0.447397072	0.002841795	2
140	4.94111E-10	0.304570946	0.677413634	0.01801542	3
141	0	0.173622307	0.819030615	0.007347078	3
142	1.74329E-06	0.627870671	0.368668847	0.003458738	2
143	0.990913437	0.008910009	0.000176474	8.01087E-08	1
144	0.705897393	0.282764857	0.01132849	9.25968E-06	1
145	0.035750861	0.87979931	0.084283798	0.000166031	2
146	0.990624963	0.009302455	7.25779E-05	4.02542E-09	1
147	0	0.000252222	0.040153093	0.959594685	4
148	2.71793E-12	0.034980231	0.520296801	0.444722967	3
149	1.30905E-09	0.39230587	0.599406401	0.008287728	3
150	8.59621E-13	0.110844151	0.811736328	0.077419521	3
151	0	2.62799E-07	0.002980735	0.997019002	4
152	0	0.003928278	0.237005031	0.759066691	4
153	0	0.883754608	0.116245372	1.98452E-08	2
154	0	0.118938304	0.858800565	0.022261131	3
155	0	0.000707205	0.882859974	0.116432821	2

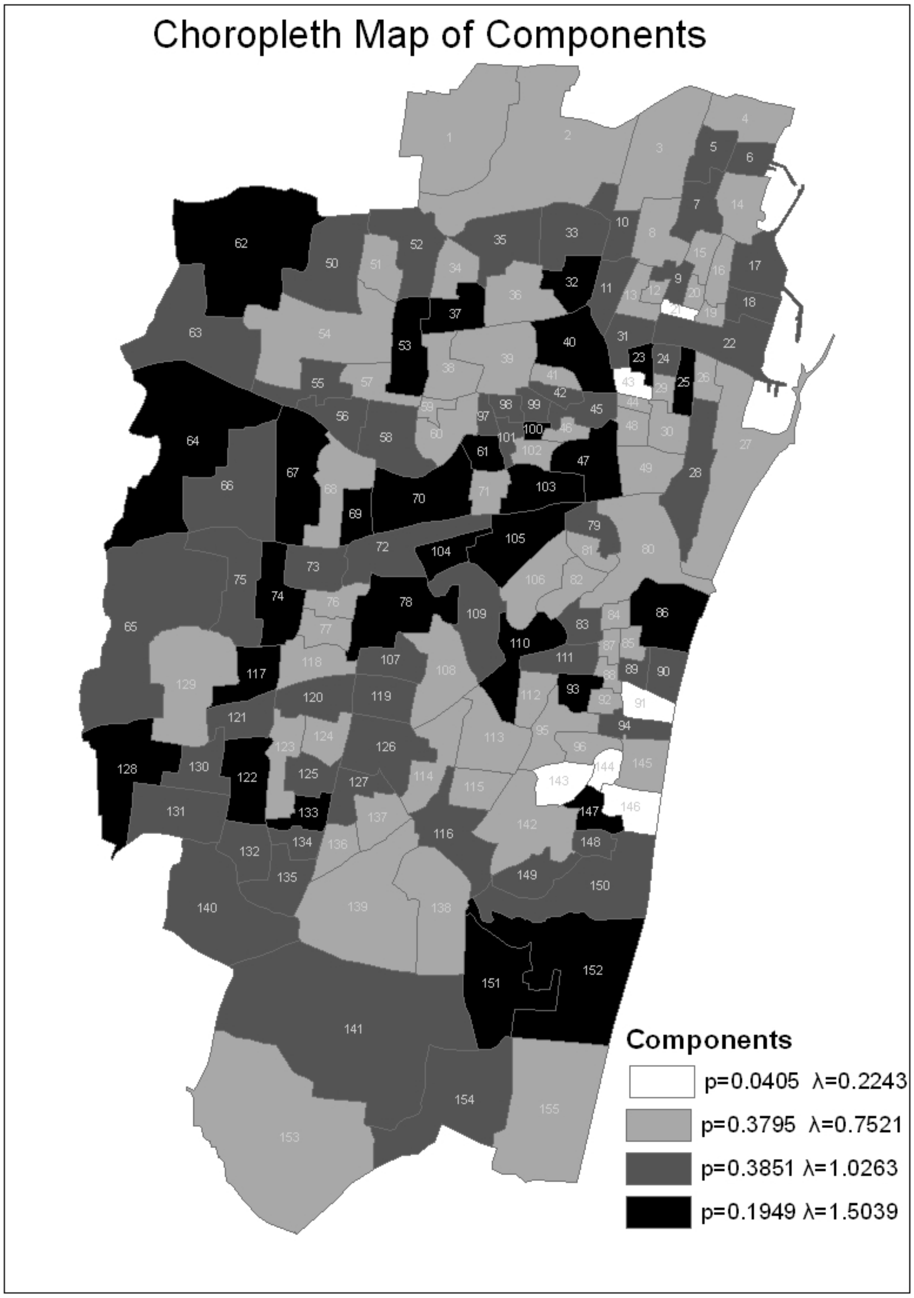

Figure 6.1: Choropleth Map of Components of Mixture Distribution