Proceedings of the Statistics Canada
Symposium on Analysis of Data in Time
October 1989

MAPPING AGGREGATE BIRTH DATA

David R. Brillinger [1]

ABSTRACT

Births by census division are studied via maps for the province of Saskatchewan for the year 1986. A principal goal of the work is to see how births are related to geography by obtaining contour maps displaying the birth phenomenon in a smooth fashion. A hierarchy of models for count-valued random variates are fit to the data by maximum likelihood. Models include: the Poisson, the Poisson with a weekday effect and the Poisson-lognormal. The last mentioned is motivated by the idea that important covariates are unavailable to include in the analysis.

KEY WORDS: Aggregate data; Contouring; Extra-Poisson variation; Locally-weighted analysis; Maps; Poisson distribution; Poisson-lognormal distribution; Random effects; Spatial data; Unmeasured covariates.

1. INTRODUCTION

The concern of this paper is data that has been aggregated over geographical regions. The analysis of such data should be "easy" because of the graphing possibilities, eg. quantity versus geography in the manner of residual plots so often employed in regression analyis; however in the present case the aggregation leads to important difficulties.

The specific data studied consists of daily births for the calendar year 1986 to women aged 25-29 for each of the 18 census divisions of the province of Saskatchewan. The corresponding population sizes, as determined in the 1986 Census, are also employed in order to compute rates. The reason that Saskatchewan was selected for this pilot study is that it is moderate sized and its boundaries and those of its census divisions are regular. (The latter was important at the early stages of the work because computer based maps were unavailable.) Women ages 25-29 were selected because that was the 5 year age group with most births. These data were provided to the author by Statistics Canada.

The data is characterized by being aggregate, by being nonGaussian and by being nonstationary in space and time.

It is wished to understand the relationship of births to geography, specifically to allow spatial patterns of fertility and possible surprises to show themselves. There are two aspects to the study; a locally-weighted analysis of grouped data is developed and random effects models are set down and fit to handle extra-Poisson variation.

It is to be emphasized that this is a preliminary report on work in progress. For example the fine structure of the data is not taken advantage of and no measures of uncertainty of the various estimates have been provided. The paper focuses principally on annual totals for the 18 census divisions. The related paper Brillinger (1990) considers both temporal and spatial aspects.

Saskatchewan has 18 census divisions. These may be seen in Figure 1. That figure also provides the total numbers of births to women aged 25 to 29 for 1986 and the corresponding female population sizes on Census Day, 3 June. (Actually because of Statistics Canada's confidentiality requirements the final digits have been rounded to the nearer of 2 and 7). The small population in the northern half of the province is evident. Figure 2 gives the annual birth rates plotted by census division. The divisions with the lowest values, .131 and .133 births per year, correspond to the cities of Saskatoon and Regina respectively. Figure 3 is a chloropleth map of the rates with intensity of hatching proportional to birth rate.

2. PATCH OR CHLOROPLETH MAPS

Maps of most quantities of direct interest that assign average values to the wholes of counties thereby lie, lie, lie.

In these graphic words Tukey (1979) deplores the use of maps such as those of Figures 2, 3 that are constant across geographic divisions. Indeed examination of Figure 2, as does common knowledge, suggests that the birth phenomenon quite likely varies smoothly across census division boundaries. One of the concerns of this work is to develop maps with smooth variation. It is hoped that such maps will prove useful in the discovery of general models and will allow insightful exploratory analyses.

A second concern is with the statistical distribution of the counts themselves. A natural special stochastic model to employ is the Poisson. Yet the birth process has been found to relate to many socio-economic quantities, eg. diet,

lifestyle, weather, environment, weekday, holidays, age structure. Further the population of the province has varied around the Census Day values throughout 1986 and lastly the women's ages range between 25 and 29. In summary it seems necessary to employ a more flexible model than the Poisson, a model able to handle omitted covariates. The Poisson-lognormal will be employed in this work. As a sideline due to the presence of the standard deviation parameter in the Poisson-lognormal, there will be a borrowing of strength that takes place in combining the data values.

3. LOCALLY-WEIGHTED ANALYSIS

In the case of nonaggregate data, locally-weighted fitting is a convenient fashion by which to estimate smoothly varying quantities. Suppose one has a variate Y with probability distribution $p(Y \mid \theta)$ depending on the finite dimensional parameter θ. Suppose one wishes an estimate of θ particular to the location with coordinates (x,y). Suppose the datum Y_i is available for location (x_i,y_i). Let $W_i(x,y)$ be a weight dependent on the distance of (x_i,y_i) to (x,y).

Consider estimating θ by maximizing the weighted loglikelihood

$$\sum_i W_i(x,y) \log p(Y_i \mid \theta) \tag{1}$$

or (often equivalently) by solving the system of estimating equations

$$\sum_i W_i(x,y)\, \psi(Y_i \mid \hat{\theta}) = 0 \tag{2}$$

with $\psi(Y \mid \theta) = \partial \log p / \partial\theta$, the score function.

To illustrate the technique consider an elementary case, specifically take Y to be normal with mean μ and variance σ^2. The locally weighted estimate of μ results from minimizing

$$\sum_i W_i(x,y)\,[Y_i - \mu]^2$$

and is given by

$$\hat{\mu}(x,y) = \sum_i W_i(x,y)\, Y_i \,/\, \sum_i W_i(x,y)$$

an expression with intuitive appeal. It is to be noted that such formulas are commonly used in computer graphics as interpolation procedures, see for example Franke (1982).

Among references we may mention Gilchrist (1967) concerned with "discounting", Pelto *et al.* (1968), concerned with least squares, Cleveland and Kleiner (1975), who suggested the use of moving midmeans and Stone (1977) focusing on regression. In the discussion of Stone's paper, Brillinger (1977) suggested the form (2) for a general distribution and justified it as a Bayes' rule. Cleveland and Devlin (1988) develop the least squares approach in real detail. Tibshirani and Hastie (1987) develop an equi-weighted local likelihood estimation procedure. Staniswalis (1989) studies and implements the general p case. Advantages of the locally-weighted technique include: no "hidden" model distribution assumption, the possibility of discerning nonaditivity, variants for resistance and influence, simple additivity of the observation component, and no matrix inversion (as, for example, kriging requires).

4. CONSTRUCTION OF THE WEIGHTS

The birth data of concern in this work is aggregate (or grouped) totals over census divisions. The procedure of the preceding section cannot therefore be employed directly. The problem is that of obtaining appropriate weights $w_i(x,y)$ evidencing the effect of the census division i on the location (x,y). Suppose $\mid R_i \mid$ denotes the area of census division i. Then the naive weight function is

$$w_i(x,y) = 1/\mid R_i \mid \qquad \text{for } (x,y) \text{ in } R_i$$

and equal 0 otherwise. In this work functions of the essential form

$$w_i(x,y) = \frac{1}{\mid R_i \mid} \int_{R_i} W(x-u, y-v)\,du\,dv \tag{3}$$

will be employed where $W(.)$ is a kernel appropriate for the nonaggregate case as studied in Cleveland and Devlin (1988). The formula (3) may be motivated by consideration of the Poisson point process case. Estimates will be determined via the criteria (1) or (2) with W_i replaced by w_i.

The specific weights employed at $\mathbf{r} = (x,y)$ are

$$w_i(\mathbf{r}) = \exp\{-(1-\rho)^2 \mid\mid \mathbf{r} - \mathbf{r}_i \mid\mid^2 / 2\tau^2\} \tag{4}$$

outside the ellipse $(\mathbf{r}_0 - \bar{\mathbf{r}}_i)S^{-1}(\mathbf{r}_0 - \mathbf{r}_i)' = d_0^2 = 5.991$ and equal 1 inside. Here $\mid\mid \mathbf{r} \mid\mid^2 = x^2 + y^2$, $\rho = d_0 / \sqrt{(\mathbf{r} - \bar{\mathbf{r}}_i)S_i^{-1}(\mathbf{r} - \bar{\mathbf{r}}_i)'}$ and $\tau = .025$, where $\bar{\mathbf{r}}_i = E\ U_i$ and $\mathbf{S}_i = var\ U_i$ with U_i a variate uniformly distributed within R_i. The logic is that the census divisions are approximated by ellipses with the same mean and variance-covariance matrix. (The specific values were chosen after a bit of experimentation, in part to make the area in the initial ellipse about .95 of the division's.)

Figure 4 displays the .50 and .99 contours of the $w_i(x,y)$ plotted for several of the census divisions. The contours are seen to follow the general shapes of the census divisions.

Other weight functions constructed with similar problems in mind may be found in Tobler (1979) and Dun and

additive and do not interact, no matrix inversion is needed, and resistance to outliers is easily built in.

Cliff and Ord (1975) Section 5.1, discusses measures of the influence of counties on each other. The concern of this present paper is the influence of a "county" on a point location.

5. THE SIMPLE POISSON

Throughout the analysis, the female population aged 25-29 and births to its members will be considered. Let $i = 1,..., 18$ index census division. Let N_i denote the census count of the women in the i-th division. (These are the counts for Census Day, 3 June 1986.) Let B_i denote the total number of births to women aged 25-29 in the year 1986.

Suppose that the probability distribution $p(.)$ of Section 3 is that B_i is Poisson with mean $N_i\mu$. The parameter μ is a birth rate. One logic for the Poisson assumption comes from the idea that birthdays are random, see Brillinger (1986).

With the Poisson assumption, the locally weighted estimate of the birth rate at location (x,y) is

$$\hat{\mu}(x,y) = \sum_i w_i(x,y)\, B_i \,/\, \sum_i w_i(x,y)\, N_i \tag{5}$$

These values are computed for (x,y) on a 40 by 40 grid. The corresponding contour plot is given in Figure 5. The contours are seen to vary smoothly. This (smoothed) rate varies from .14 to .20, with the higher values in the upper half of the province and the lower centred around the most urban part of the province.

6. THE POISSON WITH WEEKDAY EFFECTS

While the focus of this paper is on spatial analysis, it is usefult to briefly take some definite note of the temporal aspects that are present. It is common knowledge that birth rates vary with the day of the week due to medical intervention, see for example Miyaokoa (1989). The total number of births cannot therefore be reasonably expected to be a homogeneous Poisson. The following model seems worth considering. Let j be an indicator variable with $j = 1$ if the measurement is for a weekday and $j = 2$ if the measurement is for a weekend. Let B_{ij} denote the corresponding number of births in census division i. Suppose that B_{ij} is Poisson with mean $N_i \exp\{\alpha+\beta_j\}$. β_j is the weekday effect and it will be assumed that $\beta_1 + \beta_2 = 0$ to make the model identifiable. If there is no weekday effect, then $\beta_1, \beta_2 = 0$. Now, via locally-weighted estimation as described in Sections 3 and 4, one can obtain estimates of α and β as functions of location.

Figure 6 provides the estimate $\exp\{\hat{\alpha}(x,y)\}$ obtained of the annual birth rate. It is interesting to note that, relative to the constant Poisson model, the contours have expanded out from the urban area for the annual rate. Figure 7 provides the estimated weekday effect $\hat{\beta}_1(x,y)$. In its case there is bulge to the east. The order of magnitude of the $\hat{\beta}$'s is .00 to .10 while $\hat{\alpha}$ is order -2.0 to -1.6 .

The just preceding analysis suggests that there are basic variables that can affect birth rates and that modelling and analysis needs to take this circumstance into account.

7. THE POISSON-LOGNORMAL

With a multi-dimensional explanatory variable $\mathbf{x}_i$ in hand, a Poisson model that has B_i of mean $N_i \exp\{\mathbf{x}_i\theta\}$ might do a good job of explaining the data. Examples of explanatory variables include: diet, lifestyle, weather, environment, holidays, population change, age structure, vagaries of boundaries. In the present situation, these variables are not at hand. The omitted variables in the model will be assumed specifically accumulated into an error variable. It will be assumed that, given ε_i, the variate B_i is Poisson with mean $N_i\,\mu\exp\{\varepsilon_i\}$ and that ε_i is normal with mean 0 and variance σ^2. Here B is said to have a Poisson lognormal distribution. Some information on this distribution may be found in Shaban (1988).

A central difficulty, that arises in working with a Poisson-lognormal model, is that closed expressions do not exist for the probabilty function. Yet it is clearly flexible for introducing effects and handling missing variables. Following the work of Bock and Lieberman (1970) and Pierce and Sands (1975) however, one can proceed via numerical integration. The probability function may be written

$$p(y) = \frac{1}{y!}\int (\nu e^{\sigma z})^y \exp\{-\nu e^{\sigma z}\}\phi(z)dz$$

with ϕ the standard normal density, with y corresponding to B and with ν corresponding to $N\mu$. The integral is approximated by a finite number of terms involving nodes and weights.

Figures 8 and 9 provides the result of fitting employing 61 nodes. Figure 8 again shows a dip around the urban region as in Figures 5 and 6. The irregularities suggest that perhaps the estimation procedure converged to a local extremum. Figure 9 is not easily described. It suggests that the estimate is fairly variable. The estimate σ is seen to be of order of magnitude .1 and so comparable to the weekday effect of Section 6.

8. DISCUSSION

Locally-weighted analysis and random effect models appear to provide a flexible means of dealing with a broad class of problems involving geographic data. The random effect terms have two important roles: handling omitted effects and borrowing strength for improved estimates of the principal parameters. For the Poisson alone, naive totals are efficient, yet there exists extra-Poisson variability due to ommited variables in the present case. The approach is computer intensive, because of the numerical integration and the maximum likelihood estimation at many points on a grid, but proved quite mangeable on the Berkeley network of Sun 3/50's.

Much future work remains including: tools for assessing fit, uncertainty computation, weight function choice (including choice of τ in (4)), analyses for other age groups and provinces, and appropriate asymptotics. Some further results are provided in Brillinger (1990).

Other recent papers devoted to the analysis of vital statistics rates are: Clayton and Kaldor (1987), Tsutakawa (1988) and Manton *et al.* (1989). These papers are not directed at the problem of obtaining a smooth surface, which is the concern of this work.

After the analyses were completed it was learned that the birth counts were based on 1981 census divisions, while the population counts were based on 1986. The boundaries have not changed much, but this provides even more reason for wanting a procedure that can handle extra-variation.

ACKNOWLEDGEMENTS

The author would like to thank G. Brackstone, R. Gussela, R. Raby, B. Sander, P. Spector, M. Subhani, R. Villani for assistance in obtaining the data and maps, for help with computational geometry and with parallel computing. The research was supported by National Science Foundation Grant DMS-8900613.

REFERENCES

Bock, R. D. and Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. Psychometrika 35, 179-197.

Brillinger, D. R. (1977). Discussion of Stone (1977). Ann. Statist. 5, 622-623.

Brillinger, D. R. (1986). The natural variability of vital rates and associated statistics. Biometrics 42, 693-734.

Brillinger, D. R. (1990). Spatial-temporal modelling of spatially aggregate birth data. Tech. Report, Statistics Dept., University of California, Berkeley.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43, 671-681.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. J. Amer. Statist. Assoc. 83, 596-610.

Cleveland, W. S. and Kleiner, B. (1975). A graphical technique for enhancing scatterplots with moving statistics. Technometrics 17, 447-454.

Cliff, A. D. and Ord, J. K. (1975). Model building and the analysis of spatial pattern in human geography. J. Royal Stat. Soc. 37, 297-348.

Dyn, N. and Wahba, G. (1982). On the estimation of functions of several variables from aggregated data. SIAM J. Math. Anal. 13, 134-152.

Franke, R. (1982). Scattered data interpolation: tests of some methods. Math. Comp. 38, 181-200.

Gilchrist, W. G. (1967). Methods of estimation involving discounting. J. Royal. Stat. Soc. 29, 355-369.

Manton, K. G., Woodbury, M. A., Stallard, E., Riggan, W. B., Creason, J. P. and Pelom, A. C. (1989). Empirical Bayes procedures for stabilizing maps of U. S. cancer mortality rates. J. Amer. Statist. Assoc. 84, 637-650.

Miyaoka, E. (1989). Application of mixed Poisson-process models to some Canadian birth data. Canadian J. Stat. 17, 123-140.

Pelto, C. R., Elkins, T.A. and Boyd, H.A. (1968). Automatic contouring of irregularly spaced data. Geophysics 33, 424-430.

Pierce, D. A. and Sands, B. R. (1975). Extra-binomial variation in binary data. Technical Report 46, Statistics Dept., Oregon State University.

Preparata, F. P. and Shamos, I. (1985). Computational Geometry. Springer, New York.

Shaban, S. A. (1988). Poisson log-normal distributions. Pp. 195-210 in Lognormal Distributions (eds. E. L. Crow and K. Shimizu). M. Dekker, New York.

Staniswalis, J. G. (1989). The kernel estimate of a regression function in likelihood-based models. J. Amer. Statist. Assoc. 84, 276-283.

Stone, C. J. (1977). Consistent nonparametric regression. Ann. Statist. 5, 595-620.

Tibshirani, R. and Hastie, T. 91987). Local likelihood estimation. J. Amer. Statist. Assoc. 82, 559-567.

Tobler, W. R. (1979). Smooth pycnophylactic interpolation for geographical regions. J. Amer. Statist. Assoc. 74, 519-536.

Tsutakawa, R. K. (1988). Mixed model for analyzing geographic variability in mortality rates. J. Amer. Statist. Assoc. 83, 37-42.

Tukey, J. W. (1979). Statistical mapping: what should not be plotted. Proc. 1976 Workshop on Automated Cartography. DHEW Publication No. (PHS) 79-1254, 18-26. Included in The Collected Works of J. W. Tukey, Vol. 5 (1988), (Ed. W. S. Cleveland). Wadsworth, Pacific Grove.

APPENDIX

In this Appendix a few computing details are provided. The census divisions and the province boundaries are specified as polygons. To compute the weights $w_i(x,y)$ a routine was required to check whether a given point was inside a given polygon. To compute the mean and variance of a random point inside a given polygon, a procedure breaking the polygon up into triangles was required. Such routines are discussed in Preparata and Shamos (1985). The likelihood was maximized via the Harwell FORTRAN routine va09a. For the parallel computations the 40 by 40 grid was broken up into 20 disjoint segments.

FIGURE LEGENDS

Figure 1. Births for the 18 census divisions of Saskatchewan for the year 1986 to women in the 25-29 age group and corresponding total numbers of women in that age group on June 3 of the year. (As discussed in the text, the final digits of counts have been rounded to the nearer of 2 and 7.)

Figure 2. Annual birth rates for the 18 census divisions for women aged 25 to 29.

Figure 3. The rates of Figure 2 displayed via intensity of hatching.

Figure 4. The weights, $W_i(x,y)$ applied in equations (1) or (2) computed via expression (4) for four of the census divisions. They are not shown for all the divisions in the interests of clarity.

Figure 5. Expression (5) graphed for the weights of (4) with B_i the count of births in census division i and N_i the corresponding population count of women aged 25-29.

Figure 6. The estimated birth rate assuming that the number of births, B, given the population at risk, N, is Poisson with mean $N \exp\{\alpha \pm \beta\}$ with the plus sign for weekdays and minus for weekends. Local weighted fitting is carried out to obtain the estimate $\exp\{\hat{\alpha}(x,y)\}$.

Figure 7. Plot of the estimated weekday effect $\hat{\beta}(x,y)$ obtained as per Figure 6.

Figure 8. A plot comparable to Figure 6, except that now a normal error term is added to the linear predictor.

Figure 9. A plot comparable to Figure 7, except now (as in Figure 8) a normal error term has been added to the linear predictor.

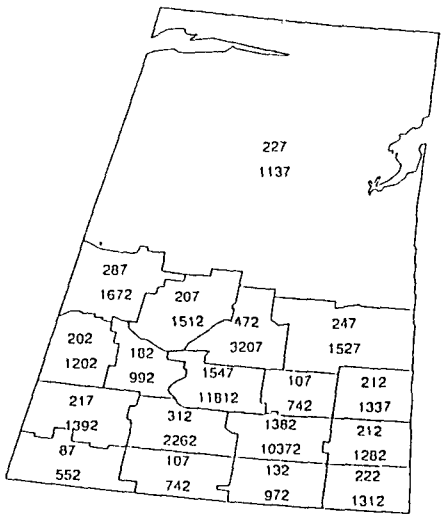


Figure 1

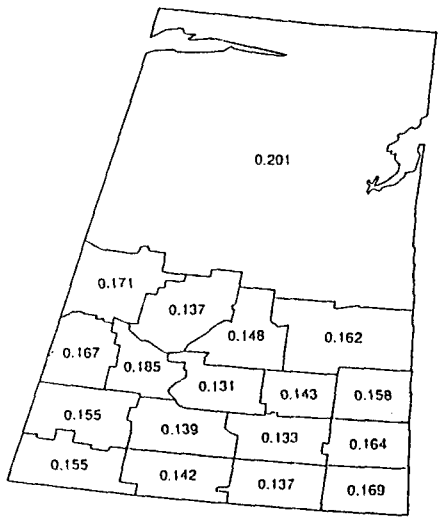


Figure 2

Annual birth rates

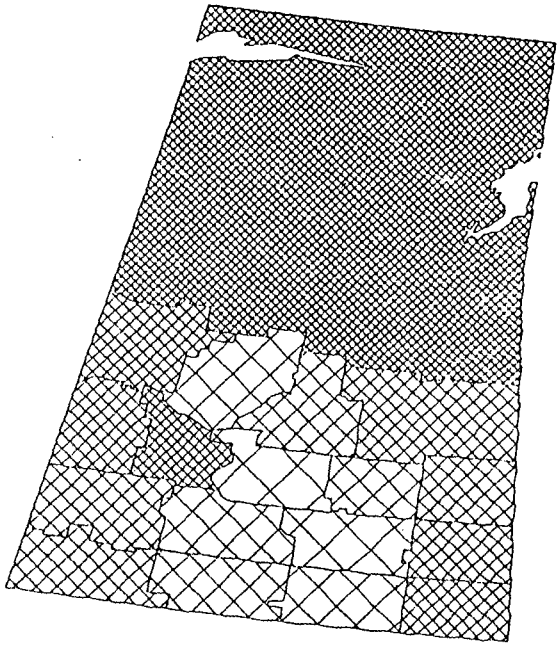


Figure 3

Census division weights

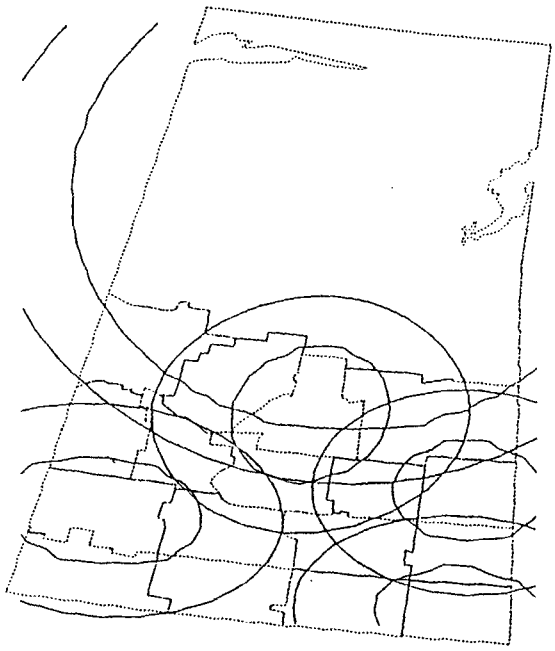
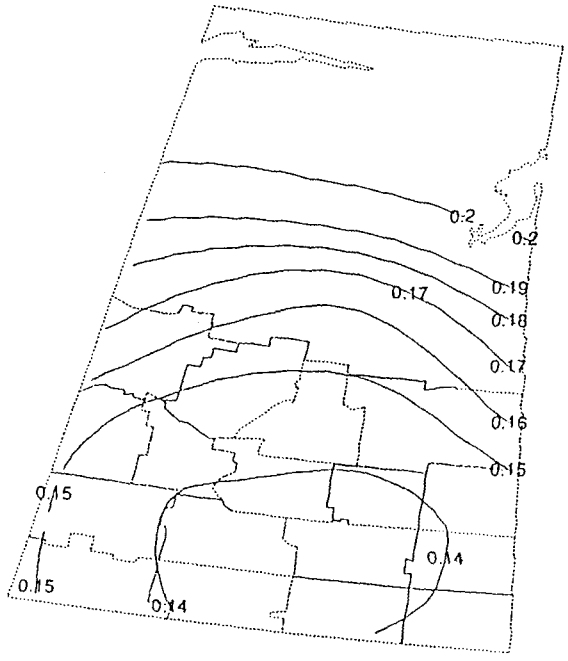


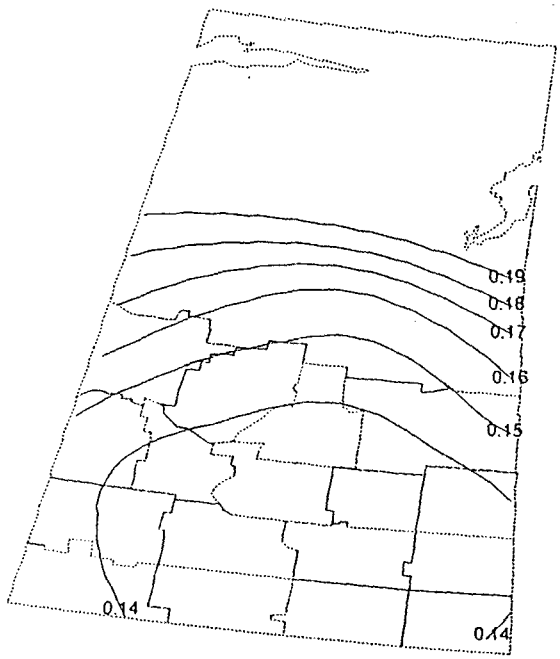
Figure 4

Annual birth rates



Simple Poisson
Figure 5

Annual birth rates

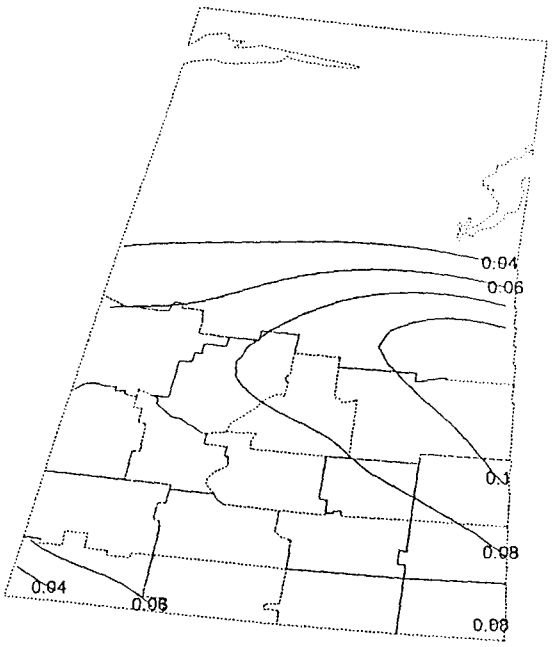


Poisson with weekday effect
Figure 6

Weekday effects

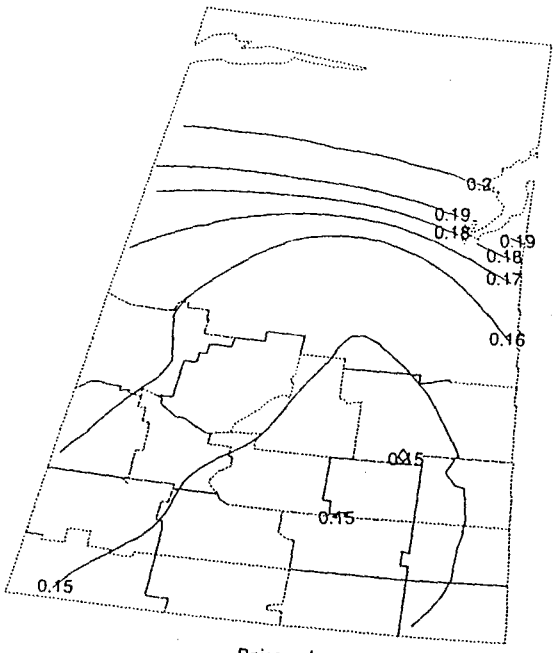
Annual birth rates

Weekday effects



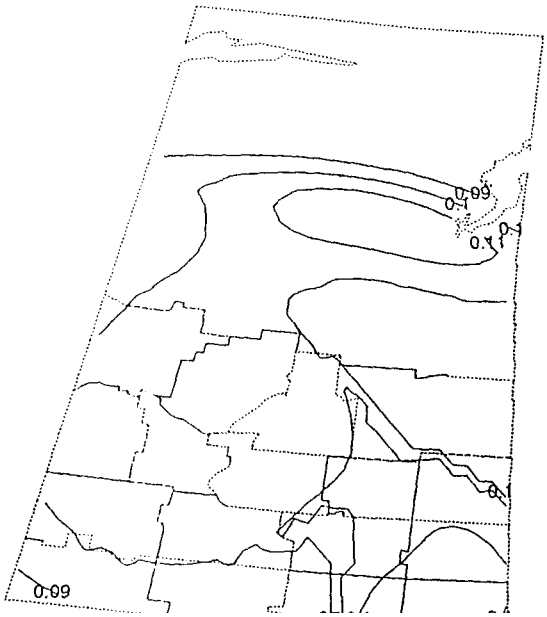
Poisson with weekday effect
Figure 7

Annual birth rates

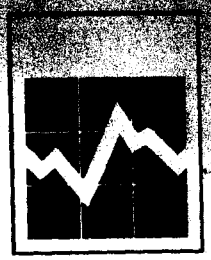


Poisson-lognormal
Figure 8

Sigma estimates



Poisson-lognormal
Figure 9



ANALYSIS OF DATA IN TIME

PROCEEDINGS OF THE 1989 INTERNATIONAL SYMPOSIUM

Edited by
A.C. Singh and P. Whitridge

Price: $20.00 Canadian

Payable to

"The Receiver General for Canada - Symposium '89 Proceedings"

Send your request to:

Statistics Canada Symposium '89 Proceedings
Social Survey Methods Division
R.H. Coats Building
Tunney's Pasture
Ottawa, Ontario
Canada
K1A 0T6