Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

LOCALLY WEIGHTED ANALYSIS OF SPATIALLY AGGREGATE BIRTH DATA: UNCERTAINTY ESTIMATION AND DISPLAY

D.R. Brillinger[1]

ABSTRACT

The concern is the analysis and display of data that is aggregate over geographic regions (such as census divisions) for phenomena that are felt to vary smoothly in space. In Brillinger (1990b) some spatial locally weighted analyses were provided for births taking place to women aged 25 to 29 in the years 1986 and 1987 for the province of Saskatchewan at the census division level. Various results were displayed via contour plots. The present work develops these ideas further and is, in particular, concerned with the computation and display of appropriate uncertainty levels for contour plots. A weekday effect is noted, but it does not vary appreciably with space. The approach provides an alternative to the empirical Bayes methods that have been proposed for similar problems.

KEY WORDS: Aggregate data; Binomial-logitnormal distribution; Contouring; Extra-variation; Interpolation; Locally weighted analysis; Logitnormal distribution; Maps; Simulation; Spatial data; Uncertainty presentation; Unmeasured covariates.

1. INTRODUCTION

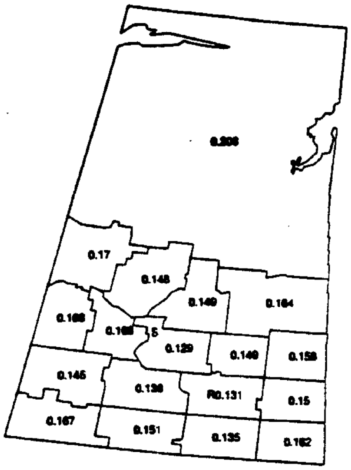
The concern is with analyzing smoothly varying spatial phenomena and with the provision of some indication of the uncertainty of the results of the analysis.

The data studied is spatially aggregate, representing totals over census divisions. The province of concern is Saskatchewan with 18 census divisions. Census division aggregate daily totals of births to women aged 25-29 for the two year period 1986-1987 are available for analysis. Also available are the census population totals for Census Day, 3 June 1986. It is desired to study and display the spatial structure of this data.

Figure 1 shows the Saskatchewan census divisions and the observed annual birth rate for each division. The letters R and S indicate the locations of the cities of Regina and Saskatoon respectively. The rates are estimated directly from the counts for the two years. More details concerning this data set may be found in Brillinger (1990a,b).

Figure 1

Saskatchewan: 1986-87 births, ages 25 to 29



annual birth rate

[1] D.R. Brillinger, Statistics Department, University of California, Berkeley, CA., U.S.A.

2. WEIGHTS AND ESTIMATION

The goal is spatial analysis and display, by means of contour plots. It is usual, in preparing contours from function values at scattered points, to first interpolate the values to a regular grid, see Pelto *et al.* (1968). Given (x_i,y_i,z_i), $i=1,...,n$ with z_i representing the value of a variate measured at location (x_i,y_i) a variety of methods have been proposed for this interpolation. A popular scheme is due to Shepard (1968). He computes z at (x,y) via:

$$z(x,y) = \sum_{i=1}^{n} w_i(x,y)\, z_i \Big/ \sum_{i=1}^{n} w_i(x,y), \tag{2.1}$$

where $w_i(x,y) = [(x-x_i)^2 + (y-y_i)^2]^{-\mu}$ with $\mu > 0$. The points are weighted inversely with their distance to (x,y). Other schemes are described in Franke (1982) and Sabin (1985).

In the present case, sampling fluctuations are present and one cannot simply interpolate. Further the data are counts, and proportions are computed from these, so the fluctuations are not elementary. Locally weighted likelihood analysis is a pertinent estimation technique for nonelementary distributions varying in space, see for example: Gilchrist (1967), Brillinger (1977), Tibshirani and Hastie (1987), Cleveland and Devlin (1988), Staniswalis (1989), Brillinger (1990a,b). Suppose a variate Z has probability function $p(z \mid \theta)$ depending on an unknown parameter θ. Let $\psi(z \mid \theta)$ denote the score function, $\partial \log p / \partial \theta$, and chose $\hat{\theta}$, the estimate of θ at location (x,y), to satisfy

$$\sum_i w_i(x,y)\, \psi(z_i \mid \hat{\theta}) = 0, \tag{2.2}$$

for some weight function $w_i(x,y)$. As in Shepard's method, $w_i(x,y)$ depends on the distance of the point (x_i,y_i) to the location (x,y).

As a simple example of a locally weighted estimate, consider the case of B_i binomial with parameters π, N_i. One computes directly the estimate

$$\hat{\pi}(x,y) = \frac{\sum_i w_i(x,y)\, B_i}{\sum_i w_i(x,y)\, N_i}. \tag{2.3}$$

This estimate is a natural extension of (2.1).

In the present case, where the data is aggregate over regions R_i, the choice for the weight is

$$w_i(x,y) = \frac{1}{|R_i|} \int_{R_i}\!\int W(x-u,y-v)\, dudv, \tag{2.4}$$

with $W(.)$ the biweight,

$$W(x,y) = (1-u^2)^2 \quad \text{for } |u| \le 1 \tag{2.5}$$

and equal 0 otherwise where $u = b\sqrt{x^2+y^2}$ for some $b > 0$. In (2.4) $|R_i|$ is the area of division i. One can view $w_i(x,y)$ here as representing the influence of census division i on a person at location (x,y), the influence resulting from items like travel, nutrition, climate, ethnicity, education, television, laws. These weights are evaluated via a Fourier transform, taking advantage of the convolutional form. A naive weight would be $w_i(x,y) = 1/|R_i|$ for (x,y) in R_i and $= 0$ otherwise. This corresponds to $W(.)$ a delta function.

Figure 2 shows the effect of varying the parameter b for Census Division 18, the northern half of the province. The values of b for the three cases illustrated in Figure 2 correspond to no smoothing, a small amount of smoothing, and a moderate amount. This last value is employed in the computations of the paper. Figure 3 gives a plot of (2.4) for all of 18 of the divisions.

Figure 2

Census division 18 - effect of smoothing

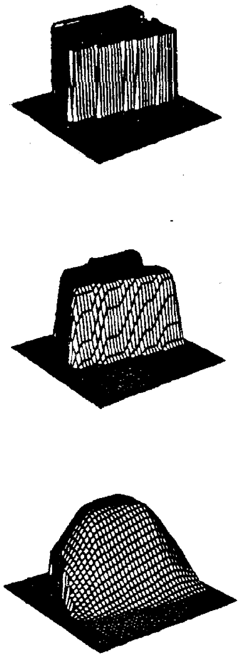
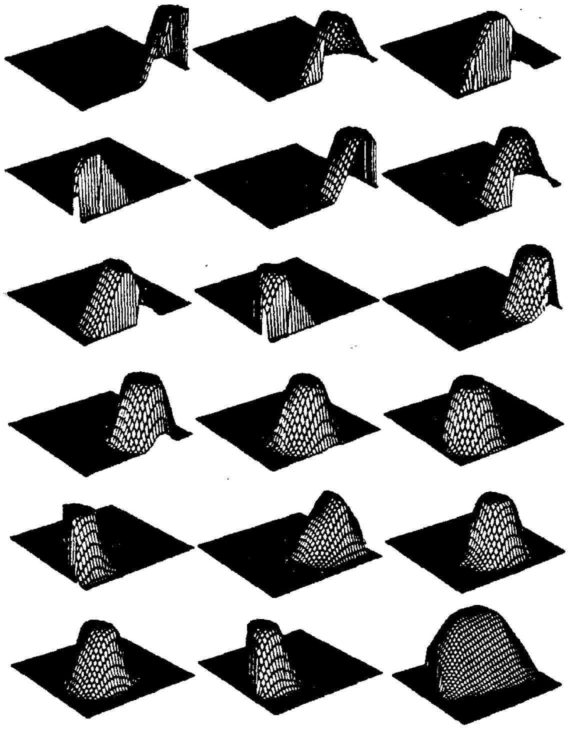


Figure 3



Alternate weights that might be employed are given in Tobler (1979) and Dyn and Wahba (1982). An advantage of the present weights is that they are local in character.

3. ESTIMATION: BINOMIAL CASE

Let B_{ijk} denote the number of births to women ages 25 to 29 in census division i and year j with $k = 1, 2$ depending on whether data is for a weekday or weekend. Let N_i denote the census population of census division i for the age group. The count B_{ijk} may be thought of as the number of births from this population. Its distribution may be approximated by a binomial. (This seems a better approximation than the Poisson used in Brillinger (1990a,b), since the chance that a woman has a baby in a year appears to be able to be as high as .2). Set $x_1 = 2$ and $x_2 = -5$. This will make the estimates orthogonal. It will be assumed specifically that B_{ijk} is binomial with parameters π_k and N_i where

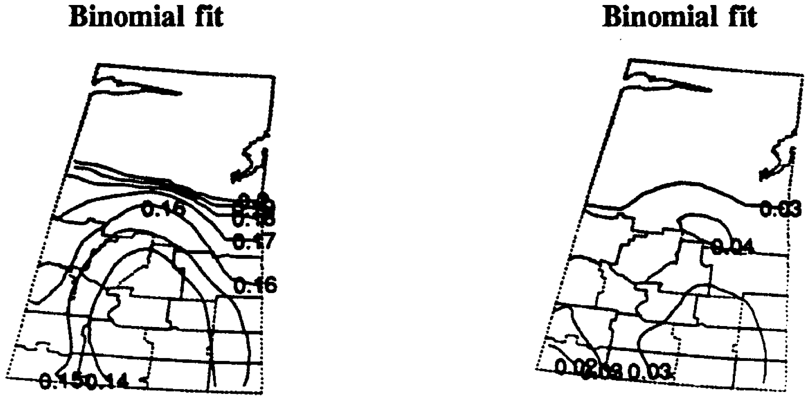
$$\mathit{logit}\ \pi_k = \log(\pi_k/(1-\pi_k)) = \alpha + \beta x_k. \tag{3.1}$$

The left-hand panel of Figure 4 provides contour plots of the annual birth rate estimated by the expression

$$\frac{5}{7}\frac{\exp\hat{\eta}_1}{1+\exp\hat{\eta}_1} + \frac{2}{7}\frac{\exp\hat{\eta}_2}{1+\exp\hat{\eta}_2}, \tag{3.2}$$

where $\hat{\eta}_k = \hat{\alpha} + \hat{\beta}x_k$, for $k = 1, 2$ and it is to be remembered that $\hat{\alpha} = \hat{\alpha}(x,y)$ and $\hat{\beta} = \hat{\beta}(x,y)$. The first term in (3.2) corresponds to weekdays, the second to weekends. One notes contours rising up and out from the census divisions including Regina and Saskatoon. The right-hand panel graphs the weekday effect estimate $\hat{\beta}(x,y)$. All contours are seen to be positive, corresponding to an increased number of births on weekdays. It must be remarked that these quantities are all subject to sampling fluctuations. The provision of an indication of sampling fluctuations will be discussed in Section 5.

Figure 4


Binomial fit

Binomial fit

annual rate

weekday effect

4. ESTIMATION: BINOMIAL-LOGITNORMAL CASE

It is argued in Brillinger (1990a,b) that a variety of pertinent explanatory variables, *e.g.* diet, lifestyle, weather, environment, holidays, age structure, urbanicity, will have gone unmeasured. This will lead to extravariation in the number of births beyond that of a binomial. One way to proceed is to introduce a random effect, σz, and replace (3.1) by

$$logit\ \pi_k = \alpha + \beta x_k + \sigma z \tag{4.1}$$

where z is a standard normal variate. The model is now binomiallogitnormal. The z's for the different census divisions are assumed independent. This model could be fit making use of numerical integration, see Bock and Lieberman (1970), Pierce and Sands (1975), Sanathanan and Blumenthal (1978), Brillinger (1990a,b) for example. In the present case the N_i are large and one can approximate the model by a logitnormal, *i.e.* by assuming the logits of the B_{ijk}/N_i are normal. Specifically this model may be written

$$\log B_{ijk}/(N_i - b_{ijk}) = \alpha + \beta x_k + \epsilon_{ijk}, \tag{4.2}$$

with the ϵ_{ijk} independent normals of mean 0 and variance σ^2. The assumption here may be checked, in part, by fitting the logitnormal and examining probability plots of the residuals. This was done for the data consisting of the annual totals for each day of the week and census division. No strong departure was noted although there was an intriguing outlier.

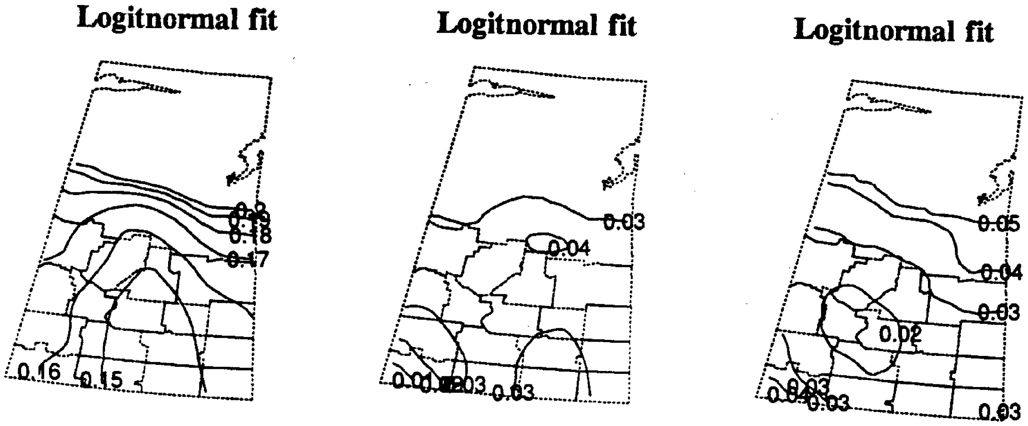
The annual birth rate is now estimated via

$$\frac{5}{7}\int \frac{\exp \hat{\eta}_1}{1+\exp \hat{\eta}_1}\phi(z)dz + \frac{2}{7}\int \frac{\exp \hat{\eta}_2}{1+\exp \hat{\eta}_2}\phi(z)dz, \tag{4.3}$$

where $\eta_k = \hat{\alpha} + \hat{\beta}x_k + \hat{\sigma}z$ for $k = 1, 2$ and with $\phi(.)$ the normal density. The first term corresponds to weekdays, the second to weekends, as in (3.2). Crouch and Spiegelman (1990) discuss the numerical evaluation of integrals such as those appearing in (4.3). In the present case Gaussian integration with 21 nodes is employed.

The top left-hand panel of figure 5 provides the estimated rate function (4.3). As compared with the binomial fit of Figure 4, the birthrate surface estimate appears flatter. The top right-hand panel is $\hat{\beta}(x,y)$. This surface appears less flat. The final panel gives the estimate $\hat{\sigma}(x,y)$. It appears less in the region around Saskatoon, but of course is subject to sampling fluctuations.

Figure 5



annual rate

weekday effect

sigma

5. UNCERTAINTY COMPUTATION AND DISPLAY

Simple maps are flaunt with difficulty of presentation and interpretation, see for example Monmonier (1991). The provision of associated indications of uncertainty seems even more difficult. This section presents a few procedures for the case of contours.

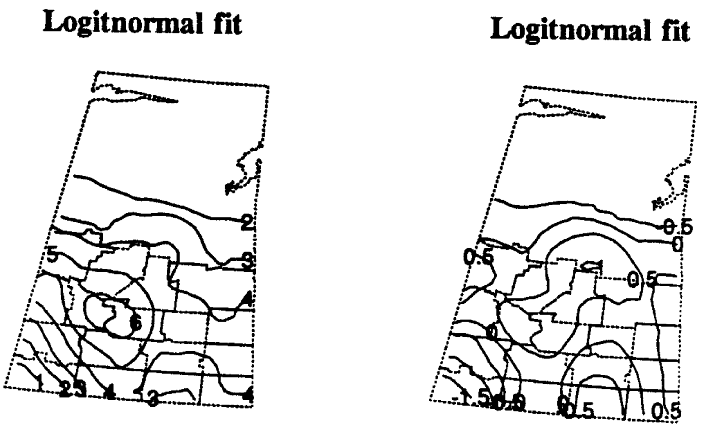
The actual computation of the uncertainties at each (x,y) seems not to be the problem. The estimates produced by the logitnormal fitting are weighted least squares so, writing in traditional notation, the variances of $\hat{\alpha}$ and $\hat{\beta}$ may be estimated by

$$\hat{\sigma}^2(X'WX)^{-1}X'W^2X(X'WX)^{-1}, \tag{5.1}$$

and the standard deviation of $\hat{\sigma}$ by

$$\frac{\hat{\sigma}}{\sqrt{2}}\frac{\sqrt{\sum w_i^2}}{\sum w_i}. \tag{5.2}$$

Figure 6



weekday effect 'O' t-statistic

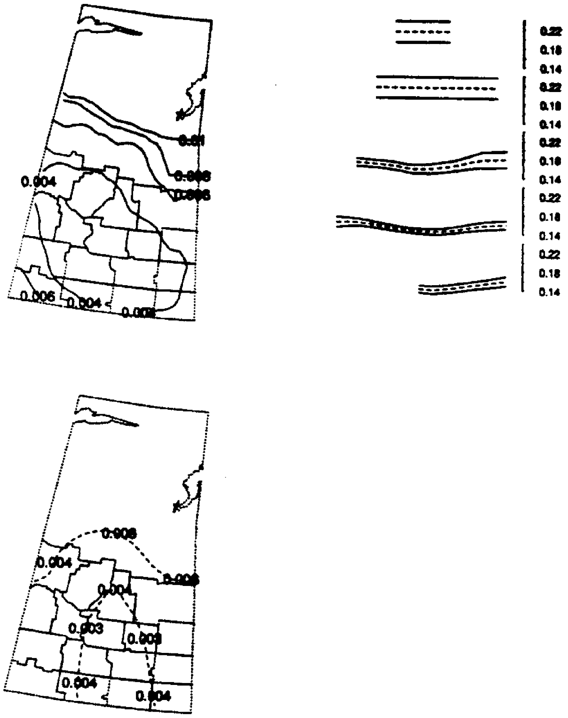
weekday effect t-statistic

The uncertainty estimates may be employed to examine hypotheses. Figure 6 is directed at the issue of whether there is a weekday effect and whether or not it varies with space. The left-hand panel of Figure 6 provides the estimate, $\hat{\beta}(x,y)$, divided by its estimated standard error. The values range from 1 to 6, providing evidence for the presence of an effect. To study whether the effect varies spatially, the t -statistic is recomputed but now with its numerator having the estimated value for the whole province subtracted. Now the t -values range from -1.5 to 1 and there is no real evidence that the effect varies spatially.

The next two figures are directed at displaying the uncertainty of the annual rate estimate given by (4.3) and graphed in Figure 5. The standard error of the estimate (4.3) is estimated by the delta-method. The first panel of Figure 7 is a contour plot of this estimate. Contour levels range from .004 to .010, and suggest a valley centered at Regina and Saskatoon. The top right-hand panel of the figure is based on five east-west slices through the province. Provided are ±2 standard error limits about the estimated rate along the slice. This is a traditional means of indicating uncertainty for functions of a single variable. The final panel graphs the .15 and .17 birthrate contours as dashed lines and superposes the estimated standard errors at selected positions along these contours. It appears simpler to take in the uncertainty this way, but one cannot read off the standard error estimates for most locations.

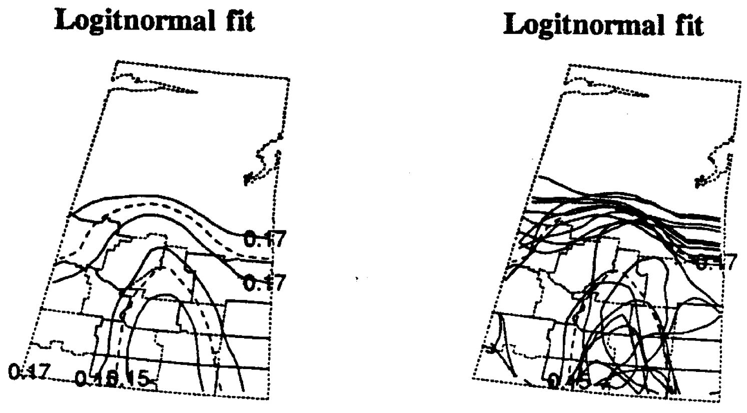
Figure 7

Logitnormal fit



The left-hand panel of Figure 8 indicates the shift in contour lines produced by adding and subtracting one standard error to the estimated rate function. Smooth bands appear about the estimated rate contours. The intention of the bands is clear, but thoughtful interpretation is needed. The right-hand panel provides 10 simulations of the process. The individual estimates $\hat{\alpha}(x,y)$, $\hat{\beta}(x,y)$, $\hat{\sigma}(x,y)$ are first aggregated over census divisions. Then independent realizations of the process are generated according to (4.2). The model is refit for each simulation, in bootstrap style, and the .15 and .17 contours determined. Diaconnis and Efron (1983) propose bootstrapping of contours. The .17 contour realizations fall in an apparent band, but the .15 move about a fair amount with some quite wild curves. The dashed lines are the original .15 and .17 contours.

Figure 8



Logitnormal fit

Logitnormal fit

annual rate +- 1 standard error

annual rate - 10 siumlations

6. DISCUSSION AND SUMMARY

The combination of the weight function, $w(.)$, and the random effect σz, allows the estimate at location (x,y) to borrow strength from the values of all census divisions, see for example Mallows and Tukey (1982). Hence this approach provides an alternative to the empirical Bayes estimates developed in Clayton and Kaldor (1987), Tsutakawa (1988), Cressie and Read (1989), Manton *et al.* (1989) for this type of data.

The computation of uncertainty has allowed examination of hypotheses of no weekday effect and weekday effect constant across the province. An advantage of the graphical approach is that were there spatial variation, then the plots might have suggested its character.

A substantial amount of work remains for the future. This includes: choice of the bandwidth, b, in (2.5), use of other weight functions, informal and formal analysis of goodness of fit, other methods to display uncertainty, including measured explanatory variables and finally appropriate asymptotics to employ in studying the technique.

ACKNOWLEDGEMENTS

The research was partially 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). *Annals of Statistics*, 5, 622-623.

Brillinger, D.R. (1990a). Mapping aggregate birth data. *Proceedings of the 1989 Symposium on Analysis of Data in Time* (Eds. A.C. Singh and P. Whitridge), Statistics Canada, Ottawa, Canada, 77-83.

Brillinger, D.R. (1990b). Spatial-temporal modelling of spatially aggregate birth data. *Survey Methodology*, 16, 255-269.

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

Cleveland, W.S., and Devlin, S.J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83, 596-610.

Cressie, N., and Read, T.R.C. (1989). Spatial analysis of regional counts. *Biometrika*, 31, 699-719.

Crouch, E.A.C., and Spiegelman, D. (1990). The evaluation of integrals of the form $\int_{-\infty}^{\infty} f\,(t)\exp(-t^2)dt$: application to logistic-normal models. *Journal of the American Statistical Association*, 85, 464-469.

Diaconnis, P., and Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, 248, 116-130.

Dyn, N., and Wahba, G. (1982). On the estimation of functions of several variables from aggregated data. *SIAM Journal of Mathematical Analyses*, 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. *Journal of the Royal Statistical Society*, 29, 355-369.

Mallows, C.L., and Tukey, J.W. (1982). An overview of techniques of data analysis emphasizing its exploratory aspects. *Some Recent Advances in Statistics* (Eds. J. Tiago de Oliveira *et al.*). Academic, London, 111-172.

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. *Journal of the American Statistical Association*, 84, 637-650.

Monmonier, M. (1991). How to Lie with Maps. Chicago Press, Chicago.

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 Department, Oregon State University.

Sabin, M.A. (1985). Contouring - the state of the art. Fundamental Algorithms for Computer Graphics (Ed. R.A. Earnshaw). *NATO ASI Series*, F17. New York: Springer-Verlag.

Sanathanan, L., and Blumenthal, S. (1978). The logistic model and estimation of latent structure. *Journal of the American Statistical Association*, 73, 794-799.

Shepard, D. (1968). A two-dimensional interpolation function for irregularly spaced data. *Proceedings of the 23rd National Conference ACM*, 517-523.

Staniswalis, J.G. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, 84, 276-283.

Stone, C.J. (1977). Consistent nonparametric regression. *Annals of Statistics*, 5, 595-620.

Tibshirani, R., and Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, 82, 559-567.

Tobler, W.R. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74, 519-536.

Tsutakawa, R.K. (1988). Mixed model for analyzing geographic variability in mortality rates. *Journal of the American Statistical Association*, 83, 37-42.

FIGURE LEGENDS

Figure 1. Annual birth rate for women aged 25-29, years 1986-7, by census division. R and S indicate the locations of Regina and Saskatoon respectively.

Figure 2. Provides (2.4) for census division 18 in the cases of no, of a small amount and of a moderate amount of smoothing.

Figure 3. (2.4) as employed for all 18 census divisions.

Figure 4. The binomial analysis. The rate estimate is from (3.2). The righthand panel gives $\hat{\beta}(x,y)$.

Figure 5. The logitnormal analysis. The rate estimate is from (4.3). The other panels give $\hat{\beta}(x,y)$ and $\hat{\sigma}(x,y)$.

Figure 6. t-statistics directed at the hypothesis of no weekday effect and of spatially constant weekday effect.

Figure 7. The first panel is a contour plot of the estimated standard error of the estimated birth rate. The second panel provides plus and minus two standard error limits about S east-west slices through the estimated birth rate surface. The final panel plots the estimated standard errors at selected points on the .15 and .17 contours.

Figure 8. The first panel gives the .15 and .17 contours as dashed lines, then the same contours for the surface increased and decreased by one standard error. The second panel shows the results of simulating realizations of the process (4.2) and determining and displaying estimates in the manner of the paper. Ten simulations are shown.

November 1991

SYMPOSIUM 91

Spatial Issues in Statistics

PROCEEDINGS

Statistics Canada Statistique Canada

Canada