Spatial Data Final Project

Kriging Illinous Breast Cancer Mortality Rate

Haicheng Tang

April 25, 1999

1 Abstract

I conduct kriging on breast cancer standardized mortality rates of Illinious counties. Since a lot of those rates are missing, and the existing rates don't display a trend over geographic regions, I carry an ordinary kriging on existing rates and predict the rates on those counties that don't have rates available. A spherical variogram seems to fit the data well and generates reasonable results. A map of existing and predicted mortality rates is plotted and compared with the map of existing rates only. From the plots we can see the kriged (predicted) rates are close to the existing rates of neighboring counties, indicating a smoothing of mortality rates by kriging.

2 Introduction

The breast cancer mortality rates data of Illinious counties are availble at

http://www.sph.umich.edu/~ aelon/etc/data/illinois/il14wm.txt.

Age adjusted cancer mortality rates per standard million population in white females in Illinois from 1970-79 are collected for Illinious counties, together with latitude and longitude (in degrees and decimal minutes) of county centers and couty fibs. Unfirtunately, rates for most of the counties are missing: for the 102 couties, 34 have rates available while the rest 68 do not have breast cancer mortality rates.

Strictly speaking, these rates are based on counts, and therefore should be assumed a discrete probabilitic distribution. That will make modelling their spatial correlations quite difficult, although not impossible (I will return to this point later). However, noticing each county has a pretty large population, and applying central limit theorem, we can view these rates as approximately follow normal distribution. Hence classical kriging method can be used.

Kriging is essentially optimal unbiased linear prediction. The idea is to find a linear combination of existing data that can best predict the observation at a certain spatial point. This prediction should be unbiased, i.e., with expectation the same as the observation to be predicted; and having the smallest prediction error, i.e., the variance of this prediction needs to be smaller than any other linear unbiased prediction.

The spatial process we consider here is the mortality rates defined on a continuous 2-dimensional space, the state of Illinious. The observed rates of some of the counties are assumed to be defined on the centers of those counties. Assuming the underlying spatial process to be a Gaussian process, as justified above, it is convenient to model the correlation between observations at different points, and from there, to krig the whole mortality rate space, with estimation of precision of the kriging. Of more interest might be those predictions on the centers of the counties with missing rates, which will be used as predicted rates for those counties.

Next I will give details on how to carry out this idea and try to justify more formally the assumptions involved. The software I use are Splus and AcrcView. A sketch of the analysis is given at http://www.sph.umich.edu/~haicheng.

3 Prelimanary Treatment of Data

First I read the data into Splus. Obviously the counties with missing values on rates are not that useful in fitting the model. Therefore I delete those counties to obtain a dataset based on which the kriging is to be condicted. There are actually more counties with missing rates (68 out of 102). Of course the center positions for the counties with missing rates will be utilized in kriging their rates.

Since Illinious does not cover a big range in latitude and longitude, I ignore the difference between spherical distance and plane distance in this small range and regard the distance in latitude and longitude as surrigate for the actual distance. The original data contain latitude and longitude in degrees and decimal minutes. I combined the degrees and minutes into one measurement each on latitude and longitude.

4 Exploratory Data Analysis

In order to give a visual display of the standardized mortality rates (SMRs), I used ArcView to plot a mortality rate map (http://www.sph.umich.edu/~haicheng/illbo.jpg). The plot shows somehow an increase in rate to southeast. However the trend is not obvious, especially with a lot missing county rates. The contour plot (http://www.sph.umich.edu/~haicheng/cont.gif) and circle plot (http://www.sph.umich.edu/~haicheng/circle.gif) reveal similar message. Notice in both plots, the horizontal axis is the latitude and the vertical one the longitude.

Trend is an issue that has to be dealt with in kriging. Generally there are two ways to handle that: one is to remove the trend and then do ordinary kriging on the detrended spatial process, the other is to integrate the trend into the kriging procedure, and conduct so-called universal kriging. In this dataset, however, upon checking the perspective plot of the mortality rates (http://www.sph.umich.edu/~haicheng/persp1.gif), I found there does not appear to be a trend. Therefore, ordinary kriging is sufficient to deal with this dataset.

Ordinary kriging is based on the idea of best linear prediction. The linear prediction is optimal among all possible forms of prediction if the process is Gaussian, which in this dataset, means the joint distribution of mortality rates on all sites follow multivariate normal distribution. Strictly speaking there's no way to check the above multivariate normal distribution since we have only one replicate of the multivariate sample. It is like checking whether a random variable is normally distributed with only one value (realization) from that distribution. However, roughly we can check the normality of the data by treating them as from one i.i.d. (independent and identically distributed) sample. From the Q-Q plot (http://www.sph.umich.edu/~haicheng/qq.gif) we can see the data are approximately normally distributed, but have a heavy tail on the left side (smaller rates). This can be expected as Poisson distribution, which has heavier left tail, is a better approximation for these counts data. And from the raw data, it is obvious that there are many small or zero counts. However, this does not provide convincing evidence against normal assumption as these rates are actually correlated and that might account for part of the deviation from normal assumption displayed by the data.

Ordinary kriging utilizes the correlation of SMRs at different locations, which is reflected in variogram, defined as $Var(Z(S+h)-Z(S))$. Before we formally model the variogram, it helps to take a look at the sample variogram, which in turn can be preceded by checking the variogram cloud (http://www.sph.umich.edu/~haicheng/vgcld.gif), which are the pairwise squared differences of county rates versus the distance between the two counties, and essentially componenets in computing sample variogram; to ensure that these components display some smooth trend. It does appear so, and I feel comfortable computing the sample variogram, whose plot appears at http://www.sph.umich.edu/~haicheng/smvg.gif, which displays an increasing trend of variogram with respect to spatial distance.

The above sample variogram estimate assumes the mortality rate process is isotropic. Hence squared differences of county rates for counties with similar distance apart but departing in different directions can be combined. Without this assumption we would have to compute the sample variogram for each direction and may not have enough sample in some directions to achieve a reliable estimate. To check this isotropic assumption more formally, I plot the variogram cloud plots for four ranges of directions (http://www.sph.umich.edu/~haicheng/isotr.gif) and find that they display very similar patterns. Therefore the isotropic assumption, and hence the sample variogram as an estimate for variogram, are valid.

5 Formal modeling

Theory of ordinary kriging tells us the prediction at a certain point s_0 with exsiting observations $Z(s_1), ..., Z(s_n)$, in this case, the observed mortality rates at 34 counties, is $\sum_{i=1}^{n} \lambda_i Z(s_i)$, where

$$(\lambda_1, \lambda_2, ..., \lambda_n) = \lambda^T = (\gamma + 1\frac{1 - 1^T\Gamma^{-1}\gamma}{1^T\Gamma^{-1}1})^T\Gamma^{-1}$$

and

$$\gamma = n \times 1 \text{ vector with} i^{th} \text{element} \gamma(s_0 - s_i),$$

$$\Gamma = n \times n \text{matrix with} (i,j)th \text{ element} \gamma(s_i - s_j).$$

The sample variogram obtained above does not suffice this prediction. The reason the sample variogram only provides estimates for variogram for certain distances (between locations). But we need variogram estimates for every possible distance (within a certain range, of course, say, in this case, within Illinois) since we want to be able to krig on any possible point s_0, in particular, the center points of those counties with mortality rates missing, and therefore need variogram of distance $|s_0 - s_i|$ for each i. All the $|s_0 - s_i|$ may not appear in all the $|s_i - s_j|$'s, whose corresponding variogram estimates we have already obtained. In particular, for those $|s_0 - s_i|$'s that are smaller than the minimun of $|s_i - s_j|$'s, it's not possible to get a direct estimate of the variogram from the sample variogram. For $|s_0 - s_i|$'s that are between some $|s_i - s_j|$'s, whose corresponding variogram estimate presumably can be obtained by intrapolation, and even for sample variogram itself, the estimates suffer from large instability since there are very few data points that are used for the original sample variogram estimate, causing large variance of the estimate.

One way to get around this problem is, based on the crude sample variogram, to fit a theoretical variogram, which can combine information on the variograms of different distances and yield a more reliable estimate. Of course, a proper variogram model is needed to achieve the goal. upon inspecting the sample variogram, I find a spherical variogram fits the sample variogram pretty well. Other options, such as linear variogram, are possible in this dataset.

With model.variogram in Splus, I fit a spherical variogram to the sample one. First notice there is a big escalation in the sample variogram at very short distance. That indicates nugget effect should be considered. Using least square as the criteria, I find a spherical variogram function with range 2.5, sill 20, and nugget 2.8 gives the best fit to the sample variogram. Graphicaly it looks a good fit (http://www.sph.umich.edu/~haicheng/vg.gif).

Now we can apply the kriging formula with function "krige" in Splus and the variogram (theoretical) fitted above. The kriged surface is displayed at http://www.sph.umich.edu/~haicheng/glb.gif. Comparing it with the perspective plot for the crude mortality rate, we can see the kriged surface reserve the shape of the crude rate surface but is more smooth. Hence we successfully smoothed out some of the random variations in the crude rates, using the autocorrelation between the underlying rates.

Any estimation or predition should be accompanied by an assessment of its precision. Therefore I calculated the standard errors of the above kriging, with its plot at http://www.sph.umich.edu/~haicheng/glbse.gif. Clearly the standard error is smaller for points close to the center and escalates when the points go to the boundaries. This should be expected as we do have more information available for points around the center, i.e., more data points colse to them. However, noticing the scale of the standard error estimate, the maganitude of difference is actually not that large between standard errors of the center and boundary points. With the standard error estimates and normal approximation, it is easy to get the lower and upper 95% confidence surfaces (http://www.sph. umich.edu/~haicheng/glbcf.gif).

For practical purpose, it is of more interest to krig on those counties with missing mortality rates. I obtained the prediction (http://www.sph.umich.edu/~haicheng/impute.txt) in Splus, and transport the imputed values to Arcview for a complete plot of the Illinois county breast cancer mortality rates (http://www.sph.umich.edu/~haicheng/illbk.jpg). The plot shows a generally smooth change of mortlity rates, which seem to be higher at the southeast corner but do not display an obvious trend. However, at some places, there appear to be strong contrast between adjcent counties. That is because in the ordinary kriging implemented in Splus, the kriged values for points with observations available are the raw data themselves. A more reasonable kriged value might be a weighted average between the raw observation and some kind of global or local mean of the rates of other counties, taking into account there is measurement error (nugget effect) in each observation.

6 Conclusion

The age adjusted breast cancer mortality rates in white female in Illinois counties between 1970 and 1979 display a smooth change geographically. The southeast region appears to have relatively higher rates but overall there is no obvious trend.

7 Discussion

Although we are able to obtain satisfactory and sensible results with ordinary kriging, there are plenty of caution and alternative options to analyse the data that are noteworthy.

First of all, as some instructors pointed out, there is discrepancy between geographical distance and distance calculated from latitude and longitude assumed to be in a Cartesian system. It is better to transfer the distance into geographical distance. In this dataset, however, the size and location of Illinous seems to justify usage of latitude and longitude directly as a pretty good approximation.

In terms of technical considerations in statistics, I have to say my analysis leaves sufficient room for possible argument and possible improvement.

As mentioned at the end of formal modeling part, a more reasonable kriging is to get rid of the measurement errors for the existing observations; in other words, to predict the "noiseless" version of the underlying process. The reason is measurement error changes wach time we make our observation and therefore is not an intrisic part of the process, therefore we are not interested in estimating the process plus a particular realization of the measurement error. It is quite straightforward theoretically to develop a "noiseless" version of kriging. Operationally, however, I have not been able to find a program or package that can achieve that.

In exploratory data analysis part, I mentioned there can be doubt about the normality of the data. This is a prevailing problem with kriging disease rates, since the rates come from bionomial distribution and normal distribution can only serve as an approximation. How well the approximation is depends on the magnitude of the rate. For rare disease it is going to be poor.

There are two ways to get around this problem. The first is to use a proper transformation, such as logarithm, square root, and an improvement based on it: Freeman-Tukey transformation (square root of SMR plus square root of (SMR+1)). The difficulty in implementing these methods lie in the fact that if we want to make the kriging unbiased in the original scale, we have to make some adjustment to the kriged value.

Another way is simply to use another distribution, such as Poisson distribution. This is actually the idea of generalized linear model: to combine the linear regression structure with any likelihood we want, not necessarily normal distribution, with the help of a link function. Poisson serves as a better approximation than normal for count data. In this dataset, we can see heavier left tail (ref: exploratory data analysis part) than expected by normal. And Poisson distribution does have a heavy left tail.

Traditionally, Poisson distribution is modeled directly. Since we need to take care of spatial correlation, that leads to conditional auto Poisson model ((6.5.6) (6.6.7) of Cressie 1993), which is hard to deal with theoretically and computationally. In particular, there are two awkward phenomena with this model: it does not allow positive correlation between neighboring sites, and the marginal distribution is no longer Poisson. Kaiser and Cressie (1997) tried to model positive spatial dependence with truncated version of conditional auto Poisson model. Still there are some problems pending, such as, how to set the truncating cap. Computation also becomes more messy.

Another approach to modeling Poisson, or any other non-Gaussian distributoin, including binomial one, is to apply the strategy of hierarchical modeling. In the case of Poisson modeling, we can model the spatial dependence not on Poisson distributions of each site directly, but on their parameters (of course, they are assumed to be random). By this approach we obtain flexibility in modeling the dependence. Also, with morden MCMC algorithms, computational issues no long bother us, although we still need to be careful in monitoring convergence. The resulting smoothed version of disease rate under the Bayesian paradigm, comes natuarally as a shrinkage of the raw rate toward to global or local mean. These estimates borrow strength from neighboring areas, with observations with smaller populations shrinked more than those with large population. Therefore the problem with the raw disease map, which is strongly impacted by rates from areas with smaller population, due to their large variance and exaggerated effects, is solved. An excellent original reference is Besag et al (1991).

When the details of the hierachical model is hard to lay out, such as when there is not enough prior information to set up the model, a GEE (generalized estimating equation) approach, which deals directly with the marginal model, can be applied to achieve robust estimates with simple computation (e.g., Heagerty & Lele (1997)).

To avoid sensitivity to model assumption, nonparametric smoothing methods have recently been applied to sptatial data (e.g., see Muller et al. (1997)).

The mainstream on smoothing of discrete data seems to have shifted from classical kriging to morden hierarchical modeling techniques.

In principle, whichever method or model we use, some covariates can be used to potentially improve the accuracy of our prediction. In kring, this is called cokriging. For this dataset, unfortuanately we do not have this option as no covariate is collected in the dataset.

8 Acknowledgment

I'm very grateful to Dr. Andy Long for his help in plotting with Arview, setting up website, and constructive suggetions on the data anlysis. Also many thanks to Dr.s Geoff Jacquez and Dr. Mark Wilson for their helpful comments.

9 References

Besag et al (1991) *Bayesian image restoration with applications in spatial statistics* Annals of the Instit. of Stat. Math., 43, 21-59

Cressie, N., *Statistics for spatial data* 1993.

Kaiser, Mark S.; Cressie, Noel (1997) Modeling Poisson variables with positive spatial dependence; Statist. Probab. Lett. 35 (1997), no. 4, 423–432.

Heagerty & Lele (1997) *A composite likelihood approach to binary spatial data* JASA, vol. 93, 1998, 1099-1111

Muller et al. (1997) *Spatial smoothing of geographically aggregated data, with application to the construction of incidence maps* JASA, vol. 92, 1997, 61-71