Literature Review

Spatial Prediction

Don Eagan & Chintan Patel
(Group 13)

[1]Paaß, G. and Kindermann, J. 2003. Bayesian regression mixtures of experts for geo-referenced data. Intell. Data Anal. 7, 6 (Dec. 2003), 567-582.

Review:

This paper identifies the need for politicians, planners and social scientist to be provided the tools to clarify and manipulate spatial distributions to predict future developments. Bayesian statistics offers a way to estimate values of a variable at locations that are not sampled. The paper tries to address a case where Tobler's law is not applicable. They are using Marcov Chain Monte Carlo analysis to develop better model based on probability distributions. and This paper considers two versions of the Bayesian variant of a flexible semi-parametric model (a mixture of experts). This consists of a gate model which divides the problem into regions of local models called expert predicts. The expert predict is responsible for providing the output. The two methods involve either providing an input variable to a single expert or to multiple experts. These two Bayesian inference models were tested using real geographic data. The results show that both Bayesian models were able to capture stochastic relation in spatial data. The multi-variable approach was able to identified nonlinearities not evident from the single dimensional approach. There were no comparisons of performance or memory usage offered by this paper. The experiments didn't involve high dimensional dataset which could be a concern. They tried to long term predict illness based on Manchester, UK census data.

[2]Baris M. Kazar, Shashi Shekhar, David J. Lilja, Ranga Raju Vatsavai, R. Kelley Pace: Comparing Exact and Approximate Spatial Auto-regression Model Solutions for Spatial Data Analysis.GIScience 2004: 140-161

Review:

Applications that use spatial auto-regression (SAR) for data mining are working with ever increasing sizes of geo-spatial databases. The explosive growth in databases coupled with the demand for exact solutions for estimating SAR parameters are both computationally expensive and memory intensive. This paper presents two candidate approximate-semi-sparse solutions of the SAR model based on the Taylor series expansion and Chebyshev polynomials. When accuracy of these new approximation algorithms and an exact algorithm were compared, both provided accurate results. However, the approximation algorithms outperformed the exact algorithm in both terms of computation and memory usage. It was also noted that the exact algorithm was unable to solve any problem with over 10K observation points. They performed experiments on satellite imagery. Authors suggested exploring better model based on this approach to get better prediction.

[3]Smirnov, O., & Anselin, L. (2001). Fast maximum likelihood estimation of very large spatial autoregressive models: a characteristic polynomial approach. Computational Statistics & Data Analysis, 35, 301–319.

Review:

This paper states that the maximization of the log-likelihood function used in spatial autoregressive models is computationally intensive and requires significant amounts of memory. This becomes problematic during analysis when very large spatial data sets are used. This papers contribution is [3]a new method for evaluating the Jacobian term based on the characteristic polynomial of the spatial weights matrix W. Comparisons made between Cholesky factorization and this characteristic polynomial algorithm showed pronounced improvement when large data sets ($n>50000$) were examined. In addition, the Cholesky algorithm failed when large data set were used due to large memory requirements. Clearly the characteristic polynomial algorithm proposed by this paper is preferred when using large data sets. This algorithm also includes a tuning variable to vary the accuracy of the result. However, increasing the accuracy of the result also increased computation times. The proposed solution is O (n) solution for regular lattices and O (nlogn) for irregular lattices.

[4]D.A.Griffiths. Quick but not so dirty ML estimation of Spatial Autoregressive Models, In Toolkits in Regional Science, M.Sonis and G.J.D Hewings (eds.), Series: Advances in Spatial Science

Review:
This paper identifies data set scalability problems for scientists implementing spatial statistical models. Specifically, this problem has been narrowed down to the determinant det(V), a normalizing constant, in the auto-Gaussian log-likelihood function:

$$\text{constant} - (n/2)\ln(\sigma^2) + \ln[\det(V)] - (Y - X\beta)^T V (Y - X\beta)/(2\sigma^2)$$

This paper states, “the normalizing constant det(V) is problematic because it (a) is a function of unknown spatial autocorrelation parameters, (b) is unwieldy, (c) fails to have a closed-form expression, and (d) almost always defies a numerical solution for sufficiently large *n* since it involves an *n*-by-*n* matrix”. This paper extends the initial work performed in [3], which proposed a new method for evaluating the Jacobian term. This paper offers four Jacobian approximations as solutions caused by the normalizing constant. Evaluation of these approximations [4]indicate that virtually any size of georeferenced data set can now be adequately described with a spatial autoregressive model. This approach doesn't work well with high dimension data (remotely sensed images)

---

[5]R.K.Pace and R.Barry. Fast CARs (Conditional Auto-Regressive Models). Journal of Statistical Computation and simulation. 59(2), 1997. pp123-147.

Review:
This paper cites the growth of spatial information and the need to process it. Obtaining traditional spatial statistics is impeded by increased size in data set requests. This paper addresses the problem of quickly computing the maximum likelihood spatial statistics for large sample sizes using the conditional auto regression (CAR) algorithm. This paper offers three enhancements to CAR processing to address performance issues. First, the CAR is modified to use sparse matrices second; the sum-of-squared errors (SSE) function is restructured to accelerate computations and third; computations are performed using vectors. Comparison were made with the eigenvalue route taken by Li (1995) showed that solutions employed by this paper allowed a personal computer on a larger problem to exceed the performance of a supercomputer on a smaller problem. They compute likelihood over the grid to make this approach more robust to local optima.

---

[6]A.S. Fotheringham, C. Brunsdon and M.E. Charlton. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships, Wiley, Chichester (2002).

Review:

Authors mentioned few techniques that incorporate local spatial relationships in to the regression framework, which is very popular and well known in the statistics community. Authors showed that hedonic price model to capture price variation in London housing market are incorrect. This is because of non-stationarity property exhibited by the dataset. . One method attempts to calibrate the geographic model based on established boundaries. Another method, identified as the moving window regression, further partitions the geographic space into regions where the data can be identified with a regression point. This approach provides a refinement, but only provides a constant weight to the data points within a region. Geographically Weighted Regression (GWR) solves this problem by assigning weights to the data points based on the distance from the regression point. For example, similar homes typically increase in value as their location moves towards a lake or river. GWR provides to ability to locally assign a unique function to each regression point that indentifies both weight and bandwidth.

The global regression model can be written as:

$$y_i = \beta_0 + \sum_k \beta_k X_{ik} + \varepsilon_i$$

GWR extends this model to include locality:

$$y_i = \beta_0(u_i,v_i) + \sum_k \beta_k(u_i,v_i) X_{ik} + \varepsilon_i$$

where: (u_i,v_i) = coordinates of the ith point in space
$\beta_k(u_i,v_i)$ = is a realization of the continuous function $\beta_k(u,v)$

The success of the GWR is demonstrated in a number of empirical studies which used GWR to identify detailed variations in spatial relationships which could not be done using global estimates. The discussion of this writing, limited to chapter 2, offers two techniques on methods to further testing. These methods include the Monte Carlo approach and the development of theoretical tests. Since both of these tests are computer intensive the authors suggested using only the Monte Carlo approach and limit the data sets and/or the iterations. No test results were provided. The authors also identified a similar study, conducted in Los Angeles County, which also used GWR. This writing fails to make a comparison of the results or techniques used in the Los Angeles County study. This leaves the reader to question the uniqueness of the solution offered by this paper.