THE ISBA BULLETIN

Vol. 16 No. 2 June 2009

The official bulletin of the International Society for Bayesian Analysis

A MESSAGE FROM THE PRESIDENT

by Mike West
ISBA President
mw@stat.duke.edu

I am writing this while on a return trip from one of the several ISBA sponsored meetings of 2009, in this case the 6^{th} Workshop on Bayesian Inference in Stochastic Processes (BISP6) held this June in Bressanone, Italy. The biennial BISP meeting has become a fixture on the Bayesian calendar and, with over 80 participants, BISP6 was intellectually vibrant and impressive in several respects. Talks and posters at the meeting continued to expand the original BISP focus on inferential and computational developments in applications coupled with theoretical and modelling research in stochastic processes. The "Bressanone blend" was very rich indeed (see web link below). A lively international gathering that was enriched with substantial numbers of junior researchers and students, BISP6 also continued the tradition of a focussed meeting in a simply lovely environment. I encourage all ISBA members to mark their diaries for BISP7, to be held (sometime, somewhere) in early summer 2011, and congratulate long-term ISBA activist Fabrizio Ruggeri for his work in establishing and running this increasingly influential series of workshops. I have more to say on meeting sponsorship below, along with updates, ... *Continue in page 2.*

A MESSAGE FROM THE EDITOR

by Raphael Gottardo
raphael.gottardo@ircm.qc.ca

Summer (at least for some of us) is finally here, and it feels good to enjoy the nice and warm, weather. Summer also means traveling, taking some vacation, finishing the few papers we put off all year long, etc. In short, we end up being more busy than during the regular year! In addition to this, a few of the Associate Editors have even gotten busier than ever with babies, taken on new jobs and more. As a consequence, it has been a bit difficult for some of them to provide their regular contributions on time. Even though a few of the regular sections are missing, you will find many interesting articles as well as updates from our President and from Herbie Lee and Hedibert Lopes on the 2010 ISBA meeting and ISBA memberships. I would also like to remind YOU that we want YOU to contribute to the bulletin, so if you have any interesting material you'd like to include in the bulletin, feel free to contact me or any of the associate editors. While I await for your next contributions, I wish you a wonderful break and wonderful summer/winter. ▲

In this issue

WORDS FROM THE PRESIDENT, *Continued from page 1.*

and news on other ISBA discussions and activities in recent weeks.

On ISBA Sponsored Meetings (both generally and specifically):

The Board and Program Council are currently discussing the roles of ISBA in sponsoring and supporting workshops and conferences. ISBA is receiving increasing numbers of requests for meeting sponsorship, and has to date had a rather informal process for considering and responding to such requests. Existing guidelines http://bayesian.org/business/meetingslocations.html stress selectivity in sponsoring meetings, and considerations of *"... quality and appropriateness, whether the conference theme is of potential interest to ISBA membership, timing, and who the other co-sponsoring organizations are."* Beyond this, the specifics of what sponsorship means at a practical level are up for grabs on a case-by-case basis, so we are considering whether a short-list of "best practices" might be beneficial. As a move towards this, recent Board approvals for sponsorship have involved more specific "contracts" with organisers, including discounted registration for ISBA members, and active and visible promotion of the ISBA logo and web link in meeting PR materials. In some, very select cases, we are continuing the tradition of offering 1-year membership to meeting participants not already members, though this is a talking point with respect to longer-term practices. In other cases, we have established agreements that ISBA will be formally involved in the program of the meeting.

I invite you all to contribute to the discussion of ISBA's roles in meeting sponsorship. If you have any specific suggestions or comments, please do communicate them to members of the Board and/or Program Council.

To update you on proximate sponsored meeting: the BISP6 meeting in mid-June http://www.mi.imati.cnr.it/conferences/bisp6.html was followed by the 7^{th} Workshop on Bayesian Nonparametrics in Turino http://bnpworkshop.carloalberto.org/. 2010 will see several ISBA sponsored or co-sponsored meetings, including: EBEB 10, the 10^{th} Brazilian Bayesian meeting organised by the Brazilian Local Chapter of ISBA http://www.ime.usp.br/~isbra/ to be held in Rio de Janeiro next March (note that this is EBEB 10, coincident with the 10^{th} birthday of ISBrA, in 2010!); the workshop on Frontiers of Statistical Decision Making and Bayesian Analysis, in San Antonio, Texas, also in March http://bergerconference2010.utsa.edu; and the CBMS/NSF conference on Bayesian Nonparametrics to be held at Santa Cruz, California, sometime in the summer of 2010. Further, we have now finalised arrangements to co-sponsor the next in the quadrennial series of international meetings of the International Chinese Statistical Association http://www.icsa.org/. The 8^{th} International Conference of ICSA, Frontiers of Interdisciplinary and Methodological Statistical Research, will be held at Guangzhou University, China, in December 2010. As part of the co-sponsorship agreement, ISBA is already developing plans for ISBA sessions of talks, and we are continuing to discuss ways in which we can develop interactions with ICSA, as with other societies, to the benefit of ISBA members and consistent with ISBA's mission.

World Meetings:

Planning is well underway for the next major ISBA meeting, ISBA10 – the 10^{th} ISBA World Meeting – to be held in conjunction with the 9^{th} Valencia International Meeting on Bayesian Statistics, in Spain next year. For more details, see the comments from Herbie Lee in this Bulletin, and begin to plan to submit papers and posters later this year once the call for submissions is announced.

It is also time to start thinking about the 11^{th} World Meeting of ISBA in 2012. In coming months, ISBA will put out a call for proposals for the 2012 meeting. The Program Council will be interested in discussions – however preliminary – with any interested group with respect to location and timing of ISBA11. Among the key factors in considering the site will be general suitability of the location for a reasonably large meeting, local organising committee, local funding potential, and balance of the representation of regions around the world over the series of ISBA World Meetings. So, please put this on your radar screens and – as the mood takes you – get involved in thinking and planning initiatives for the next in *the* major series of Bayesian conferences. More on this in a later Bulletin.

Membership:

I am very pleased to report renewed buoyancy in membership sign-ups in the early months of 2009; We are now well over 500 members and have an increasing numbers of Life Members (new and converting from standard memberships). As reported in the first Bulletin of 2009,

the new ad-hoc *Committee on Membership* is working on various topics related to membership services; you can read comments from the committee chair, Hedibert Lopes, elsewhere in this Bulletin.

ISBA IT & Virtual Office:

One of the major developments in recent weeks has been the start of a process to develop a professional ISBA "virtual office", beginning with the establishment of a membership and event management system. This began in earnest at the start of June, following Board approval of a proposal developed in late spring. This is a non-trivial effort but is now nicely underway, and we expect to be in touch again within a few weeks when the first phase – automated membership management – is fully functional, fully tested and integrated into the ISBA web site. This development is being carried out with the services of a professional data base developer and with expert IT support, but is being done at very modest cost to ISBA and with substantial institutional support from a host university department.

On specifics, the ISBA system will shortly have a fully operational version of the membership management system in the CiviCRM software www.civicrm.org. CiviCRM provides facilities for membership management, including dues payments and general record keeping, membership email and subscription services, and event management including workshop/conference registration processing, scheduling, etc. These are the key areas of current priority interest for ISBA. Following broad consultation with IT professionals on this, the ISBA Board approved development of the CiviCRM as a well-regarded and effective system that would serve these interests. As a result of these discussions and consultations, we are now midway through the installation and integration of CiviCRM into the ISBA web space/server, migration of current (spreadsheet-based) membership records into the system, with development for routine use of CiviCRM for all aspects of ISBA membership management, ISBA membership email and subscription services, as well as aspects of ISBA-related workshop and conference management.

The next time you hear of this should be in an invitation to visit the ISBA web site and log into the system to check or update details of your own membership, and then in the end-of-year membership renewal process that will be – from that point – automated. Beyond membership management, we will be piloting the use of the system for workshop registrations in one of the ISBA sponsored workshops in early 2010. Keep posted!

Committees:

The last few weeks have seen personnel changes on several active ISBA committees. All new appointees will be giving valuable time and effort to ISBA activities on a voluntary basis, and should be thanked by us all for their willingness to serve.

First, the 2009 Nominating Committee: chaired by past-president Christian Robert, the regular members of this committee are Marilena Barbieri, Carlos Carvalho, Simon French, David Higdon, Eduardo Gutiérrez-Peña and Judith Rousseau. This committee will select candidates to take office in 2010 as President Elect (to serve as ISBA President in 2011), Executive Secretary (for 2010-2012) and Board members (four members for 2010-2012).

Second, David Madigan has joined the ISBA Constitution & Bylaws Committee, serving from July 2009 for a five year year. Thanks are due to Steve Fienberg for his work on the committee since inception, and whose term ended in 2009.

Third, in July 2009 Michael Jordan and Dongchu Sun begin three year appointments on the ISBA Prize Committee that oversees the processes and people involved in the adminstration of all ISBA professional awards. Thanks are due to Raquel Prado and Chris Carter whose terms ended in June 2009.

One other, key aspect of ISBA organisation is the international Local Chapters that are very active in meeting organisation and professional development for the Bayesian communities around the world. The vitality and growth of Bayesian thinking and applications is reflected in the growth of the profession worldwide, and I believe the Local Chapters will – and certainly should – become more and more central to the role of ISBA as a "hub" of professional activism. Two examples on my mind from recent interactions and discussions are the Indian and Brazilian chapters, each of which is growing and very active in interfacing with national statistical societies as well as the international Bayesian community. I was recently invited to write to members of each of these two chapters in the forthcoming chapter newsletters. With apologies for duplication, I thought it worth restyling some of what I wrote on the growth and vitality of Bayesian analysis, as part of a message aiming to promote ISBA membership and engagement

through the chapters.

Most of us are very well aware of the vast growth and adoption of Bayesian methods in applications in many fields over the last couple of decades. From basic biology to frontier information technology, highly structured stochastic models of increasing realism – often with high-dimensional parameters and latent variables, multiple layers of hierarchically structured random effects, and nonparametric components – are increasingly adopted. For all of us, this is reason for celebration of the success of Bayesian methods in applications. More importantly, in my opinion, is the increased engagement in Bayesian thinking and philosophy that this pragmatic adoption of methodology can engender. We should be increasingly recognising, and celebrating, the progressive breakdown of historical prejudices against Bayesian thinking that has been driven by the increasing adoption of Bayesian models and methods by non-statisticians, and applied statistical researchers from many fields.

Most of us are all also well aware that much of the impetus behind this growth and success of applied Bayesian methods has been access to the increasingly rich array of advanced computational strategies for Bayesian analysis; this has led to increasing adoption of Bayesian methods from heavily practical and pragmatic perspectives. However, I believe that we are now experiencing, or in some fields perhaps beginning to experience, change in statistical science at a more fundamental level – and I believe this is the real reason to celebrate the increasingly adoption of Bayesian methods, since it portends change at deeper, foundational levels.

This view stems, in part, from numerous personal experiences with collaborators and colleagues. As applied researchers become increasing involved in more complex stochastic model building enabled by advanced Bayesian computational methods, they also become more and more engaged in foundational thinking. This engenders an appreciation for the inherent logic and directness of Bayesian model building. Scientifically relevant, highly structured stochastic models are often simply naturally developed from Bayesian formalisms and have overt Bayesian components. Hierarchical models with layers of random effects, random processes in temporal or spatial systems, and large-scale latent variables models of many flavours are just a few generic examples of nowadays standard stochastic structures in wide application, and that are all inherently Bayesian models. The adoption of Bayesian methods from pragmatic viewpoints will "stick" as it engenders deeper, foundational change in scientific philosophy towards a more holistically Bayesian perspective. And this, in turn, has important implications for the core of the discipline; bringing Bayesian methods of stochastic modelling center-stage - with models of increasing complexity and structure for reasons of increased realism - will inevitably re-energize the core of the discipline, presenting new conceptual and theoretical challenges to statistical researchers as applied problems scale in dimension and complexity.

We have seen this in recent years, in several areas. In nonparametric methodology, for example, applied developments have led to a substantial focus on the need for new nonparametric modelling concepts and new theoretical questions; this is energizing theoretical research in Bayesian statistics, and laying foundations for much "core" research in years to come, as well as responding to more immediate applied challenges. Another example is developments in applied probability and stochastic process theory emerging from very practically oriented innovations in Bayesian computation via simulation methods, including Monte Carlo Markov chains and sequential Monte Carlo approaches. The latter, in particular, is an area that is currently exploding in application while driving researchers to dig deeply into novel theoretical and conceptual areas. These are but two examples of how major applied growth and success is beginning to feed back to the core of the discipline. Many of us will extrapolate to predict great vitality in core, foundational and theoretical areas of Bayesian statistics over the coming couple of decades, consistent with the increasing vitality in broader ranges of important and deep applications.

ISBA has a critical place and critical roles in ensuring this; in particular, in promoting and supporting the engagement of new researchers entering our professional playground, and in aiding developments of interconnections to increasingly broaden the interdisciplinary and international presence of Bayesian analysis. ▲

ISBA 2010 World Meeting

by Herbie Lee
Program Council Chair
herbie@ams.ucsc.edu

The ISBA 2010 World Meeting will be held in conjunction with the Ninth Valencia International Meeting on Bayesian Statistics, June 3-8 2010 in Benidorm, Spain. There will be tutori-

als on the first day, contributed talks in the late afternoons, and poster sessions in the evenings. For consideration for a contributed talk, the presenter must be a current member of ISBA, and abstracts will be due December 1, 2009 and the program committee will select among those. All others are encouraged to present their research in the poster sessions. Contributed papers (whether oral or poster) are encouraged to be submitted for publication in Bayesian Analysis, and if published they are eligible for the Lindley Prize. A formal call for submissions is going out via email, and more details will be available at the ISBA 10 web site.

Report from the Ad-Hoc Committee on Membership

by Hedibert Lopes
Ad-Hoc Membership Committee Chair
hedibert.lopes@chicagobooth.edu

As reported in the first Bulletin of 2009, this committee was created to help to increase our societal focus on ensuring growth of the membership. The committee was charged to embark on discussions about strategies for consolidating and expanding the membership, and on expanding the appreciation of "what ISBA does for the members". The committee was asked to present biannual reports to the Board to, in part, promote further discussion by the Board for activities to enrich the "member services" and hence the membership appeal.

We have had discussions on several specific topics as well as general questions, summarized in our June 2009 report to the ISBA Board, recently submitted (and that will be updated over the coming 6 months). The broad areas of discussion, with some specifics, are as follows.

1. Membership cost when compared to other statistical associations, with no suggestions for change.

2. Outreach to other societies/organizations and their members, as well as graduate departments, with a primary suggestion that a formal ISBA "flyer" could be useful, as well as asking all members of ISBA to help in promoting ISBA membership among other inter-related professional communities and groups.

3. The potential for a 4-year, reduced fee membership for certified students.

4. Potential coordination of recurrent Bayesian meetings supported by ISBA in "odd numbered" years to coordinated with the "even numbered" World Meetings.

5. Regularizing the benefits to ISBA members offered in connection with World Meetings and also ISBA sponsored meetings. Among the benefits are or might be regular reduced registration fees, and ISBA member-only eligibility, as well as ISBA member-only junior researcher travel awards.

6. Potential to expand interactions with other professional societies concerning collaborative membership sign-up with discounts for parallel sign-up.

7. Broader and expanded use of the ISBA web site to deliver member-only access and services.

At this point, this committee is about to embark on the creation of an initial ISBA flyer as mentioned above. This flyer (to be emailed broadly via membership, other organizations ISBA links to, broader email lists maintained by ISBA, statistical and allied departments around the world, etc.) will be electronic, and should contain permanent information high-lighting benefits of ISBA membership, linking to the ISBA web site and some nice "visuals". It could also contain year-to-year updated information on major events - awards, meetings, etc. In the coming months a subgroup of committee members will draft an initial flyer. Members interested in helping - we are looking for Bayesian graphic designers! - or sending comments or ideas should feel free to contact one or more members of the committee.

Ad-hoc Membership Committee:
David Dunson
Sylvia Fruhwirth-Schnatter
Lurdes Inoue
Brunero Liseo
Hedibert Lopes (Committee Chair)
David van Dyk

ANNOTATED BIBLIOGRAPHY

BAYESIAN METHODS FOR PROTEIN STRUCTURE PREDICTION

David B. Dahl
dahl@stat.tamu.edu

Protein structure prediction methods seek to predict the three-dimensional structure of a protein from its amino acid sequence. Since the structure of a protein determines its function in the cell, protein structure prediction is very important in biomedical, pharmaceutical, and biotechnology applications. Technological developments in biochemistry have led to an explosion of genomic data, but experimental methods to solve protein structures have not kept pace. To accelerate the process, protein structure prediction methods aim to construct accurate models of a target protein's native state using only the protein's amino acid sequence. The training dataset for such methods is the Protein Data Bank, a repository of proteins whose amino acid sequence is known and whose structure has been determined experimentally.

By way of background, protein structure is typically described in terms of four categories: primary through quarternary. Primary structure consists of the linear sequence of covalently bonded amino acids that make up a protein's polypeptide chain. Secondary structure describes the regularly repeating local motifs of α-helices, β-strands, turns, and coil regions. For a single polypeptide chain, tertiary structure describes how the secondary structure elements arrange in three-dimensional space to define a protein's fold. By allowing the polypeptide chain to come back on itself, the loops and turns effectively define the arrangement of the more regular secondary structure of α-helices and β-strands. Quarternary structure describes how multiple folded polypeptide chains interact with one another. In a typical structure prediction problem the primary structure (i.e., amino acid sequence) is known, and the goal is to use this information to predict the secondary or tertiary structure.

Bayesian methods have proven to be very successful in integrating the various genomic data and the known structures from the Protein Data Bank to make good structure predictions. There are, however, many challenges and opportunities that remain. This annotated bibliography seeks to highlight several Bayesian papers to get interested Bayesian statisticians started in the field of protein structure prediction. Also listed (without annotation) are methods for the closely related problem of protein alignment. The lists focus on Bayesian papers and, even then, is in no way comprehensive. I apologize to those whose work I may have omitted. Lastly, I would like to thank my biochemistry collaborator Jerry Tsai (University of the Pacific) for introducing me to this field and my statistical collaborators Kristin Lennox (Ph.D. candidate at Texas A&M University) and Marina Vannucci (Rice University).

Tertiary Structure Prediction

- K. T. Simons, C. Kooperberg, E. Huang, D. Baker (1997). Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions, *Journal of Molecular Biology*, 268, 209-225.

  The David Baker lab has been highly successful in the protein structure prediction field. This paper uses Bayes rule to score candidate predictions generated from a simulated annealing algorithm that assembles native-like structures from fragments of unrelated protein structures with similar local sequences.

- S. C. Schmidler, J. S. Liu, D. L. Brutlag (2001). Bayesian Protein Structure Prediction, *Case Studies in Bayesian Statistics*, 5, 363-378.

  This paper, presented at the Carnegie Mellon University's 1999 Case Studies in Bayesian Statistics Workshop, generalizes the author's 2000 paper (see the next section) and models the β-strand pairing to predict the three-dimensional contacts of proteins.

- H. Singh, V. Hnizdo, E. Demchuk (2002). Probabilistic Model for Two Dependent Circular Variables, *Biometrika*, 89, 719-723.

Although not a Bayesian paper itself, this paper provides the "sine model," a particular bivariate von Mises distribution used in the Bayesian nonparametric models of Lennox, et. al. (2009a, 2009b) and Dahl et. al. (2008).

- K. V. Mardia, C. C. Taylor, G. K. Subramaniam (2007). Protein Bioinformatics and Mixtures of Bivariate Von Mises Distributions for Angular Data, *Biometrics*, 63: 505-512.

  This non-Bayesian paper is an important reference to which the Bayesian models below are naturally compared.

- D. B. Dahl, Z. Bohannan, Q. Mo, M. Vannucci, J. W. Tsai (2008). Assessing Side-Chain Perturbations of the Protein Backbone: A Knowledge Based Classification of Residue Ramachandran Space, *Journal of Molecular Biology*, 378, 749-758.

  Using a Dirichlet process mixture of bivariate normal distributions, this paper provides a method for nonparametric density estimation of the (ϕ, ψ) torsion angles of the three-dimensional protein backbone at a given amino acid residue. The results imply that side-chain steric effects strongly influence a residue's backbone torsion angle conformation.

- K. P. Lennox, D. B. Dahl, M. Vannucci, J. W. Tsai (2009). Density Estimation for Protein Conformation Angles Using a Bivariate von Mises Distribution and Bayesian Nonparametrics, *Journal of the American Statistical Association*, 104, 586-596.

  Recognizing that (ϕ, ψ) torsion angles exhibit periodic behavior (e.g., that angles $-\pi$ and π are the same), this paper uses a bivariate von Mises sine model as the component distribution in a Dirichlet process mixture model for nonparametric density estimation. The paper demonstrates that half position data provides a better approximation for the distribution of conformational angles at a given sequence position, therefore providing increased efficiency and accuracy in structure prediction.

- D. B. Dahl, R. Day, J. W. Tsai (2008). Distance-Based Probability Distribution on Set Partitions with Applications to Protein Structure Prediction, *preprint, available from the author*.

  This paper defines a new class of Bayesian nonparametric models that utilizes distance-based probability distribution over partitions as a prior clustering distribution. The method is applied to a model for protein structure prediction and is shown to substantially improve predictive accuracy.

- K. P. Lennox, D. B. Dahl, M. Vannucci, R. Day, J. W. Tsai (2009). A Dirichlet Process Mixture of Hidden Markov Models for Protein Structure Prediction, *preprint, available from the author*.

  Previous work has modeled (ϕ, ψ) torsion angles at a single sequence position. This paper proposes a new semiparametric model for the joint distributions of angle pairs at multiple sequence positions, permitting the sharing of information across sequence positions. Results show this strategy successfully models the notoriously difficult loop and turn regions.

Secondary Structure Prediction

- S. C. Schmidler, J. S. Liu, D. L. Brutlag (2000). Bayesian Segmentation of Protein Secondary Structure, *Journal of Computational Biology*, 7, 233-248.

  This seminal paper presents a secondary structure prediction method based on a Bayesian model of protein sequence/structure relationships in terms of structural segments. The model is Markovian in the segments, permitting efficient exact calculation of the posterior probability distribution over all possible segmentations of the sequence using dynamic programming.

- W. Chu, Z. Ghahramani, A. Podtelezhnikov, D. L. Wild (2006). Bayesian Segmental Models with Multiple Sequence Alignment Profiles for Protein Secondary Structure and Contact Map Prediction, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3, 98-113.

  This paper extends the work by Schmidler, Liu, and Brutlag (2000, 2001) to incorporate the multiple alignment sequence pro-

files into the semi-Markov model to improve secondary structure prediction.

- Z. Aydin, Y. Altunbasak, H. Erdogan (2007). Bayesian Protein Secondary Structure Prediction With Near-Optimal Segmentations, *IEEE Transactions on Signal Processing*, 55, 3512-3525.

  This paper proposes an alternative decoding technique for the hidden semi-Markov model originally employed in the BSPSS algorithm of Schmidler, Liu, and Brutlag (2000). The alternative is based on an N-best paradigm to compute the most likely segmentations.

Protein Alignment

- J. Zhu, J. Liu, C. Lawrence (1998). Bayesian Adaptive Sequence Alignment Algorithms, *Bioinformatics*, 14, 25-39.

- J. Liu, C. Lawrence (1999). Bayesian Inference on Biopolymer Models, *Bioinformatics*, 15, 38-52.

- B. Webb, J. Liu, C. Lawrence (2002). Balsa: Bayesian Algorithm for Local Sequence Alignment, *Nucleic Acids Research*, 30, 1268-1277.

- P. J. Green, K. V. Mardia (2006). Bayesian Alignment Using Hierarchical Models, with Applications in Protein Bioinformatics, *Biometrika*, 93, 235-254.

- A. Rodriguez, S. C. Schmidler (2009). Bayesian Protein Structure Alignment, *Annals of Applied Statistics*, *submitted*.

BAYESIAN HISTORY

A CALL FOR CONTRIBUTORS

by Timothy D. Johnson
tdjtdj@umich.edu

The editor of the ISBA Bulletin and I would like to make a general call for contributors to the Bayesian History section of this bulletin. We would like to propose that this section be devoted to the historical development of Bayesian theory and methods in particular countries—yours, for example.

This one to two page article could consist of an overview of the development of Bayesian methods in your particular country, a synopsis of the contributions to a particular aspect of Bayesian methods/theory that have taken place in your country or by fellow countrymen or an interview with a prominent statistician.

We are also open to suggestions and articles pertaining to the general history of Bayesian statistics. So, if you have an idea, or wish to contribute to this section of the ISBA Bulletin, please contact me.

APPLICATIONS

HIERARCHICAL SPATIAL MODELS FOR LARGE DATASETS IN FORESTRY

Sudipto Banerjee
sudiptob@biostat.umn.edu

Recent advances in Geographical Information Systems (GIS) and Global Positioning Systems (GPS) enable accurate geocoding of locations where scientific data are collected. This has encouraged formation of large spatiotemporal datasets in many fields and has generated considerable interest in statistical modelling for location-referenced spatial data. Among other fields, forestry offers numerous intesting scientific questions that require models for spatial data. In order to better understand the role of forests in the global carbon cycle, scientists collect measurements on several important variables so as to identify new agents of environmental change and also to illuminate relationships that may exist between them. Such findings can be instrumental in helping policy mak-

ers chart out future plans for environmental protection as well as energy production. In recent times, I have had the opportunity to collaborate with Professor Andrew Finley on developing hierarchical (Bayesian) spatial process models on a variety of problems in forestry. Here, I briefly discuss some of those projects.

One challenging problem is to understand how certain scientific variables generated from different species of trees are related to other forest attributes and how to accurately predict these relationships over space. For instance, forest biomass, which provides a measure of carbon accumulation in the trees, is an important scientific variable that is central to understanding the global carbon cycle. Spatial modelling of forest biomass and other variables related to measurements of current carbon stocks and flux have recently attracted much attention for quantifying the current and future ecological and economic viability of forest landscapes. Interest often lies in detecting how biomass changes across the landscape (as a continuous surface) and how homogeneous it is across the region.

Another area of application concerns the construction of spatially explicit data layers of tree species assemblages, referred to as forest types or forest type groups (FTG). These constitute a key component in large-scale assessments of forest sustainability, biodiversity, timber biomass, carbon sinks, and forest health monitoring. National Forest Inventories (NFIs) sample populations of interest and report plot-based estimates of forest resources. Spatial model-based approaches to mapping are attractive here as they can depict spatial distributions of forest attributes and can easily incorporate ancillary variables or flexible spatial dependence structures to improve the accuracy and precision of parameter estimates and/or prediction.

The above problems share the underlying theme of analyzing spatially-referenced variables, but are also different in some aspects. For instance, forest biomass can be treated as continuous variables, while forest types are categorical. Furthermore, in terms of meeting the scientific objectives, the statistical methods would tend to focus more upon improved predictive and estimative performances in the first example, and better classificaton in the second.

Hierarchical (Bayesian) spatial process models (e.g. Cressie, 1993; Stein, 1999; Banerjee et al., 2004) provide a powerful and very flexible framework to model the underlying physical processes generating spatial data from diverse settings and with varied objectives – such as those we briefly described above. Spatial process models envision a *random surface* $w(\mathbf{s})$ that conceptually exists in continuum over the study domain $\mathbf{s} \in D$. Using the partial information available from the observed data over a finite set of locations, we can estimate this surface and carry out spatial interpolation/prediction at arbitrary locations. By interpolating at arbitrarily fine resolutions, these models *estimate* the random surface accounting for correlation at locations closer to each other and produce a *response surface* for the dependent variable. Such interpolation from statistical models is often referred to as "kriging" and the response surfaces are called kriged surfaces.

The spatial process $w(\mathbf{s})$ also determines the distribution of spatial random effects, providing local adjustment (with structured dependence) to the mean, often interpreted as capturing the effect of unmeasured or unobserved covariates with spatial pattern. The customary process specification for $w(\mathbf{s})$ is a mean 0 Gaussian Process with covariance function, $C(\mathbf{s}_i, \mathbf{s}_j)$, which is denoted $GP(0, C(\mathbf{s}_i, \mathbf{s}_j))$. This implies that for an arbitrary collection of locations, say $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, the distribution of $\mathbf{w} = \{w(\mathbf{s}_i)\}$ is a multivariate normal distribution with zero mean vector and a dispersion matrix $\Sigma_{\mathbf{w}}$ whose ij-th element is given by $C(\mathbf{s}_i, \mathbf{s}_j)$. The covariance function $C(\mathbf{s}_i, \mathbf{s}_j) = \text{cov}\{w(\mathbf{s}_i), w(\mathbf{s}_j)\}$ needs to be a symmetric positive definite function (that ensures positive definiteness of $\Sigma_{\mathbf{w}}$, not just for a specific set of locations, but for *any* set of locations). The $C(\mathbf{s}_i, \mathbf{s}_j)$ are often expressed as $\sigma^2\rho(\mathbf{s}_i, \mathbf{s}_j)$, where σ^2 is a spatial variance term and $\rho(\mathbf{s}_i, \mathbf{s}_j)$ is a spatial *correlation* function. Such functions have been well-studied in complex analysis and are characterized by a well-known theorem due to Bochner (see, e.g., Stein, 1999) as characteristic functions of symmetric random variables. Choices range from very simple exponential decay functions to quite complicated nonstationary Matérn functions (Paciorek and Schervish, 2006) providing for a range of modelling needs. When the covariance depends only upon the separation between the sites, the underlying process is said to be weakly stationary; otherwise it is said to be nonstationary.

As a simple and more concrete example, suppose $y(\mathbf{s}_i)$ is an observation of a contniuous outcome (e.g. forest biomass) at location $\mathbf{s}_i$. Then, using standard notations, and the above proper-

ties of a Gaussian process, we can write down a spatial hierarchical model as

$$p(\tau^2) \times p(\beta) \times p(\theta) \times N\{\mathbf{w} \,|\, \mathbf{0}, \Sigma_{\mathbf{w}}\} \times \prod_{i=1}^{n} N\left\{y(\mathbf{s}_i) \,|\, \mathbf{x}(\mathbf{s}_i)'\beta + w(\mathbf{s}_i), \tau^2\right\}, \quad (1)$$

where θ denotes the set of parameters specifying the spatial process and τ^2 is the measurement error variance (often called the "nugget"). Full Bayesian inference and accurate assessment of uncertainty for models such as (1) will require Markov chain Monte Carlo methods (Banerjee et al., 2004). For a large number of spatial locations, fitting customary geostatistical models becomes prohibitive with necessary matrix factorizations of cubic order complexities. Without further specifications, estimating (1) will involve linear solvers or Cholesky decompositions of computational complexity $O(n^3)$, once every MCMC iteration, to produce estimates of θ. With large n, this is simply infeasible.

Modelling large spatial datasets have received much attention in the recent past and an exhaustive survey of the various approaches is beyond the scope of this article. In seeking an approach that would be rich, flexible and would adapt seamlessly to the diverse settings tackled by hierarchical spatial process models, we arrived at a class of models that are variants of the so-called "subset of regressors" methods used in Gaussian process regressions in machine learning (Wahba, 1990; Rasmussen and Williams, 2006). The essential idea is to consider a smaller set of locations, or "knots", say $\mathcal{S}^* = \{\mathbf{s}_1^*, \ldots, \mathbf{s}_{n^*}^*\}$, where the number of knots, n^*, is much smaller than the number of observed sites, and to express the spatial process realizations over $\mathcal{S}$ in terms of its realizations over the smaller set of knots. Different authors have proposed and investigated alternate strategies to build sparser models using these knots. These are often referred to as low-rank or reduced-rank spatial (see, e.g., Kamman and Wand, 2003; Stein, 2007; Cressie and Johannesson 2008). Banerjee et al. (2008) suggest looking at a spatial process that operates on a lower-dimensional subspace. They refer to $w(\mathbf{s})$ as the *parent process* and derive a *predictive process* simply as a conditional expectation of the parent process given realizations of the original process. This is equivalent to setting $\tilde{w}(\mathbf{s}) = \mathbf{c}(\mathbf{s})'\Sigma_{\mathbf{w}^*}^{-1}\mathbf{w}^*$, where $\mathbf{c}(\mathbf{s};\theta)$ is an $n^* \times 1$ vector with i-th element given by $C(\mathbf{s}, \mathbf{s}_i^*)$, $\Sigma_{\mathbf{w}^*}$ is an $n^* \times n^*$ matrix with $C(\mathbf{s}_i^*, \mathbf{s}_j^*)$ as its ij-th element and $\mathbf{w}^* = \{w(\mathbf{s}_i^*)\}$ is the $n^* \times 1$ vector of realizations of the parent Gaussian process over the knots. Hence $\mathbf{w}^*$ is distributed as a multivariate normal with covariance matrix $\Sigma_{\mathbf{w}^*}$.

Replacing $w(\mathbf{s})$ in (1) with $\tilde{w}(\mathbf{s})$, the corresponding predictive process model becomes

$$p(\tau^2) \times p(\beta) \times p(\theta) \times N\{\mathbf{w}^* \,|\, \mathbf{0}, \Sigma_{\mathbf{w}^*}\} \times \prod_{i=1}^{n} N\left\{y(\mathbf{s}_i) \,|\, \mathbf{x}(\mathbf{s}_i)'\beta + \mathbf{z}(\mathbf{s}_i)'\mathbf{w}^*, \tau^2\right\} \quad (2)$$

where $\mathbf{z}(\mathbf{s}_i)' = \mathbf{c}(\mathbf{s}_i)'\Sigma_{\mathbf{w}^*}^{-1}$, we see that $\tilde{w}(\mathbf{s})$ is a spatially varying linear transformation of $\mathbf{w}^*$. The dimension reduction is seen immediately. In fitting the model in (2), the n random effects $\{w(\mathbf{s}_i), i = 1, 2, ..., n\}$ are replaced with only the n^* random effects in $\mathbf{w}^*$; we can work with an n^* dimensional joint distribution involving only $n^* \times n^*$ matrices. No new parameters are introduced in (2), hence one need not worry about identifiability issues. In fact, with X the $n \times p$ matrix of regressors and Z the $n \times n^*$ matrix with $\mathbf{z}(\mathbf{s}_i)'$ as its rows, the predictive process likelihood can be cast into a linear mixed model framework: $\mathbf{y} = X\beta + Z\mathbf{w}^* + \epsilon$, with $\mathbf{w}^* \sim N(\mathbf{0}, \Sigma_{\mathbf{w}^*})$ and $\epsilon \sim N(\mathbf{0}, \tau^2 I)$. Note that the matrix Z depends upon the spatial process parameters. The text by Ruppert, Wand and Carroll (2003) provides an excellent treatment of the different statistical methods (Bayesian and classical) to analyze such models.

The predictive process approach applies to multivariate spatial and spatiotemporal process models with equal ease. In Banerjee et al. (2008), we extended (1) to a spatially-varying regression model for biomass, where the regression parameters are jointly specified through a multivariate Gaussian process. The Bayesian approach is attractive here as it offers exact inference for the random spatial coefficients by delivering an entire posterior distribution at both observed and unobserved locations. Spatial interpolation for processes that are neither observed nor arise as residuals appears inaccessible with classical likelihood-based methods. While this allows us to capture how the impact of specific covariates vary over space, the models become prohibitively expensive to estimate. In fact, we used $9,500$ locations obtained from the USDA Forest Service Forest Inventory and Analysis (FIA) program. Each location yields measurements on biomass from trees in that location and two regressors: the cross-sectional area of all stems above 1.37 meters off the ground (basal area), and

the number of tree stems (stem density) at that location. Along with an intercept, this results in three spatially-varying regression parameters and, in full generality, would involve $9,500 \times 3$ (i.e. $28,500$) dimensional matrix computations! Therefore, we used the predictive process versions of these models to handle the computations (see Banerjee et al., 2008, for further computational details and the selection of knots). Figure 1 displays three digital image-plots, overlaid with contour lines. The left panel shows a predicted spatial map of forest biomass from the species "beech". Yellow zones represent higher levels of biomass while red zones are lower levels. The other two panels show how the impact of basal area and stem density varies spatially – yellow zones represent regions of less significant impact, while red zones represent regions of more significant impact. Wheeler and Waller (2008) provide some nice insight into inference and interpretations of spatially-varying coefficient models.

Since, the predictive process, or, for that matter, any other low-rank smoother, tends to oversmooth the data, this results in an underestimation of spatial variability. This residual variability is often absorbed by the residual unstructured variance component (e.g. τ^2 in model 2) and is manifested by a sytematic upward bias in the measurement error variance. A simple rectification is to add a rather special structured noise to the predictive process. This leads to a modified predictive process $\tilde{w}_{\tilde{\epsilon}}(\mathbf{s}) = \tilde{w}(\mathbf{s}) + \tilde{\epsilon}(\mathbf{s})$, where $\tilde{\epsilon}(\mathbf{s})$ are independently distributed as zero-mean normal distributions with variance given by $C(\mathbf{s},\mathbf{s}) - \mathbf{c}(\mathbf{s})'\Sigma_{\mathbf{w}^*}^{-1}\mathbf{c}(\mathbf{s})$. Finley, Banerjee, Waldmann and Ericsson (2009) demonstrate the improved performances of the noise-added predictive process models in analyzing spatial datasets arising from tree genetic improvement initiatives here in the United States and Sweden. Recent interest in promoting bio-economies has spurred renewed interest in wood fiber production and in genetic improvement studies. Field measurements from a 26-year-old scots pine (*Pinus sylvestris* L.) progeny study in northern Sweden served as a trial dataset for our proposed methods. The analysis involved over 8000 spatial locations and preliminary explorations suggested somewhat more complex models that accommodated *anisotropy*, i.e. the strength of spatial correlation depended upon direction as well as distance. Again, predictive process models were necessary to resolve the computational bottleneck.

A classic paper by Diggle et al. (1998) discussed the use of spatial process models for non-Gaussian data within the framework of generalized linear models. For non-Gaussian data, the likelihood in (1) will simply be changed to an exponential family likelihood and the regression (including the spatial process) modeled through a link function. The corresponding predictive process models would then arise exactly analogous to (2), but again with the likelihood a member of the exponential family.

One important distinction between the models for Gaussian and non-Gaussian variables is worth pointing out here. For the latter, we introduce spatial random effects in the transformed mean encourages the means of spatial variables at proximate locations to be close to each other – not the observed values of the variables themselves. While marginal spatial dependence between the observed variables is induced, their observed values need not be close. Therefore, unlike in (1) and (2), it does not make sense to incorporate measurement error in the likelihood and, hence, there is no τ^2. Only one (spatial) variance component is modeled using the spatial process. Finley, Banerjee and McRoberts (2009) explored the utility of coupling georeferenced national forest inventory (NFI) data with readily available and spatially complete environmental predictor variables. They achieve this by developing a class of spatially-varying multinomial-logistic regression models to predict forest type groups (seven categories) across large forested landscapes. The richness of these models incurs onerous computational burdens and, again, a rectified predictive process is employed to achieve dimension reduction.

Finally, it is worth pointing out that we continue to make our modeling advancements available to the scientific community through the `R` package `spBayes`. This is available through the Comprehensive R Archive Network (CRAN) (see `http://cran.r-project.org/web/packages/spBayes`). A particularly useful function in the package is `spLM`. It uses MCMC algorithms to estimate Bayesian spatial regression models as well as their predictive process counterparts for larger datasets. Multivariate spatial regression models (with multivariate predictive processes) and some generalized spatial linear models are also accommodated through the `spMvLM` function. Functions that assist in knot-selection for predictive processes are also provided. Full posterior predic-

tive distributions can be computed for prediction and spatial interpolation and the Deviance Information Criteria (DIC) (Spiegelhalter et al., 2002) is available for model comparisons. To harness maximum computing speed, all the underlying functions are all written in C++ which are called by the R interface. The user, however, does not need to know C++ to use these functions.

References

- Banerjee S., Carlin, B.P. and Gelfand, A.E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: Chapman and Hall/CRC Press.
- Banerjee, S., Gelfand, A.E., Finley, A.O. and Sang, H. (2008). Gaussian predictive process models for large spatial datasets. *Journal of the Royal Statistical Society Series B* **70** 825–848.
- Cressie, N., (1993). *Statistics for Spatial Data*. Second edition. New York: Wiley.
- Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B* **70** 209–226.
- Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998). Model-based geostatistics (with discussion). *Applied Statistics* **47** 299–350.
- Finley, A.O., Banerjee, S., Waldmann, P. and Ericsson, T. (2009). Hierarchical spatial modeling of additive and dominance genetic variance for large spatial trial datasets. *Biometrics*, **65**, 441–451.
- Finley, A.O., Banerjee, S. and McRoberts, R.E. (2009). Hierarchical spatial models for predicting tree species assemblages across large domains. *Annals of Applied Statistics* (in press).
- Kamman, E.E. and Wand, M.P. (2003). Geoadditive models. *Applied Statistics* **52** 1–18.
- Paciorek, C.J. and Schervish, M.J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, **17**, 483–506.
- Rasmussen, C.E. and Williams, C.K.I. (2006) *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.
- Ruppert, D., Wand, M.P. and Caroll, R.J. (2003). *Semiparametric Regressgion*. Cambridge, UK: Cambridge University Press.
- Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B* **64** 583–639.
- Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory of Kriging*. New York: Springer.
- Stein, M.L. (2007). Spatial variation of total column ozone on a global scale. *Annals of Applied Statistics* **1** 191–210.
- Wahba, G. (1990). Spline models for observational data. SIAM, Philadelphia.
- Wheeler, D.C. and Waller, L.A. (2008). Comparing spatially varying coefficient models: a case study examining violent crime rates and their relationships to alcohol outlets and illegal drug arrests. *Journal of Geographical Systems*, **11**, 1–22.

SOFTWARE HIGHLIGHT

RAMPS: UNIFIED BAYESIAN GEOSTATISTICAL MODELING OF COMPLEX SPATIOTEMPORAL DATA

by Brian Smith, Kate Cowles, and Jun Yan

The **ramps** R package consists of a suite of tools for Bayesian geostatistical analysis of spatiotemporal data (1). In particular, it allows for linear modeling of point-source and/or areal measurements as a function of fixed covariates, cluster-specific random effects, spatiotemporal random effects, and measurement error. MCMC methods are used to sample from the posterior distribu-

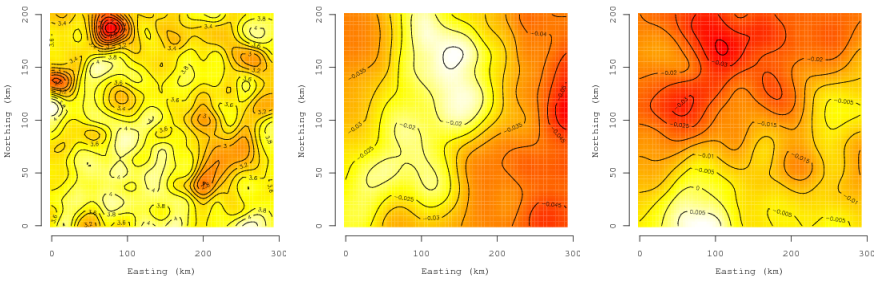


Figure 1: Left: Image-plots of forest biomass of beech; Center: impact of basal area; Right: impact of stem-density

tion (2), thus making possible fully Bayesian inference of all model parameters and model-based predictions. Predictions can be made at any arbitrary site, measured or unmeasured, and visualized with 2D and 3D graphics functions supplied by the package. Additional integration with the **coda** package (3) is provided to facilitate summary, plotting, and convergence diagnostics of generated MCMC output.

Advantages of the **ramps** package include its general class of geostatistical model, object oriented interface, extensible correlation structures, and efficient MCMC routines based on the Reparameterized and Marginalized Posterior Sampling (RAMPS) algorithm of Cowles et al. The package is open-source and publicly available from the Comprehensive R Archive Network at

http://cran.R-project.org/package=ramps

Model Specification

Implemented in **ramps** is the geostatistical model of the general form

$$Y = X\beta + W\gamma + KZ + \varepsilon$$
$$\gamma \sim N(0, \Sigma_\gamma),\ Z \sim N(0, \Sigma_Z),\ \varepsilon \sim N(0, \Sigma_\varepsilon)$$

where the response vector Y can include point-source measurements, areal measurements, or a combination thereof. The model components include the following:

- *Main effects* (β) to account for mean effects of fixed covariate or transformations of covariate values given in the design matrix X.
- *Exchangeable random effects* (γ) for cluster-specific random intercepts (optional), mapped to measurement values via an indicator matrix W. Associated variances may be allowed to vary between groups of clusters.
- *Spatiotemporal random effects* (Z) characterized by a Gaussian random field in which correlation is a function of distance (and potentially direction) between points. Each measurement can be associated with a single random effect or a linear combination of random effects via the design matrix K. Spatial variances may vary between regions.
- *Measurement error* (ε) to capture unexplained measurement-specific variability. Measurement error variances may vary between groups of measurements.

A wide range of parametric spatial correlation functions is supplied with the package, including Gaussian, exponential, powered exponential, spherical, Matérn, rational quadratic and sine wave. Separable and non-separable spatiotemporal functions are also supplied. Moreover, spatial distance may be calculated as Euclidean or great-circle (haversine formula) distance to accommodate measurements from flat or spherical surfaces. Correlation functions are implemented as R objects, so that users are free to define their own without needing to make changes to the package's source code. Consequently, users can extend the model to make available new correlation structures as they are developed or needed in practice.

Data Fusion Example

Consider county-averaged (4) and point-source (5) uranium measurements (ppm) available for the U.S. state of Connecticut. Since the source

for both types of measurements is uranium deposited in the soil, a joint analysis of the two should allow for better estimation of the underlying spatial distribution. County-averaged measurement Y_{1i} is modeled as

$$\begin{aligned} \ln(Y_{1i}) &= \beta_1 + K_{1i}^\top Z + \varepsilon_{1i} \\ \varepsilon_{1i} &\sim N(0, \sigma_1^2/m_i) \end{aligned}$$

and point-source measurement Y_{2i} as

$$\begin{aligned} \ln(Y_{2i}) &= \beta_2 + K_{2i}^\top Z + \varepsilon_{2i} \\ \varepsilon_{2i} &\sim N(0, \sigma_2^2) \end{aligned}$$

where vector $K_{1i}^\top$ averages over spatial sites giving rise to the i^{th} county-averaged measurement, m_i is the land mass area of the county, and indicator vector K_{2i} associates the i^{th} point-source measurement with the spatial random effect for that location. Thus, separate intercepts and error variances are allowed for the two types of measurements, but a common distribution is assumed for the spatial random effects, i.e.

$$Z \sim N(0, \sigma_Z^2 R(\phi))$$

with spatial correlation matrix $R(\phi)$ defined by an exponential correlation function with decay parameter ϕ. To complete the Bayesian model specification, flat prior distributions are placed on the β intercepts, inverse gamma $IG(2.0, 0.1)$ for the error and spatial variances, and a uniform for ϕ.

Illustrated below is the usage of **ramps** functions to model jointly the two types of measurements. The datasets used in this example include one containing measurements and covariates (`NURE`) and another containing grid locations over which county measurements are assumed to be averaged (`NURE.grid`). Common to the two datasets is the variable `"id"` which identifies the counties associated with averaged measurements and grid locations, respectively. Model fitting is carried out with a two step process. Step one is to create an object that defines MCMC control parameters, initial parameter values, and prior distributions, as follows.

```
NURE.ctrl <- ramps.control(
  iter = 2500,
  beta = param(c(0, 0), "flat"),
  sigma2.e = param(c(1, 1), "invgamma",
                   shape=2.0, scale=0.1),
  sigma2.z = param(1, "invgamma",
                   shape=2.0, scale=0.1),
  phi = param(10, "uniform", min=0, max=35)
)
```

Step two is to define the model itself and to generate MCMC samples from the posterior distribution.

```
NURE.fit <- georamps(
  log(ppm) ~ factor(measurement) - 1,
  correlation = corRExp(form = ~ lon + lat,
                        metric = "haversine"),
  variance = list(fixed = ~ measurement),
  weights = area * (measurement == 1)
            + (measurement == 2),
  data = NURE,
  aggregate = list(grid = NURE.grid,
                   blockid = "id"),
  control = NURE.ctrl
)
```

Detailed descriptions of these functions and the datasets can be found in the documentation supplied with the package. Posterior summaries, model goodness-of-fit measures, and plots of the predicted spatial distribution can be produced from the fitted model object. An example contour plot from the uranium analysis is shown in Figure 2.

References

[1] Smith, B.J., Yan, J. and Cowles, M.K. (2008) Unified Geostatistical Modeling for Data Fusion and Spatial Heteroskedasticity with R package **ramps**. *Journal of Statistical Software*, **25**(10), 1–21, http://www.jstatsoft.org/v25/i10

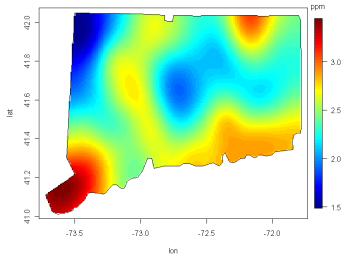
[2] Cowles, M.K., Yan, J. and Smith, B.J. (2009) Reparameterized and Marginalized Posterior and Predictive Sampling for Complex Bayesian Geostatistical Models. *Journal of Computational and Graphical Statistics*, (In Press)

[3] Plummer, M., Best, N., Cowles, M.K. and Vines, K. (2006) **CODA**: Convergence Diagnostics and Output Analysis for MCMC. *R News*, **6**(1), 7–11, http://cran.R-project.org/package=coda

[4] Duval, J.S., Jones, W.J., Riggle, F.R. and Pitkin, J.A. (1989) Equivalent uranium map of conterminous United States. USGS Open-File Report 89-478

[5] Smith, S.M.(2006) National Geochemical Database Reformatted Data from the National Uranium Resource Evaluation (NURE) Hydrogeochemical and Stream Sediment Reconnaissance (HSSR) Program. USGS Open-File Report 97-492

Figure 2: Posterior mean predicted uranium concentrations.



STUDENTS' CORNER

Luke Bornn
l.bornn@stat.ubc.ca

This Students' Corner features an article from Gareth Peters detailing some excellent tips on succeeding in graduate school. We also feature thesis abstracts from two recent PhD graduates. As always, if you (or your student) have recently graduated and would like your dissertation abstract published, please contact me.

GETTING THE MOST OUT OF GRADUATE SCHOOL

by Gareth Peters
peterga@maths.unsw.edu.au
http://web.maths.unsw.edu.au/~peterga
University of New South Wales

I was asked by Luke Bornn to write a short section on the graduate school experience and to share some perspectives from my experiences. I thought it would be good to do this before my graduate experience comes to an end in July, when I begin a new chapter as a lecturer in statistics at University of NSW. I have been lucky enough to have studied at universities in Australia (Monash University; The University of Melbourne; Univeristy of New South Wales), UK (Cambridge) and Canada (University of British Columbia), in addition during this time I have spent extended research stays at LSE, Toulouse, France; RiskLab in ETH, Zurich; SAMSI, in North Carloina, US; and ISM, Tokyo, Japan. I will share some thoughts on what I have learned from these very different experiences. I will then finish with a brief list of some additional resources and guides that may prove valuable.

The graduate experience can be deeply rewarding, exciting, stimulating academically, challenging and frightening all at the same time. For the first time in your life you will be responsible directly for your own fate as a researcher, this can be both exhilarating and daunting. What you learn in graduate school will set the foundation for a life-long learning experience and a successful career. My experience has taught me that it takes a true dedication to get the most out of graduate school, this is more than just a good undergraduate preparation and a desire to get an advanced degree. Success in this regard depends to a large extent on what expectations, commitment and discipline you have and how much you demand from yourself.

The ability to do well in graduate school environment requires certain personality traits which I have found to be universal in all good graduate students. These typically include dedication and a strong sustained work ethic, motivated by ones own sense of desire to learn and achieve in research. The discipline required to spend the extra time reading the literature, books, going to talks not just in ones own research field but in related fields, as often great ideas come at the interface of disciplines. Forming strong research practices such as documenting regularly ideas, thoughts and comments on papers you have read, discussing ideas and thoughts regularly with your supervisors and other student colleagues. Below I will split the comments into two sections which include things to consider about graduate school selection and then things to consider once in graduate school.

Pre-graduate school selection.

I will begin with some "universal truths" that I have found to be important throughout the

world. These may seem obvious and you may wonder why I would even comment on them, but I am still constantly surprised by the number of graduate students entering a program who have not carefully thought through their research plan. I think to be successful in any graduate program you must first consider three important aspects which should help guide you in selecting the appropriate program: the first is what area of research do you feel excites you and makes you want to endure the ups and downs that will come with a graduate school experience; which university and which supervisor would place you in a situation to maximise your learning potential in this research field; and thirdly what are the requirements of the graduate program. Whilst considering these, one should also realise that as with anything in life, your research interests will develop and change as a function of your environment, so be flexible and open to all the possibilities that graduate school can provide.

The first aspect mentioned is what I would consider the most important, though surprisingly not often the key element on the agenda of a graduate student. Before making the all important choice of where to spend the next 3 to 6 years, you should seriously consider your answers to the following questions:

If I was asked to outline an area of research that truly excites me and motivates me, what would it be and what aspects would I discuss?

This is really a personal choice and the answer should be influenced largely by your goals especially since you will be directly responsible as a graduate student for directing your life in this direction over the next few years. My experience here involved considering the undergraduate courses and graduate courses I had taken prior to research level study, and thinking which aspects of these courses motivated me. Then this lead to an extensive literature review to examine what exactly was going on in this research field in the past, in the present and who were the main contributors. As a result of this I was able to narrow down some potential supervisors and universities that could provide expertise in this area.

Do you know what to expect from each graduate school and supervisor you are considering?

It is very important to know what to expect from graduate school and what commitments and responsibilities are expected of you. Typically, this may include finding out about aspects such as expected coursework components, qualifying exams, tutoring load, restriction on outside work, scholarships available, research assistant positions and teaching assistant positions. Find out how many students are on average working with each supervisor and check to see how regularly they attend international and national research meetings and conferences. Does the group have an active research presence and how is the structure and support network maintained in the department.

Try to find out about a laboratory or a supervisor by emailing them with your interests. If possible arrange to meet and discuss potential research supervision and projects that you think fit their expertise. If this goes well, you can also ask if they would mind you emailing other graduate students and postdocs for their thoughts on the experience in the labrotory and work environment. This will allow you to gauge to what extent you would integrate into such a work environment. Try to establish an understanding of the research culture and find out how often students meet with supervisors? Is there an active collaboration network and or seminar / reading group in your area of interest ?

Success in a chosen graduate school.

It is critical to start the graduate program with a strong research ethic. Depending on the university one either begins the first day with a research proposal in mind, this is more common in Australia and the UK, or as is the case of Canada, one takes several graduate courses before settling on a thesis topic and supervisor. What will be important in either case is to begin to build a research network and set of collaborators with whom you can work and share ideas. This involves surrounding yourself with other graduates and colleagues who will push you, make you think and provide interesting discussion.

It is important to remember that graduate research is about generating and implementing new concepts. It is not a 9-to-5 job with a well structured direction and a manager who provides deliverables. You must learn to become your own boss and your own project manager. Additionally, it is important to recognize upfront that with the academic freedom that graduate school presents there is also a downside, which for some can prove too much. One must realize that the reality of graduate school involves long hours at a low pay. Hence, contrary to the jokes about the lazy graduate student, to be successful one should expect to work hard, and expect to spend long hours.

Some useful tips that have worked for me have

included keeping detailed research notebooks. You may consider sections related to research papers you have read and your thoughts, your own research interests and ideas to pursue. It is also critical in a computational statistics area to begin early with good coding practice, using programs such as cvs server and always commenting your code, with updates dated and commented will save you a lot of time towards the end of your research degree.

Other tips could include developing a research ethic and routine which involves regularly spending time reading the literature. I cannot stress this enough, it is critical to know what has been done as the wheel does not need to be reinvented. Always remember that as a junior researcher it is now your personal responsibility to ensure you are well read. Attend regularly seminars and make a point to ask relevant questions at conferences. It is also wise to attend the seminars even if the talks are not directly on your research topic. Another useful tip is that whilst you should stay focused on your research, it is also useful to learn about areas other than your own. This is not only motivating, but can make you significantly more employable at the end of your program.

Finally, take as many opportunities to travel and collaborate internationally as possible. It is both a great way to develop a research network and also to experience the world outside of your office. Traveling to work with international collaborators can be both exciting and thought provoking, sometimes all you need to make a research leap is a little fresh air!

I wish you the best of luck in your graduate experience.

References. 1. Careers in Science and Engineering: A Student Planning Guide to Grad School and Beyond, National Academy Press, Washington, DC, 1996. www.nap.edu. 2. On Being A Scientist: Responsible Conduct in Research, National Academy Press, Washington, DC, 1995. www.nap.edu. 3. P. J. Feibelman, PhD Is Not Enough: A Guide to Survival in Science, Addison-Wesley, New York, NY, 1994. 4. P. B. Medawar, Advice to a Young Scientist, Basic Books, New York, NY, 1981. 5. S. R. Covey, A. R. Merrill, and R. Merrill, First Thing First, Simon & Schuster, New York, NY, 1995.

Dissertation Abstracts

BAYESIAN NONPARAMETRIC ANALYSIS OF CONDITIONAL DISTRIBUTIONS AND INFERENCE FOR POISSON POINT PROCESSES

by Matthew Taddy
matt.taddy@chicagobooth.edu
http://faculty.chicagobooth.edu/matt.taddy/
Booth School of Business,
The University of Chicago
PhD Supervisor: Athanasios Kottas & Herbert Lee (UC Santa Cruz)

This thesis provides a suite of flexible and practical nonparametric Bayesian analysis frameworks, together related under a particular approach to Dirichlet process (DP) mixture modeling based on joint density estimation with well chosen kernels and inference through finite stick-breaking approximation to the random mixing measure. Development of a novel nonparametric mean regression estimator serves as an introduction to a general modeling approach for nonparametric analysis of conditional distributions through initial inference about joint probability distributions. Three regression modeling frameworks are proposed: quantile regression, hidden Markov switching regression, and regression for survival data. A related approach is adopted in modeling for marked spatial Poisson processes. This class of models is then expanded to a full nonparametric framework for inference about marked or unmarked dynamic spatial Poisson processes which occur at discrete time intervals. This involves the development of a version of the dependent DP as a prior on the space of correlated sets of probability distributions. Posterior simulation methodology is contained throughout and numerous data examples have been provided in illustration.

HETEROGENEITY IN CAPTURE-RECAPTURE: BAYESIAN METHODS TO BALANCE REALISM AND MODEL COMPLEXITY

by Simon Bonner
s.bonner@stat.ubc.ca
http://www.simon.bonners.ca
University of British Columbia
PhD Supervisor: Carl Schwarz (Simon Fraser)

Capture-recapture experiments are important for monitoring many endangered animal pop-

ulations, such as salmon threatened by overharvesting and migratory songbirds impacted by habitat loss. An important consideration in the analysis of capture-recapture data is potential variation in the probabilities of capture and survival. Failure to account for this variation can lead to incorrect inference, but traditional models incorporating heterogeneity may be very complex. This thesis presents three Bayesian methods that balance realistical modelling of variation in the capture and survival probabilities and increasing model complexity.

In the first project, I consider the analysis of data from two-sample experiments used in estimating the number of juvenile salmon leaving their spawning grounds. These migrations may last for several weeks and standard models may require many parameters to account for variations over time. My solution is to model the population size as a smooth function of time by fitting a Bayesian penalised spline. The method is applied to two datasets from the migration of juvenile salmon and provides more precise estimates of the population size that are less affected by outliers in the data than previous methods.

My second project addresses estimation of the size of an open population when individual capture or survival probabilities are functions of a time-dependent, continuous covariate. The main challenge is that these covariates can only be observed on occasions when an individual is captured. I develop a two-stage Bayesian method that first examines the covariate's effect by analysing the capture of marked individuals, and then applies the results to estimate the total population size. The model is used to study the dynamics of a population of Soay sheep (*Ovis aries*) whose survival is affected by body mass.

Finally, I develop a method to allow more flexibility in modelling the relationship between a covariate and individual survival probabilities. Standard methods assume that the relationship is linear on some scale. My model incorporates Bayesian adaptive splines to allow smooth but local fitting of the linear predictor. I apply this model to study the effect of body condition on the survival of reed warblers (*Acrocephalus scirpaceus*) breeding in Holland.

NEWS FROM THE WORLD

Announcements

I would like to encourage those who have any announcements or would like to draw attention to an up-coming conference, to get in touch with me and I would be happy to place them here.

Savage Award

The finalists for this year's Savage award have been announced. A special session at the 2009 JSM, in Washington, DC, will showcase talks from the four finalist (Session 359; 04/08/2009; 2:00-3:50pm). The winners will be announced at the SBSS mixer later that day. Additional details are available at http://www.bayesian.org/awards/Savage.html#winners

2010 Valencia Conference

This is to announce that the Ninth Valencia International Meeting on Bayesian Statistics and the 2010 ISBA World Meeting will jointly be held in Benidorm (Alicante, Spain), June 3rd to June 8th, 2010. As already announced in Valencia 8, this will be the last Valencia meeting personally organized by José M. Bernardo (who will be 60 when the conference takes place). After Valencia 9, the Valencia meetings will become regular ISBA World Meetings (which will not necessarily take place in the State of Valencia). ISBA world meetings will therefore take place every two years.

For more information visit the website, http://www.uv.es/valenciameeting and for other ISBA sponsored meetings please visit the following page http://bayesian.org/business/meetings.html.

Events

Case Studies in Bayesian Statistics and Machine Learning, Carnegie Mellon University, Pittsburgh, PA. 15-17th October, 2009.

The Workshop will focus on applications of Bayesian statistics and Machine Learning to problems in science and technology. It will

feature three different tracks: In-depth contributed presentations and discussions of substantial research, shorter presentations by young researchers and poster presentations. In conjunction with the workshop, the Department of Statistics' Eleventh Morris H DeGroot memorial lecture will be delivered by Professor Michael Jordan, University of California at Berkeley.

For more information visit the website, http://bayesml1.stat.cmu.edu/, or contact Pierpaolo De Blasi *bnp@carloalberto.org*.

2010 Bayesian Biostatistics Conference, Houston, Texas. 27-29th January, 2010.

Current and prospective users of Bayesian biostatistics are invited to join experts in the field for a three-day conference sponsored by the Department of Biostatistics at The University of Texas M. D. Anderson Cancer Center in Houston, Texas, USA. Attendees will have the opportunity to attend two courses on the first day of the conference (Wednesday): The Use of Bayesian Statistics in Clinical Trials, and Applications of Bayesian Methods to Drug and Medical Device Development. On Thursday and Friday, invited presentations will cover a variety of topics, possibly including comprehensive decision modeling; using predictive probabilities in clinical studies and drug development; roles for hierarchical modeling; how Bayesian methods can be used to augment traditional methods; Bayesian methods in epidemiology; the Bayesian approach and medical ethics; how to assure good quality and scientific rigor in taking a Bayesian approach; and guidelines for publishing Bayesian analyses. Registration fees will be modest. Program co-chairs: Donald A. Berry, Ph.D., The University of Texas M. D. Anderson Cancer Center, and Telba Z. Irony, Ph.D., Center for Devices and Radiological Health, U.S. Food and Drug Administration.

Information will be available at http://biostatistics.mdanderson.org/BBC2010/

Frontier of Statistical Decision Making and Bayesian Analysis, San Antonio, Texas. 17-20th March, 2010.

This conference consists of plenary, invited and poster sessions. Plenary speakers include Donald Berry, Lawrence Brown, Persi Diaconis, Stephen Fienberg, and Alan Gelfand. The conference will provide an overview of past, present and future developments of statistical decision making and Bayesian analysis. Prior to the conference, short courses on various statistical topics will be offered.

For more information visit the website, http://bergerconference2010.utsa.edu/.