Joint simulation of regional areas burned in Canadian forest fires: A Markov Chain Monte Carlo approach

Steen Magnussen[1]

***Abstract**: Areas burned annually in 29 Canadian forest fire regions show a patchy and irregular correlation structure that significantly influences the distribution of annual totals for Canada and for groups of regions. A binary Monte Carlo Markov Chain (MCMC) is constructed for the purpose of joint simulation of regional areas burned in forest fires. For each year the MCMC prediction is a binary vector with regions classified to a large fire year (LF) or a small fire year (SF). The regional area burned is then obtained from empirical quantile functions; separately for LF and SF years. The MCMC results were unbiased with respect to: the annual number of LF regions, national totals, and variances of area burned. Approximately 65% of the observed regional covariance was captured in the results.*

Keywords: binary correlation, multivariate simulation, marginal distribution, transition kernel.

Introduction

Forest fires affect forest resources and the global cycling of carbon and greenhouse gasses (Amiro et al. 2001, Bergeron et al. 2004, Gillett et al. 2004). They are a dominant driver in Canada's boreal forest carbon balance (Bond-Lamberty et al. 2007). Forecasting areas burned annually in forest fires (*BA*) at a regional and a combined regional scale is therefore important to predicting future greenhouse gas emissions (Kurz and Apps 2006, Kurz et al. 2008).

In Canada *BA* varies dramatically between years and regions. Large fires tend to occur during periods of stable high pressure (Skinner et al. 1999, Skinner et al. 2008). These atmospheric patterns are sub-continental in scale and may impose

[1] Canadian Forest Service, Natural Resources Canada, 506 West Burnside Road, Victoria, British Columbia V8Z 1M5, Canada. Phone: +1 250 363 0712, Fax: +1 250 363 0775, Email: steen.magnussen@nrcan.gc.ca

In: McWilliams, Will; Moisen, Gretchen; Czaplewski, Ray, comps. 2009. 2008 Forest Inventory and Analysis (FIA) Symposium; October 21-23, 2008: Park City, UT. Proc. RMRS-P-56CD. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 1 CD.

some regional synchronization in *BA*. Yet snow-cover or rain during the winter prevents the emergence of a strong temporal autocorrelation by saturating the forest fire fuels.

The simplest approach to forecasting *BA* is by recasting historic records. Recasting is attractive on grounds of expediency, simplicity, low costs, and transparency; however, this approach must take into account any interregional correlation structure. In Canada regional *BA* from 1955 to 1999 show an irregular pattern of weak and strong interregional correlations that exert a significant effect on the variance, and thus the shape, of distributions of sums of regional *BA*-values.

This study demonstrates a Markov Chain Monte Carlo (MCMC) procedure for joint forecasting of *BA* in 29 Canadian forest fire regions. The 29 regions account for about two-thirds of the areas burned in Canada. The rationale for the MCMC procedure rests with the fact that interregional correlations of *BA* are - to a large degree - shaped by a few years favorable to large fires.

Material and Methods

Data

Estimates of annual areas burned in forest fires (*BA*) from 1959 to 1999 in Canada’s 29 forest fire regions were used as data for forecasting purposes (Magnussen 2008, Kurz and Apps 2006, Stocks et al. 2002).

Forecasting objectives

The objective is to forecast a sequence of *BA*-values for each of the 29 fire regions consistent with historic data from 1959-1999. Forecasted data should also conserve the regional correlations pattern so that the distribution of sums of regional *BA*-values matches the historic distribution of these sums.

Model premise

Regional correlations reflect the number of concurrent large values of *BA*. Consequently, a binary classification of region years to a large fire year (*LF* = 1) or a small fire year (*SF* = 0) is used to capture the regional correlations. To simplify the correlation structure it was decided to: *i*) form 29 balanced five-member ρ-cliques by maximizing the average within-clique correlation, and *ii*) assume that the *LF*-status of a region is only influenced by regions in the same ρ-clique. Accordingly, a two-stage process for the joint forecasting of *BA* is formulated: In stage one, the total number and regional allocation of *LF* years is determined in a MCMC step (Robert and Casella 1999). In stage two, regional *BA*-values are drawn from empirical quantile functions, separately for *LF* and *SF* years.

Classification of region-years to LF or SF

The classification of region-years to either *LF* (1) or *SF* (0) was done by a *k*-means clustering (*k* = 2) routine. Following the classification, the probability (θ_i) that a *LF* year occurs in region *i* was estimated as:

$$\hat{\theta}_i = \frac{\sum_{j=1}^{41} LF_{ij}}{\sum_{i=1}^{29}\sum_{j=1}^{41} LF_{ij}} \qquad [1]$$

The MCMC (stage I)

Every forecast begins with a random draw of the number nLF^* of regions with a *LF* = 1 status. The draw is from a zero-truncated beta-binomial (Griffiths 1973) of nLF fitted to the classified data. An initial random allocation of the nLF^* is done with probability proportional to the regional probability of a *LF* year $\left\{\hat{\theta}_1,...,\hat{\theta}_{29}\right\}$ and modified in a sequence of switches $\left(s = 0,1,2,...\right)$ of *LF* status between two regions with opposite *LF*-status. The sequence maximizes the conditional likelihood of the allocation. Let $\mathbf{LF_0}$ denote the binary vector of the initial random allocation of regional *LF*-years. A switch involves two regions (*i* and *j*) for which $LF_i^* = 1$ and $LF_j^* = 0$, and $LF_i^* \to 0$ and $LF_j^* \to 1$. Let $\mathbf{LF}_s^*$ denote the vector of regional *LF* status after *s* switches, and let $\mathbf{LF}_{new}^*$ denote a proposed new configuration obtained from $\mathbf{LF}_s^*$ and applying a proposed switch. $\mathbf{LF}_{new}^*$ is accepted with a probability $\alpha_{s,s+1}$ $\left(\mathbf{LF}_s^* = \mathbf{LF}_{new}^*\right)$ or rejected with probability $1-\alpha_{s,s+1}$ in which case $\mathbf{LF}_{s+1}^* = \mathbf{LF}_s^*$. We have

$$\alpha_{s,s+1} = \text{Min}\left[1, K\left(\mathbf{LF}_{new}^*, \mathbf{LF}_s^*\right)\right], U_{s+1}^* \sim \text{Uniform}[0,1]$$
$$\mathbf{LF}_{s+1}^* = \begin{cases} \mathbf{LF}_{new}^* \text{ if } U_{s+1}^* \le \alpha_{s,s+1} \\ \mathbf{LF}_s^* \text{ if } U_{s+1}^* > \alpha_{s,s+1} \end{cases} \qquad [2]$$

where U_{s+1}^* is a random draw from a uniform distribution on the unit interval [0,1] and $K\left(\cdot,\cdot\right)$ is defined in Equation 3

$$K\left(\mathbf{LF}^*_{new},\mathbf{LF}^*_s\right)=\frac{\ell_p\left(\mathbf{LF}^*_{new}\right)}{\ell_p\left(\mathbf{LF}^*_s\right)}\times\frac{\ell\left(LF^*_{i,s}=1\,,\,LF^*_{j,s}=0\right)}{\ell\left(LF^*_{i,new}=0\,,\,LF^*_{j,new}=1\right)} \qquad [3]$$

with

$$\ell_p\left(\mathbf{LF}^*_{new}\right)\propto\ell\left(LF^*_{i,new}\mid LF^*_{i'\sim i,new}\right)\ell\left(\sum_{i'\sim i}LF^*_{i',new}\right)\ell\left(LF^*_{j,new}\mid LF^*_{j'\sim j,new}\right)\ell\left(\sum_{j'\sim j}LF^*_{j',new}\right)$$

$$\ell_p\left(\mathbf{LF}^*_s\right)\propto\ell\left(LF^*_{i,s}\mid LF^*_{i'\sim i,s}\right)\ell\left(\sum_{i'\sim i}LF^*_{i',s}\right)\ell\left(LF^*_{j,s}\mid LF^*_{j'\sim j,s}\right)\ell\left(\sum_{j'\sim j}LF^*_{j',s}\right)$$

[4]

where ℓ denotes a likelihood and ℓ_p a pseudolikelihood and $i'\sim i$ denotes regions $\left(i'\right)$ in the same *ρ*-clique as region *i*. Likelihoods $\ell\left(LF^*_{i,new}\mid LF^*_{i'\sim i,new}\right)$ were estimated from maximum likelihood estimates (MLE) of clique-specific autologistic functions. Conversely, $\ell\left(\sum_{i'\sim i}LF^*_{i',new}\right)$ were estimated from MLE of a clique-specific probability mass function (binomial, zero-inflated-binomial, or beta-binomial). Finally $\ell\left(LF^*_{i,s}=\delta\,,\,LF^*_{j,s}=1-\delta\right),\ \delta=\{1,0\}$ was estimated from the classified data as outlined in Congdon (2006, p 395).

After approximately 1000 accepted switches the Markov Chain reached a steady state so that the current value of the vector $\mathbf{LF}$ could be viewed as sampled from the joint distribution of regional *LF*-status (Robert and Casella 1999). As a safeguard, the vector $\mathbf{LF}^*_S$ after 3000 accepted switches was retained. A total of 41×100 random replicates of $\mathbf{LF}^*_S$ were generated, representing 100 replications of 41-year forecasts. Without a temporal autocorrelation, years and replicates are interchangable.

Forecasting BA (Stage II)

For a given *LF* forecast for region *i* the associated *BA* was determined as

$$BA^*_i=\begin{cases}\hat{F}_i^{-1}\left(u^*\mid LF^*_{i,S}=0\right),\ u^*\in\left[0,\hat{u}_{SFi}\right)\\ \hat{F}_i^{-1}\left(u^*\mid LF^*_{i,S}=1\right),u^*\in\left[\hat{u}_{SFi},1\right]\end{cases} \qquad [5]$$

where $\hat{F}_i^{-1}$ is the empirical quantile function of *BA* for region *i* (i.e. the inverse to the empirical distribution function), u^* is a random draw from a uniform distribution from a specified interval, and $\hat{u}_{SFi}$ is the MLE estimate of the regional cut-off quantile between *LF* = 1 and *SF* = 0 on the empirical quantile function. A simulation study (Magnussen 2008) suggested the following endpoints for the quantile functions: $\hat{F}_i^{-1}(0) = 0.54 \times \mathrm{Min}(BA_i)$, and $\hat{F}_i^{-1}(1) = 1.35 \times \mathrm{M}\,ax(BA_i)$.

Results

Regional correlations of *BA* averaged 0.06 but were statistically significant (5% level) at a rate of 0.14. Only five regions (NB1, NF1, PQ1, PQ2, SK4) were not significantly correlated with at least one other region. A bootstrap simulation study confirmed that significant correlations were almost always due to a few concurrent years of large *BA*.

Examples of the classification of *BA* to *LF* (1) and *SF* (0) by the *k*-means procedure are in Figure 1. The average regional relative frequency of *LF* years was 0.19 but varied from a low of 0.05 in PQ4 to a high of 0.39 in SK2. The average cut-off point for classifying a *BA*%-value to *LF* was 0.7% (range: 0.04% in AB4 to 3.3% in AB3). Regional correlations of *BA* in shared *SF* years were, on average, about 84% below the correlations for the entire period of 41 years and the rate of significant correlations was consistent with the null hypothesis of a zero correlation.

In the MCMC forecasts the average rate of *LF* years was 2% below the rate in the classified data (*P* = 0.12, bootstrap *t*-test). The number of regional *LF* years forecasted for a 41-year period was positively correlated (0.77) with the observed number (Figure 2). There is a tendency to overestimate nLF_i in regions with lower frequencies of *LF* years and to underestimate in regions with higher rates. A bootstrap *t*-test with 36 degrees of freedom identified three regions (BC2, NS1, NWT1) with a significant difference $(0.018 \leq P \leq 0.026)$ between the classified and the forecasted number of *LF*-events during a period of 41 years.

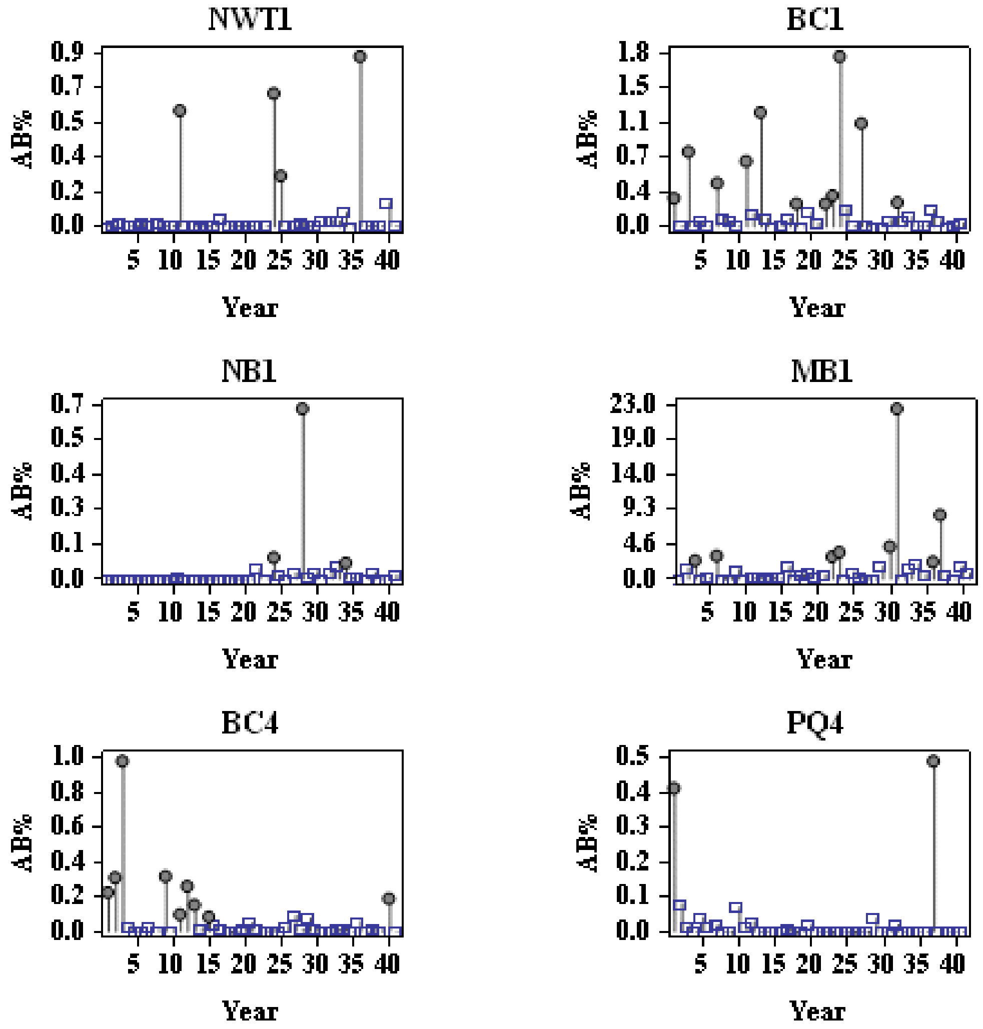

Figure 1: Percent of forested area burned annually (*BA%*) in six randomly chosen regions. Year 1 = 1959, year 41 = 1999. The classification of *BA%* to *LF* (large fire) or *SF* (small fire) is indicated by squares (*SF*) and circles (*LF*).

The bias pattern in Figure 2 carries over to regional correlations of *LF* years in the MCMC results (Figure 3) and created an inflation in cliques with a below average interregional correlation and vice versa. Across all regions, the average correlation of *LF* years was 0.04 in the forecasts and 0.05 in the data, and the relationship between the two sets of correlation coefficients was consistent with a slope of 1.0 and a zero intercept (P = 0.16).

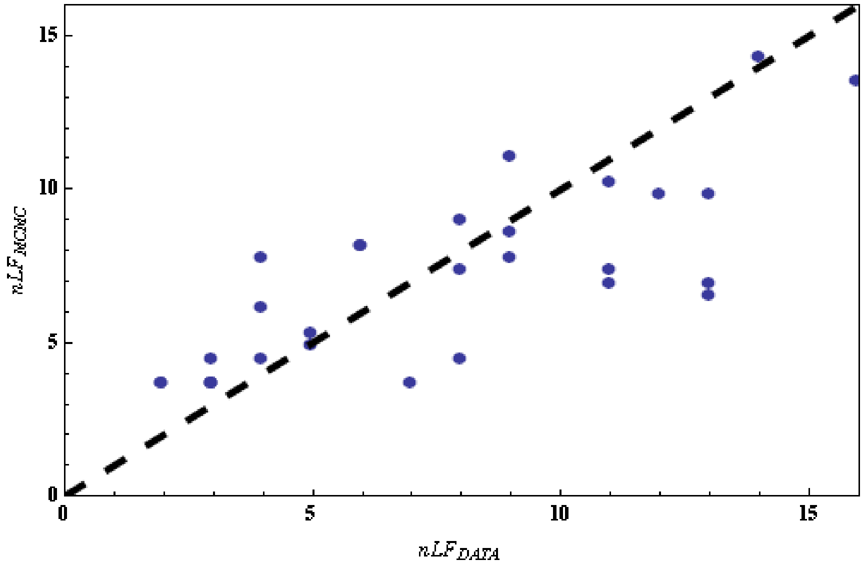

Figure 2: Forecasted total number of regional *LF* years during a period of 41 years (*nLF_{MCMC}*), plotted against the number in the classified data (*nLF_{DATA}*).

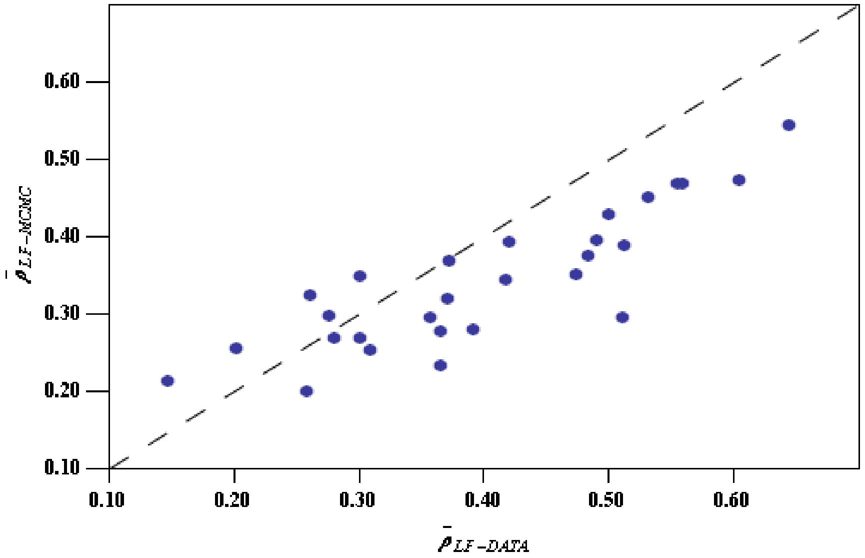

Figure 3: Forecasted average inter-regional correlation of *LF* years $\left(\bar{\rho}_{LF-MCMC}\right)$ plotted against the average correlation in the data $\left(\bar{\rho}_{LF-DATA}\right)$. The average is over regions in a ρ -clique.

The mean and variance of forecasted regional *BA* matched fairly closely their historic values. Scatter plots in Figures 4 and 5 convey a strong correlation (0.98) between forecasted and historic values. A linear relationship with a slope of 1.08 $\left(\pm 0.02\right)$ and an intercept not significantly different from zero (*P* = 0.31) captures the relationship. For all regions combined the average *BA* in the forecasts was 886 976 ha versus 817 308 ha in the data. The bias is attributed to the asymmetric capping of the empirical quantile functions. The standard deviation of the regional totals of *BA* was 849 596 ha but only 742 973 ha (-14%) in the MCMC forecasts.

The conversion to a binary variable (*LF*) and the ensuing attenuation of the regional correlations is the main factor behind the bias. Regarding regions as independent would generate a standard deviation of 486 990 ha (-43%). In other words, the MCMC procedure captured 65% of the regional covariance of *BA*.

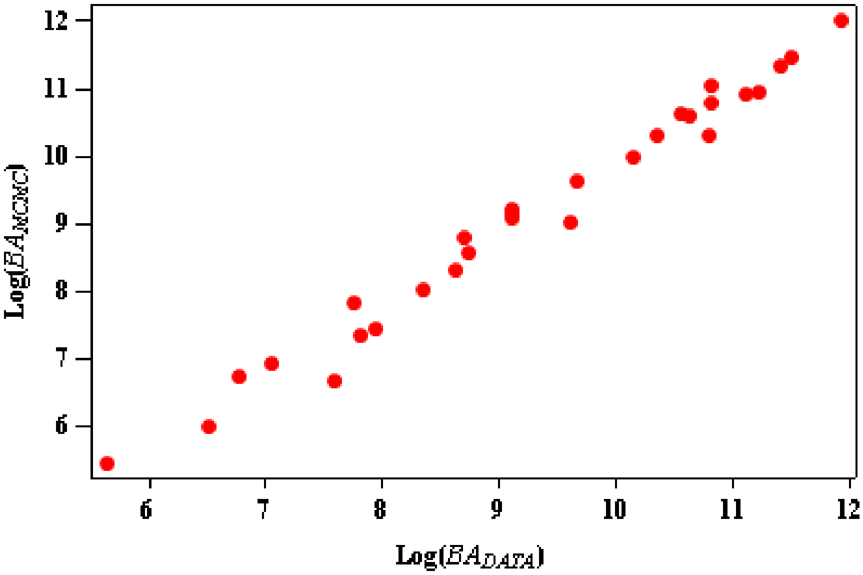

Figure 4: Forecasts of average annual regional *BA*-values $\left(BA_{MCMC}\right)$ plotted on a logarithmic scale against historic values $\left(BA_{DATA}\right)$.

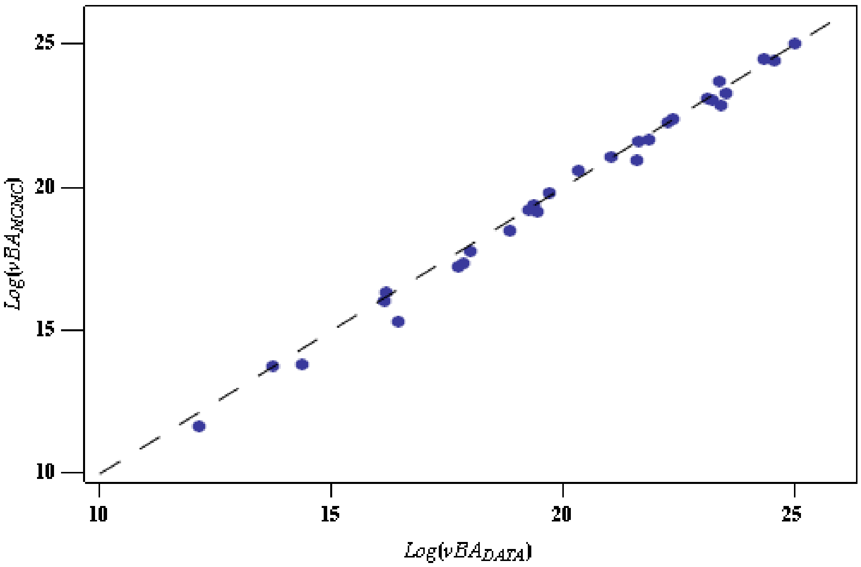

Figure 5: Forecasts of variance of annual regional *BA*-values $\left(vBA_{MCMC}\right)$ plotted on a logarithmic scale against historic values $\left(vBA_{DATA}\right)$.

Discussion and conclusions

A joint forecast of regional *BA* must take the apparent correlation structure into account or the variation of sums of regional values will be biased downward. Without a suitable multivariate distribution function, the task of a joint forecast becomes a complex challenge (Aalo and Piboongungon 2005, Carpenter and Diawara 2007).

A binary classification of *BA* as either large or small facilitates an interpretation of the regional correlation structure as it changes the focus from areas to years. Common environmental factors in region-years classified as *LF* may be identified. Modeling at the binary level also facilitates an integration of expected trends in *LF* years (Bergeron et al. 2004, Beverly and Martell 2003, Larsen 2007).

Modeling a regional distribution of correlated binary variables is commonly done by formulating the probability of an event in a region conditional on the number of events in some defined neighbourhood (ρ-cliques) composed of interdependent regions, usually a group of first-order spatial neighbours (Gilliland and Schabenberger 2001, Sherman et al. 2006).

The proposed MCMC procedure was simplified by conditioning on the marginal distribution of the total number of *LF* events in a year. Without this simplifying step, the transition kernel would have been considerably more complicated (Smith and Smith 2006). We surmise that our MCMC results reflect the constraints on the covariance structure inherent in all multivariate distribution functions (Johnson 1987). A restriction of first-order regional interactions to ρ-cliques limited our ability to capture the observed interregional correlation structure. The proposed MCMC approach is capable of reproducing the main features of observed marginal distributions and an irregular and patchy correlation structure.

References

Aalo, V.A.; Piboongungon, T. 2005. On the multivariate generalized gamma distribution with exponential correlation. In: Global telecommunications conference IEEE 3(28): 1229-1233.

Amiro, B.D.; Todd, J.B.; Wotton, B.M.; Logan, K.A.; Flannigan, M.D.; Stocks, B.J.; Mason, J.A.; Martell, D.L.; Hirsch, K.G. 2001. Direct carbon emissions from Canadian forest fires, 1959-1999. Canadian Journal of Forest Research 31: 512-525.

Bergeron, Y.; Flannigan, M.; Gauthier, S.; Leduc, A.; Lefort, P. 2004. Past, current and future fire frequency in the Canadian boreal forest: implications for sustainable forest management. Ambio 33:356-360.

Beverly, J.L.; Martell, D.L. 2003. Characterizing historical variability in boreal forest fire frequency: Implications for fire and forest management. In: Arthaudf, G.J.; Barrett, T.M. (eds). Proceedings Systems Analysis in Forest Resources. Volume 7: Managing Forest Resources. p 3-14. Kluwer, Dordrecht NL. 326 p.

Bond-Lamberty, B.; Peckham, S.; Ahl, D.; Gower, S. 2007. Fire as the dominant driver of central Canadian boreal forest carbon balance. Nature 450: 89-92.

Carpenter, M.; Diawara, N. 2007. A multivariate gamma distribution and its characterization. Technical Report. Dep Math & Statistics. Auburn Univ.

Congdon, P. 2006. Bayesian statistical modelling. Wiley, Chichester, England. 571 p.

Gillett, N.P.; Weaver, A.J.; Zwiers, F.W.; Flannigan, M.D. 2004. Detecting the effect of climate change on Canadian forest fires. Geophysical Research Letters 31: L18211.

Gilliland, D.: Schabenberger, O. 2001. Limits on pair-wise associations for equi-correlated binary variables. Journal of Statistical Science 4: 279-285.

Griffiths, D.A. 1973. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. Biometrics 29: 637-648.

Johnson, M.E. 1987. Multivariate statistical simulation. Wiley, New York. 230 p.

Kurz, W.A.; Apps, M.J. 2006. Developing Canada's national forest carbon monitoring, accounting and reporting system to meet the reporting requirements of the Kyoto protocol. Mitigation and Adaptation Strategies in Global Change 11: 33-43.

Kurz, W.; Stinson, G.; Rampley, G.; Dymond, C.; Neilson, E. 2008. Risk of natural disturbances makes future contribution of Canada's forests to the global carbon cycle highly uncertain. Proceedings National Academy of Science USA 105: 1551-1555.

Larsen, C.P.S. 2007. Fire and climate dynamics in the boreal forest of northern Alberta, Canada, from AD 1850 to 1989. The Holocene 6: 449-456.

Magnussen, S. 2008. Joint Regional Simulation of Annual Area Burned in Canadian Forest Fires. The Open Forest Science Journal 17: 37-53.

Robert, C.P.; Casella, G. 1999. Monte Carlo statistical methods. Springer, New York. 507 p.

Sherman, M.; Apanasovich, T.V.; Carroll, R.J. 2006. On estimation in binary autologistic spatial models. Journal of Statistical Computation and Simulation 2: 167-179.

Skinner, W.; Skinner, M.; Flannigan, M.; Stocks, B.; Martell, D.; Wotton, B.; Todd, J., Mason; J.A., Logan, K.; Bosch E.M. 2008. A 500 hPA synoptic wildland climatology for large Canadian forest fires, 1959-1996. Theoretical and Applied Climatology 71: 157-168.

Skinner, W.; Stocks, B.; Martell, D.; Bonsal, B.; Shabbar, A. 1999. The association between circulation anomalies in the mid-troposphere and area burned by wildland fire in Canada. Theoretical and Applied Climatology 63: 89-105.

Smith, D., and Smith, M. 2006. Estimation of binary Markov random fields using Markov chain Monte Carlo. Journal of Computational and Graphical Statistics 15: 207-227.

Stocks, B.J.: Majson, J.A.; Todd, J.B.; Bosch, E.M.; Wotton, B.M.; Amiro, B.D.; Flannigan, M.D.; Hirsch, K.G.; Logan, K.A.; Martell, D.L.; Skinner, W.R. 2002. Large forest fires in Canada, 1959-1997. Journal of Geophysical Research 108: 1-12.