Applicability of Spatial Duration Model for Forest fire Duration Toward Haze Alart System

Written by Makoto TSUKAI, M.N.B.JAAFAR, and Kiyoshi KOBAYASHI
Translated by Daisuke HARADA

1. Introduction

The forest fires in Kalimantan and Sumatra, Indonesia has caused large effects to ASEAN nations in dry season when the southwest monsoon blows harder. That monsoon including some pollutants is called Haze[1)], and ASEAN nations receive some damages from Haze in each year such as the close of the airports, respiratory diseases and so on. Haze disaster especially in 1997 is one of the most terrible disasters, and the total loss by Haze amounted to 9million USD in all damaged nations[2)].

Malaysia near Sumatra Island is one of the most suffering nations from Haze. Western area in Malaysia, which has some metropolis and some active Industries, is particularly easy to be affected. On Haze disaster in 2005, schools, Industries, and offices temporally had to close[3)]. To prevent from that disaster, government in Malaysia has made efforts to build a system which can give us the alarm by estimating the density of air pollution in real time. Now private company, Alam Sekitar Malaysia (ASMA)[4)] has managed the monitoring network (Continuous Air quality monitoring, called CA station), set at 50 places in Malaysia, which can continuously measure the air pollutant such as CO, CO_2, NOx, SOx. Based on ASMA's data, Malaysian government in department of environment established API (Air Pollution Index)[5)] standard and judge whether air pollutants exceed API standard or not. But now they don't have a system which enables them to forecast air pollution density. To provide alarm information, they need to estimate future data on Haze, and statistic models are effective to forecast.

On this motivation, Kobayashi has already built a statistic model which can forecast the air pollutant density, and analyzed correlation between space and time on Haze. His research shows the characteristics of Haze, long memory, seasonality, and space-time correlation. Furthermore he developed his model by adopting regime switching in addition to long memory. Regime switching makes us possible to forecast more accuracy so that we can adjust the sharp change of the density of Haze and to overcome the occurrence of error when Haze depends on past data in long term. Kobayashi shows the behavior of Haze by setting regimes in each station

in Malaysia. However, regime switching in his model required us to forecast the number of hotspots in Sumatra Island, because time series data as to hotspots in Sumatra Island decided the regime of Haze. In Kobayashi's research he considered the number of hotspots as given values and inserted them exogenously into the model, but exogenous values bring some error to each regime and bring a limitation to forecasting. So we can't set up the accurate regime on Haze, unless we can predict the number of hotspots.

The purpose of this paper is to forecast the number of hotspots statistically, and we try to compensate the lack of Kobayashi's previous theory. As a concrete explanation, we build a statistical duration model to enable us to estimate the number of hotspots in Sumatra Island by using the duration model based on hazard function on space-time data given by satellite. Finally we apply this duration model to the observed data in Sumatra Island, and then we reveal the effectiveness of the model.

Outline of this paper is as follows. In Chapter 2, we explain the fundamentals of this research. Chapter 3 shows the duration model. Chapter 4 is result of applying this duration model to observed data of hotspots.

2. Fundamentals

Some researcher has proposed CA model (Cellular Automata model) is effective to forecast the process of landscapes on forest fire. CA model is a statistical simulation model which can take correlation of space-time into consideration. CA model can describe the landscape as a grid in which the dimension of each cell is 50m ($2500m^2$), and we can consider several processes which affect the landscape of fire spreading, including diffusive spread from cell to cell. With the use of CA model we can find the change of landscapes during time series, besides can see when forest fire happens and disappears. However, CA model demands us much information about landscapes. It is certain that the more the number of explanatory variables are, the more accurate we can forecast. But the objective of this paper is to cover the theoretical problem in Kobayashi's paper. He used only the number of hotspots in Sumatra Island as a explanatory variable. Besides we want to build a model as simple as possible because of computational problems.

As an approach to forecast easily, we try to adjust the duration model. The duration model can show a spell of landscapes, and specify the turning points in the course of spreading of forest fire. Furthermore duration model makes computational tasks more fewer than CA model, and lead us to express more flexibly.

In this research, we regard a spell in forest fire as a random variable, and then we

build a statistical duration model which can update the spread of fire with use of the number of past hotspots data.

3. Duration model

3.1 Geo additive hazard function

Geo additive hazard function is a kind of duration model which can express the correlation of space-time. First we define a hazard function before setting up a geo additive function.

Let t, a non-negative random variable, be a spell length for forest fire in the absence of censoring, and let T be the censoring time measured from time origin for the spell. Then the distribution function and the density function following t are given as follows.

$$F'(t) = f(t)$$

Moreover, we suppose a probability distribution S(t) that means forest fire has not yet occurred during $T + t \geq T$. In that sense, probability distribution S(t), called survivor function, is the probability that the random variable T will equal or exceed the value t. A particular useful function for duration analysis is the hazard function h(t), and survivor function S(t) is related so closely to hazard function h(t) giving the rate at which spell will be complicated at duration t, given that they last until t. Then we can define hazard function and see the relation below

$$\begin{aligned} h(t) &= \lim_{h \to 0} \frac{\Pr\{t \leq T < t + h\}}{h} \\ &= -\frac{d\ln S(t)}{dt} \\ &= \frac{f(t)}{S(t)} \end{aligned} \qquad (1)$$

To express the characteristics of each spot, we will adopt the proportional hazard function to h(t).

$$\begin{aligned} h_i(t) &= h_0(t)\exp(\eta_i) \\ &= \exp(\log h_0(t) + \eta_i) \end{aligned} \qquad (2)$$

Where index i means the number in each spot and $h_0(t)$ is a baseline hazard independent to the factor of spots. In formula (1), we can estimate the log-likelihood function as follows,

$$L = \sum^{i} \{\delta_i \log h_i(t) - \int_0^t h_i(u)du\} \qquad (3)$$

Where δ_i is a dummy variable, when right censoring in i-th hotspot occurs, δ_i will

get 1, otherwise δ_i will be 0.

Kneib and Fahrmeir proposed a Geo additive function, which can adjust to the relationships hold everywhere within spots. The characteristic of Geo additive function is 1) to express the correlation in spatial dimensions, 2) to adjust to time-varying effects, 3) to estimate baseline function simultaneously with covariate effects.

$$h_i(t) = \exp\left\{g_0(t) + \sum^{j} \gamma_j u_{ij} + \sum^{k} g_k(t) u_{ik} + \sum^{l} f_l(u_{il}) + f_i(s)\right\} \qquad (4)$$

Here, $g_0(t)=\log h_0(t)$ means a baseline hazard function with liner, and u_{ij}, u_{ik}, u_{il} are fixed covariates on i-th spot whose effect are represented by time-independent parameter γ_j and time-dependent parameter $g_k(t)$. Where $f_l(u_{il})$, which is normally applied to three dimensional splines function, is a non-liner effect of covariate u_{il}, and $f_i(s)$ shows the spatial effect in i-th spot.

Furthermore Kneib and Fahrmeir suggest the new way of parameter estimating to add penalized differences between parameters adjacent basis function. For the spatially correlated and structured effect, we choose Markov random field common in spatial statistics of the form

$$f_i(s) = \frac{1}{N_i} \sum^{s \in D_i} \beta_s + u_s \qquad u_s \sim N\left(0, \frac{\delta_s^2}{N_i}\right) \qquad (5)$$

Where N_i is the number of adjacent area s with spatial effect and $s \in D_i$ denotes that area s is a neighbor of area D_i. Therefore $f_i(s)$ is an average of evaluations of spatial effect for neighboring areas. u_s is a random variable and δ_s^2 is a variance of u_s. The N×S design matrix $\mathbf{L_i}$ giving us the adjacent relation between every spot and area s is now a 0/1 incidence matrix. Its value in the i-th row and s-th column is 1 if observation i is located in area s, and zero otherwise. Now N means the number of spots, while S means the number of areas with spatial effect.

The S×S penalty matrix $\mathbf{K_s}$ has the form of an adjacency matrix.

In a penalized log-likelihood setting, the difference penalty can be expressed as

$$P_s = -\frac{\boldsymbol{\beta}' \mathbf{K_s} \boldsymbol{\beta}}{2\delta_s^2} \qquad (6)$$

Where the penalty matrix is of the form $\mathbf{K_s} = \mathbf{L_i}'\mathbf{L_i}$.

Then we can get a penalized log-likelihood function as follows

$$L_p = \sum^{i} \left(\delta_i \log h_i(t) - \int_0^t h_i(t)dt \right) - P_s \qquad (7)$$

Here, to maximize the formula (7) means to maximize the posterior likelihood on the perspective of Bayes estimation and we can estimate structured parameters with spatial effect simultaneously.

But some problem happens. We can't obtain the marginal distribution to estimate the variance δ_s^2, unless the penalty matrix $\mathbf{K_s}$ is regular.

As to this problem, Kneib and Fahrmeir proposed a procedure to estimate variances in a structured hazard regression model. That is, first to apply a Laplace approximation to the marginal log-likelihood (See Kneib and Fahrmeir[10]), second to set up initial variance δ_m^2 in the following likelihood function, then we get the parameters in hazard function, and finally to keep updating variance until parameters in hazard function become converge. That marginal log-likelihood function is represented as

$$L(\delta_s^2) = -\frac{1}{2}\log|\delta_m^2| - \log|\mathbf{H}| - P_s \qquad (8)$$

Where $\mathbf{H}$ is a Fisher-information-matrix and then we can obtain the variance δ_s^2 by maximizing the likelihood.

3.2 Application geo additive to forest fire

Geo additive hazard function is an ideal match for forest fire model. That is why it can express the landscape of widespread forest fire accuracy with considering the characteristics of each spot and interaction between spots. To illustrate the usefulness and flexibility of geo additive hazard function, we apply geo additive hazard function to forest fire in Sumatra.

$$h_i^o(t) = \exp\{\gamma_0 + \sum^{k} \rho_k u_{i,t-k} + f_s(s_i)\} \qquad (9)$$

Let γ_0 and ρ_k be fixed parameters and $f_s(s_i)$ is a spatial effect in i-th spot. Then we include a time-varying covariate $u_{i,t-k}$ in hazard function. Time-varying components enable us to update covariates by each time, and bring the forecast ability improved.

This model is not only used to judge the duration of forest fire(called **Death model**) but also used to determine the time of occurrence of hotspot(called **Birth model**). But we need some attention here. On this progressive process, Samples used to the

Birth model are different from those of Death model, because the start time of the spell is different. For example in case of Birth model we use samples at T_i , but in other case we can use samples at $T_i + t_i$ after the duration of forestfire finished. Besides we can consider the rebirth of fire in hotspots. Hotspots are defined by temperature on the ground. Therefore even if the temperature at each spot become lower than the standard level once, some spots are defined as hotspot again by the effect of interaction in spatial dimension. Now we modify the log-likelihood function reflected the rebirth of fire in hotspots.

$$L_p = \sum^{i}\sum^{q}\left(\delta_i^q \log h_i^q(t) - \int_0^t h_i^q(t)\right) - P_s \qquad (10)$$

Let δ_i^q is a dummy variable in addition to the effect of the occurrence of fire at q times. When we estimate the parameters in formula (10), we can use the data observed at time T_i^q which shows the start time after (q-1)th spell of the fire.

3.3 Estimation method for the number of hotspots

This section shows you the procedure of estimation of the number of hotspots. The estimation method is followed by the survivor function and input data are based on observed data up to time t. Here, we employ a 3 steps method The procedure is as follows

a) Estimation of the number of pre break-out spots

The number of pre break-out spots at future time $t + \tau$, $\Delta x_t^o(\tau)$ is represented as

$$\Delta x_t^o(\tau) = \sum_{i=1}^{N_t^o}\{S_i^o(t) - S_i^o(t+\tau)\}$$

Where N_t^o is the number of pre break-out spots at time t, and $S_i^o(t)$ shows the survivor function for Birth model at time t.

b) Estimation of the number of still fired spots

The number of fired spots at future time $t + \tau$, $\Delta x_t^d(\tau)$ is represented as

$$\Delta x_t^d(\tau) = \sum_{i=1}^{N_t^d}\{S_i^d(t) - S_i^d(t+\tau)\}$$

Where N_t^d is the number of spots still fired at time t, and $S_i^d(t)$ shows the survivor function for Death model at time t.

c) Estimation of the number of hotspots

We can obtain the number of hotspots at future time $t + \tau$ by using formula a),b).

The number of hotspots $x_{t+\tau}$ can be expressed as

$$x_{t+\tau} = x_t + (\Delta x_t^o(\tau) - \Delta x_t^d(\tau))$$

Here, x_t is the number of hotspots at time t.

4. Application to the data in Sumatra

4.1 Data base

NOAA circulates the earth every one hundred minutes at an altitude of 850km. Data from NOAA can be received everyday at specific time. NOAA is equipped with a sensor called AVHRR (Advance Very High Resolution Radiometer). AVHRR detects the temperature at ground level by using mainly near infrared rays. Hot Spot is the terminology for a pixel, which has a higher temperature than the particular threshold captured by satellite digital data. The size of a pixel is 1.1 km times, 1.1 km and the threshold values applied for the infrared channel are 315K (42oC) for day capturing and 310K (37oC) for night capturing When cloud covers the land, Hot spots cannot be detected. Satellite picture shows us the distribution of the temperature on the ground. On the basis of satellite picture, hotspots are counted.

The data used here is published on web page in JICA, and we can obtain observed data in Sumatra Island, Malay, and Borneo Peninsula. But our research is focused on the East area in Sumatra, such as Riau state, Jambi state, and Lamnpung state because hotspots there mainly have caused the occurrence of Haze. Data base contains 22,002 times occurrence of hotspots in total at those 3 spots during July 26th to October 19th.

4.2 Result of estimation

The modeling framework is already explained at chapter 3, and we need to estimate mainly two kinds of parameters 1) for hazard regression, 2) for spatial effects. First, the Markov random field must be set up in order to estimate the parameters for spatial effects, then we can estimate all parameters simultaneously by using Maximum Likelihood Estimation. As mentioned in Introduction, the purpose of this paper is to build a predictable model including the spatially correlated effects in hotspots. This is a common way in spatial statistic to introduce the spatially correlated effects by assuming neighboring sites.

Now, we set up the field divided a target district into 100 meshes, and area s given a spatial effect to neighbors is determined following the number of occurrence of hotspots, in that sense, we distribute it by picking up 10 meshes in order as to the number of occurrence 1st, 11th, 21th,…,91th. Beside neighboring sites D_i is defined as the area where mesh i is surrounded by 8 meshes.

Table-1, Table-2 show the result of estimation on parameters for Birth and Death model on hotspots. Here we adopt the Weibull function to baseline hazard g_0. Parameters in the table are all in hazard function except for parameters as to Weibull function and penalized variance, and the value of parameters provides us the strength of the relation to the event. For example, the parameter ρ_k in **Table-1** gets the plus value, this means that the longer the duration of the death in hotspots is, the more frequent hotspots occurred. But we can't say definitely geo additive hazard function fit the phenomenon in Sumatra, because t-statistic and the likelihood value are low in both cases. However, we can determine, as a result, this model is effective. We can see from the average likelihood Birth model is more effective to apply than Death model and from the table spatial effects at each spot vary widely. By the way the index number for each area in the table doesn't follow the frequency of hotspots, and make sure that the lower the value of parameter as to spatial effects is, the more frequently hotspots occur at each spot. That is why hotspots occurred in short term or in small scale are not counted because these hotspots never affect the event occurrence.

Let discuss the number of hotspots. **Figure-1, Figure-2** show the result of 1 and 2 days forecast number of hotspots in addition to the real number of hotspots. We can see from **Figure-1** the forecasted value at 1 day later can be enough applied to the real one, but the value at 2 days later in **Figure-2** is not fit well. Besides forecasted values at 1 day later are precedent to the real ones, while values at 2 days later fall behind the real one. That means this model is effective to forecast only 1-day-later-hotspots. However, the purpose of this model is to be able to apply to Kobayashi's paper, and his paper required us 12 hours forecasting. Therefore we can affirm this model is enough applicable and has a sufficient forecast ability.

5. Conclusion

In this research, we attempt to develop the forecasting model for hotspots in Sumatra. The approach proved to be useful in a real data example on the number of hotspots in Sumatra Island and showed satisfactory statistical properties in a simulation study. For accurate forecasting, the accumulation of data base is required and some extension of the proposed method might be desirable.

Table-1　Estimated result (Birth model)

	CSS estimate	t-statistic
q	0.0034	1.77
ρ_k	0.022	2.72
γ_0	−0.764	−9.64
β_1	0.084	0.98
β_2	−0.192	−0.67
β_3	0.105	0.65
β_4	0.054	0.63
β_5	0.035	0.29
β_6	0.086	0.27
β_7	0.057	0.41
β_8	−0.144	−1.01
β_9	0.001	0.01
β_{10}	−0.048	−0.34
α (Weibull function)	0.362	22.92
δ	1.866	2.77
Converge CSS	−2066.3	
Average CSS	−2.364	
sample number	874	

Table-2　Estimated result (Death model)

	CSS estimate	t-statistic
q	0.012	2.35
ρ_k	0.023	2.59
γ_0	−0.665	−8.35
β_1	−0.03	−0.23
β_2	0.34	1.02
β_3	−0.093	−0.52
β_4	−0.151	−1.7
β_5	0.003	0.05
β_6	−0.038	−0.15
β_7	−0.105	−0.76
β_8	0.137	0.75
β_9	−0.028	−0.23
β_{10}	0.062	0.39
α (Weibull function)	0.718	23.08
δ	1.212	0.76
Converge CSS	−1413.3	
Average CSS	−1.717	
sample number	823	

Reference

1) Langman, B. : A model study of smoke-haze influence on clouds and warm precipitation formation in Indonesia 1997/1998, *Atmospheric Environment*, (article in press), 2007.
2) The global fire monitoring centre : Forest fire situation in Malaysia, *International forest fire news country notes*, vol.26, pp66-74, 2001.
3) Nichol, J. : Smoke haze in SE Asia : A predictable recurrence. *Atmospheric Environment*, vol. 32, pp 2715-2716, 1998.
4) Abdul, R.K. : Observation of PM 10 Readings in Relation to Forest Fire Events from ASMA's Continuous Air Quality Monitoring Stations, ASMA, 2000.
5) Department of Environment Malaysia : *A Guide to Air Pollutant Index (API) in Malaysia,* 2000.
6) Kobayashi, K. M.N.B.JAAFAR, Ogata, S. and Tsukai, M. : Statistical Warning Model of Trance-Boundary Haze Disaster. *JSCE*. D, Vol. 63, pp. 478-497, 2007.
7) Hargrove, R. Guardner, R. Tuener, M. Romme, W. and Despain, D. : Simulating firepatterns in heterogeneous landscapes, *Ecological Modelling*, vol .135, pp.243-263, 2000.
8) Kneib, T. and Fahrmeir, L. : A mixed model approach for multicategorical space-time data : A mixed model approach, *Biometrical,* vol.69, pp.109-118, 2006
9) Kneib, T. and Fahrmeir, L. : A mixed model approach for geoadditive hazard regression, *Scandinavian Journal of Statistics*, vol.34, pp.207-228, 2007
10) Fahrmeir, L. and Kneib, T. : Penalized structured additive regression for space-time data : A Bayesian perspective, Statistical Scinica, vol.14, pp.731-761, 2004.
11) Fujiwara, A. Sugie, T. Chong. And Sigematsu, H. : The analysis of Duration model for Park and Ride, *JSCE*. Mo.14, pp.671-678, 1997.(Japanese only)
12) Sarvador, R. Loret, F. Pons, X. and Pinol, J. : Does fire occurrence modify the probability of being burned again? A null hypothesis test from Mediterranean ecosystem in NE Spain, *Ecological Modelling*, vol.188, pp.461-469, 2005.
13) JICA FFPMP2. : http://ffpmp2.hp.infoseek.co.jp/ewds.htm.