AN APPLICATION OF NOVEL ZERO-ONE INFLATED DISTRIBUTIONS WITH SPATIAL DEPENDENCE FOR THE DEFORESTATION MODELING

Ryuei Nishii

Faculty of Mathematics, Kyushu University
Motooka, Fukuoka, Japan

Shojiro Tanaka

Faculty of Science and Engineering, Shimane University
Nishikawatsu, Matsue, Japan

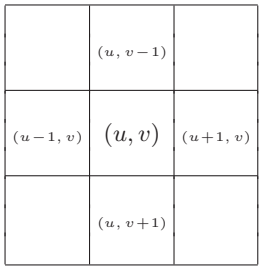
1. INTRODUCTION

Deforestation is caused by cultivated for cash crops, cattle grazing, shifting cultivation, logging, and fuel-wood use in developing countries, as well as airborne pollutants and acid rain in developed nations mainly [3]. Almost all of the causes of deforestation are related to human activities [2]. Forest fires are thought to be the biggest cause of deforestation in places such as Indonesia; such fires often result from the activities of nearby human populations, e.g., unextinguished embers of bonfires, cigarettes, etc. [1].

We incorporated the human factor of population into the forest coverage ratio by using grid-cell data. We assumed that the process of deforestation has a strong dependency on the human population in the area. Although human related factors such as land price and availability of transportation are potentially related to deforestation, we chose human population to be the first factor to be scrutinized. Forests tend remain on steep hillsides in populated areas since those areas are not suitable for cultivation or residence and are necessary to prevent landslides. We hence considered a second factor of slope steepness. Data on these two factors as well as on world-wide forest coverage are available in the form of grid cells (Fig. 1).

Let $N = N(s)$ be the population and $R = R(s)$ be the relief energy at site s. Here, the relief energy means the difference between the maximum and minimum altitudes (see Fig. 1). Also let $F = F(s)$ be the forest areal ratio including open forest ($0 \leq F \leq 1$). Our main concern here is to specify a regression model of F by explanatory variables: human population N and relief energy R.

Regression models: $F = g(N)+h(R)+error,\ \log\{F/(1-F)\} = g(N)+h(R)+error,\ \log\{(F+0.5)/(1-F+0.5)\} = g(N)+h(R)+error$ were examined [5], where $g(N)$ and $h(R)$ are non-linear regression functions and $error$ follows a spatially dependent/independent normal distributions. In our case, the forest ratio F is restricted to values from 0 to 1. The



- $s = (u, v)$
- $F = F(s)$: forest areal ratio $(0 \leq F \leq 1)$
- $N = N(s) \geq 0$: human population density
- $R = R(s) \geq 0$: relief energy (meter)

Fig. 1. Site specification in grid-cell system and variables on the site

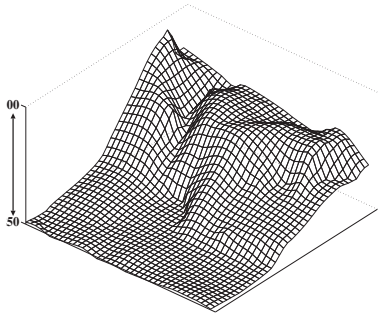


Fig. 2. Relief energy (R)

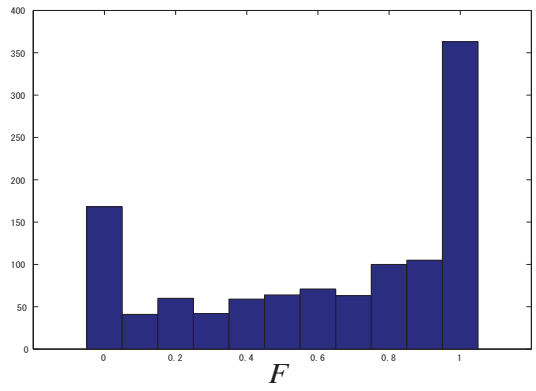


Fig. 3. Data distribution of F in Wuhan, China, data size: 1136

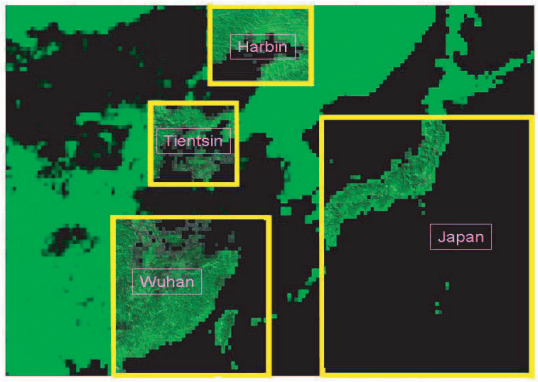


Fig. 4. Test areas; Japan, Harbin, Tienshin and Wuhan

last transform aims to use all values of forest areal rates F. Unfortunately, the predicted value of F based on these models will exceed this limit in many cases.

Fig. 3 shows the frequencies of the forest areal rate F. It is seen that the the probabilities such that F takes zero or one are fairly large (**zero-one inflated**). Therefore, it is natural to fit the zero-one inflated beta distribution [4] with the following cumulative distribution function $P(F \leq f)$:

$$P(F \leq f) = \begin{cases} 0 & \text{if } f < 0 \\ c_0 & \text{if } f = 0 \\ c_0 + c_1 \displaystyle\int_0^f \nu(t)dt & \text{if } 0 < f < 1 \\ 1 & \text{if } f \geq 1 \end{cases}$$

where c_0 and c_1 are positive constants with $c_0 + c_1 \leq 1$, and $\nu(t)$ is a probability density function of beta distribution Beta$(\mu\phi, (1-\mu)\phi)$ defined by $\nu(t) = \dfrac{t^{\mu\phi-1}(1-t)^{(1-\mu)\phi-1}}{B(\mu\phi,(1-\mu)\phi)} I_{(0,1)}(t)$. Here, $0 < \mu < 1$ and $\phi > 0$ are constants, $I_{(0,1)}(t)$ is the indicator function of interval $(0,1)$, and $B(\cdot,\cdot)$ denotes the beta function. This implies that F follows a discrete-type distribution with $P(F=0) = c_0$ and $P(F=1) = 1 - c_0 - c_1 \geq 0$, and the conditional distribution of F given the condition $0 < F < 1$ follows the continuous-type distribution with density $\nu(t)$.

The remaining parts of the work are organized as follows. In Section 2, a new family of zero-one inflated distribution based on Gamma and binomial distributions is proposed. Section 3 gives the further idea to propose zero-one inflated beta-type distributions with spatial dependence based on Markov random fields (MRF) for ratio data. This family will be an important tool for analyzing the zero-one inflated ratios which are spatially correlated. Section 4 concludes the work and states the issue under examination.

2. A NEW FAMILY OF ZERO-ONE INFLATED BETA DISTRIBUTIONS

Let $X_1, \ldots, X_n$ be independent random variables following the same gamma distribution $\Gamma(\alpha, 1)$ for positive constant α, and let M be a random variable following the binomial distribution with n: number of trials, and p: probability of success. Then,

we define a random variable Y by the mixture distribution as follows.

$$Y = \frac{X_1 + \cdots + X_M}{X_1 + \cdots + X_n}.$$

Obviousely, it satisfies that

$$P(Y = 0) = P(M = 0) \equiv p(0; n, p, \alpha) = (1 - p)^n \text{ and } P(Y = 1) = P(M = n) \equiv p(1; n, p, \alpha) = p^n.$$

Also, Y takes continuous values in $(0, 1)$. Hence Y is **zero-one inflated**. If $0 < m < n$, it is known that $(X_1 + \cdots + X_m)/(X_1 + \cdots + X_n)$ follows the beta distribution Beta$(m\alpha, (n - m)\alpha)$. Hence for $0 < y < 1$, the distribution function of Y is obtained by

$$P(0 < Y < y) = E\, P(0 < Y < y \mid M > 0) \quad = \quad \sum_{k=1}^{n-1} \frac{\int_0^y t^{\alpha k - 1}(1 - t)^{\alpha(n-k)-1} dt}{B(\alpha k, \alpha(n - k))} \times {}_nC_k p^k (1 - p)^{n-k}.$$

Taking the derivative by y, we have the probability density function at $y \in (0, 1)$:

$$p(y; n, p, \alpha) = \sum_{k=1}^{n-1} \frac{{}_nC_k p^k (1 - p)^{n-k}}{B(\alpha k, \alpha(n - k))} y^{\alpha k - 1}(1 - y)^{\alpha(n-k)-1}.$$

Furthermore, the density function and moment generating function can be expressed in the closed form, and we have

$$E\,(Y) = p \text{ and } Var\,(Y) = \frac{(\alpha + 1)p(1 - p)}{\alpha n + 1}.$$

Generalized linear models for the mean value p will give a new tool for analyzing the forest areal rates. If $n = 2$, this distribution is a special case of zero-one inflated beta distribution [4].

Suppose that the random variable Y denoting a ratio is related to the feature vector $\boldsymbol{x}$. Then, the mean function $E(Y)$ could be specified by the logistic function:

$$E(Y) = \frac{1}{1 + \exp(-\boldsymbol{\beta}^T \mathbf{x})} \text{ or } \frac{1}{1 + \exp\{-g(N) - h(R)\}}. \tag{1}$$

The latter case is for the forest areal rate with feature variables N and R.

Let r be a positive constant. Then, a new random variable defined by $Z \equiv Y^r$ has also a moment generating function expressible in a closed form. This transformation makes the family rich enough for application to ratio data.

3. ZERO-ONE INFLATED BETA-TYPE DISTRIBUTIONS WITH SPATIAL DEPENDENCE

Now, we propose zero-one inflated beta-type distributions with spatial dependence. Let $\mathcal{D}$ be a set of sites in a multispectral image. The pixels are numbered from 1 to k, and $\mathcal{D} = \{1, \ldots, k\}$. Note that the pixels are small unit areas in the surface of the earth. Assume that a pair $(y_i, \boldsymbol{x}_i)$ is observed at each pixel i in $\mathcal{D}$, where $y_i \in [0, 1]$ is a forest areal rate, and $\boldsymbol{x}_i$ is a multidimensional feature (explanatory) vector. We fit the hidden Markov model for the ratio y_i.

Let $M_i; i \in \mathcal{D}$ be a binomial random variable which determines the zero-one inflated beta-type distribution. Let $U_i \subset \mathcal{D}$ be

a set of neighbors of i to which y_i is dependent. Then, we assume that M_i's follow the Markov random field:

$$P(M_1 = m_1, ..., M_k = m_k) = p(m_1, ..., m_k) = \frac{1}{Z} \prod_{i=1}^{k} {}_nC_{m_i} p_i^{m_i} (1 - p_i)^{n - m_i} \exp \left\{ -\beta \sum_{j \in U_i} (m_i - m_j)^2 \right\} \tag{2}$$

where $m_i \in \{0, 1, ..., n\}$, Z is a normalizing factor, $p_i = f(\boldsymbol{x}_i) \in (0, 1)$ is a function of the feature vector $\boldsymbol{x}_i$, and $\beta \geq 0$ gives a spatial dependency parameter of M_i's. If $\beta = 0$, the joint distribution of M_i's is reduced to the independent binomial distributions.

Furthermore, we assume **the conditional independence**: $p(y_1, ..., y_k \mid m_1, ..., m_k) = \prod_{i=1}^{k} p(y_i \mid m_i)$ with

$$p(y|m) = \begin{cases} \delta(y) & \text{if } m = 0 \\ \delta(y - 1) & \text{if } m = n \\ \dfrac{y^{\alpha m - 1} (1 - y)^{(n-m)\alpha - 1}}{B(m\alpha, (n - m)\alpha)} I_{(0,1)}(y) & \text{if } 0 < m < n \end{cases} \tag{3}$$

where $\delta(y) := 1$ if $y = 0$, $:= 0$ otherwise. Now, $p \in (0, 1)$ is the success probability of the binomial random variate M_i. Using the formulas (2) and (3), we have the following joint distribution:

$$p(y_1, ..., y_k, m_1, ..., m_k) = \left[\prod_{i=1}^{k} p(y_i \mid m_i) \right] p(m_1, ..., m_k). \tag{4}$$

If the number of trials equals to $n = 2$ (the simplest case), we can see that $M_i = 0$ if $y_i = 0$; $M_i = 2$ if $y_i = 1$; and $M_i = 1$ if $0 < y_i < 1$. In this case, $p(y_1, ..., y_k, m_1, ..., m_k)$ can be expressed in the closed form, and it is possible to obtain the estimates specifying the model (4).

4. DISCUSSION

We consider stochastic distributions of ratio data with inflation at zero and one. After the review of the zero-one inflated beta distribution [4], new families of the inflated beta-type distributions with spatial dependence are proposed. We note here that logistic regression models will be promising candidates. The proposed methods are now examinined at the four areas in Fig. 4. It can be seen as a general trend that the proposed methods show an excellent result in comparison of the result [5].

5. REFERENCES

[1] P. Cottle, "Insuring southeast Asian commercial forests: Fire risk analysis and the potential for use of data in risk pricing and reduction of forest fire risk," *Mitigation and Adaptation Strategies for Global Change*, vol. 12, no. 1, pp. 181–201, 2007.

[2] E. Lambin, "Modelling and monitoring land-cover change processes in tropical regions," *Progress in Physical Geography*, vol. 21, pp. 375–393, 1997.

[3] N. Myers, "The world's forests and human populations: the environmental interconnections," *Population and Development Review*, vol. 16 (supplement), pp. 1–15, 1990.

[4] R. Ospina and S. L. P. Ferrari, "Inflated beta distributions," *Statistical Papers*, vol. 51, no. 1, pp. 111–126, 2010.

[5] S. Tanaka and R. Nishii, "Non-linear regression models to identify functional forms of deforestation in East Asia," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2617–2626, 2009.