Optimal sampling system for estimating geographical
distributions of natural resource and
environmental pollution

Yasushi Taga Kazumasa Wakimoto Minoru Ichimura
Shizuoka Univ. Okayama Univ. Okayama coll.of Sci.

1. Introduction

Suppose some kind of natural resource or chemical substance be distributed geographically according to an unknown density function $f(\underline{x})$ in a certain region D .

Suppose $f(\underline{x})$ be approximated fairly well in D by $\exp\{P_k(\underline{x})\}$ where $P_k(\underline{x})$ is a polynomal of degree k , and let $z_1, \dots, z_m$ be observed values of the density $f(\underline{x})$ at m points $\underline{x}_1, \dots, \underline{x}_m$ to be selected randomly in the region D . Then the coefficients of $P_k(\underline{x})$ may be estimated by the least squares method such that the sum of squares of differences between u_i 's and $P_k(\underline{x}_i)$'s is to be minimized where $u_i = \log z_i$, $1 \leq i \leq m$.

Thus we can get the estimated density function $\hat{f}(\underline{x})$ of $f(\underline{x})$ in D , and then the whole region D may be divided into sub-regions $D_1, \dots, D_l$ by the method for optimum stratification in sampling theory.

Then n_i points are randomly selected in each D_i ($1 \leq i \leq l$) so that the integral $I = \int_D f(\underline{x}) d\underline{x}$ is to be estimated precisely.

2. Estimation of $f(\underline{x})$

In the approximate expression $\exp\{P_k(\underline{x})\}$ of the density function $f(\underline{x})$, we may take the degree k of polynomal $P_k(\underline{x})$ for some suitable even number. For simplification we shall

state our method in the case where $k=2$ and the dimension of $\underline{x}$ is also two, i.e. $\underline{x}=(\xi, \eta)$.

Let $P_2(\underline{x})$ be quadratic function of $\underline{x}$ expressed as

$$(2.1) \quad P_2(\underline{x}) = a\xi^2 + b\xi\eta + c\eta^2 + d\xi + e\eta + g,$$

and m random points $\underline{x}_1=(\xi_1, \eta_1), \dots, \underline{x}_m=(\xi_m, \eta_m)$ be mutually independently and identically distributed according to the uniform distribution in the region D . Then the values of density $f(\underline{x})$ at each point $\underline{x}_i$ are measured, and denoted by z_i for $i=1, 2, \dots, m$. Here we assume that measurements of density could be done without errors, i.e. $z_i=f(\underline{x}_i)$ for $i=1, 2, \dots, m$.

Now we consider that $u_i = \log z_i$ may be approximated fairly well by $P_2(\underline{x}_i) = a\xi_i^2 + b\xi_i\eta_i + c\eta_i^2 + d\xi_i + e\eta_i + g$ for $i=1, 2, \dots, m$, and so that coefficients $a, b, \dots, g$ could be estimated by the least squares method, namely by minimizing the sum S of squares of differences between u_i and $P_2(\underline{x}_i)$. Let us denote estimates of $a, b, \dots, g$ by $\hat{a}, \hat{b}, \dots, \hat{g}$ which can be obtained by solving the following equations:

$$(2.2) \quad \frac{\partial S}{\partial a} = 0, \quad \frac{\partial S}{\partial b} = 0, \quad \frac{\partial S}{\partial c} = 0, \quad \frac{\partial S}{\partial d} = 0, \quad \frac{\partial S}{\partial e} = 0 \quad \text{and} \quad \frac{\partial S}{\partial g} = 0,$$

$$\text{where } S = \sum_{i=1}^m [u_i - P_2(\underline{x}_i)]^2.$$

This system of equations reduces to the following one:

$$(2.3) \quad \begin{aligned} & \hat{g} = \bar{u} - (\hat{a}\bar{\xi}^2 + \hat{b}\bar{\xi}\bar{\eta} + \hat{c}\bar{\eta}^2 + \hat{d}\bar{\xi} + \hat{e}\bar{\eta}), \\ & \hat{a}s_{20 \cdot 20} + \hat{b}s_{20 \cdot 11} + \hat{c}s_{20 \cdot 02} + \hat{d}s_{20 \cdot 10} + \hat{e}s_{20 \cdot 01} = s_{\cdot 20} \\ & \hat{a}s_{11 \cdot 20} + \hat{b}s_{11 \cdot 11} + \hat{c}s_{11 \cdot 02} + \hat{d}s_{11 \cdot 10} + \hat{e}s_{11 \cdot 01} = s_{\cdot 11} \\ & \hat{a}s_{02 \cdot 20} + \hat{b}s_{02 \cdot 11} + \hat{c}s_{02 \cdot 02} + \hat{d}s_{02 \cdot 10} + \hat{e}s_{02 \cdot 01} = s_{\cdot 02} \\ & \hat{a}s_{10 \cdot 20} + \hat{b}s_{10 \cdot 11} + \hat{c}s_{10 \cdot 02} + \hat{d}s_{10 \cdot 10} + \hat{e}s_{10 \cdot 01} = s_{\cdot 10} \\ & \hat{a}s_{01 \cdot 20} + \hat{b}s_{01 \cdot 11} + \hat{c}s_{01 \cdot 02} + \hat{d}s_{01 \cdot 10} + \hat{e}s_{01 \cdot 01} = s_{\cdot 01} \end{aligned}$$

$$\begin{aligned} \text{where } S_{\alpha\beta} &= \frac{1}{m} \sum_{i=1}^m (u_i - \bar{u})(\xi_i^\alpha \eta_i^\beta - \overline{\xi^\alpha \eta^\beta}) \\ S_{\alpha\beta \cdot \gamma\delta} &= \frac{1}{m} \sum_{i=1}^m (\xi_i^\alpha \eta_i^\beta - \overline{\xi^\alpha \eta^\beta})(\xi_i^\gamma \eta_i^\delta - \overline{\xi^\gamma \eta^\delta}) \\ \bar{u} &= \frac{1}{m} \sum_{i=1}^m u_i, \quad \overline{\xi^\alpha \eta^\beta} = \frac{1}{m} \sum_{i=1}^m \xi_i^\alpha \eta_i^\beta, \end{aligned}$$

and $\alpha, \beta, \gamma, \delta = 0, 1, 2$.

Solving the system of equations (2.3) we can obtain the estimated function $\hat{f}(\underline{x})$ of density function $f(\underline{x})$ in D . The precision of approximation may be given by the correlation coefficient $r_m = \text{Cov}(\hat{f}, f) / S_{\hat{f}} S_f$,

$$\text{where } S_f^2 = \frac{1}{m} \sum_{i=1}^m [f(\underline{x}_i)]^2 - \left[\frac{1}{m} \sum_{i=1}^m f(\underline{x}_i) \right]^2,$$

$$S_{\hat{f}}^2 = \frac{1}{m} \sum_{i=1}^m [\hat{f}(\underline{x}_i)]^2 - \left[\frac{1}{m} \sum_{i=1}^m \hat{f}(\underline{x}_i) \right]^2,$$

$$\text{and } \text{Cov}(\hat{f}, f) = \frac{1}{m} \sum_{i=1}^m \hat{f}(\underline{x}_i) f(\underline{x}_i) - \left[\frac{1}{m} \sum_{i=1}^m \hat{f}(\underline{x}_i) \right] \left[\frac{1}{m} \sum_{i=1}^m f(\underline{x}_i) \right].$$

The above method may apply to the density function $f(\underline{x})$ with a polynomial $P_k(\underline{x})$ of any degree k , but $f(\underline{x})$ with $P_4(\underline{x})$ of degree four is well applicable to various situations by our experiences.

3. Estimation of I and Optimum Subdivision of D

The total amount or integral $I = \int_D f(\underline{x}) d\underline{x}$ may be estimated by

$$(3.1) \quad T = \frac{A}{n} \sum_{j=1}^n f(\underline{x}_j)$$

where A denotes the area D and $\underline{x}_1, \dots, \underline{x}_n$ are independently and identically distributed according to the uniform distribution $U(D)$ in the region D .

It is easily shown that T is an unbiased estimator of I , i.e.

$$(3.2) \quad E\{T\} = I,$$

and the variance of T is given by

$$(3.3) \quad V(T) = (A/n) \int_D [f(\underline{x}) - \mu]^2 d\underline{x},$$

where $\mu = I/A$ is the mean density in D .

Now let us suppose D is divided into l subregions $D_1, \dots, D_l$, and n_i points $\underline{x}_{i1}, \dots, \underline{x}_{in_i}$ are taken randomly in each D_i ($1 \leq i \leq l$). Then an unbiased estimator of I is given by

$$(3.4) \quad T_l = \sum_{i=1}^l (A_i/n_i) \sum_{j=1}^{n_i} f(\underline{x}_{ij}),$$

and the variance of T_l is given by

$$(3.5) \quad V(T_l | D) = \sum_{i=1}^l (A_i/n_i) \int_{D_i} [f(\underline{x}) - \mu_i]^2 d\underline{x},$$

where A_i denotes the area of D_i , $\mu_i = I_i/A_i$ the mean density in D_i , and $D = \{D_1, \dots, D_l\}$ a subdivision of D .

Let us call a subdivision D^* optimum if $V(T_l | D^*) \leq V(T_l | D)$ for any subdivision D under the condition that l , $n = \sum_{i=1}^l n_i$ and allocation $(n_1, \dots, n_l)$ are preassigned.

In case of proportional allocation ($n_i = nA_i/A$), an optimum subdivision $D^* = \{D_i^*\}$ is given such that

$$(3.6) \quad D_i^* = \{\underline{x}; t_{i-1}^* < f(\underline{x}) \leq t_i^*\},$$

where t_i^* 's are $(l-1)$ increasing real numbers satisfying the relations

$$(3.7) \quad t_i^* = \frac{1}{2} (\mu_i^* + \mu_{i+1}^*), \quad \mu_i^* = I_i^*/A_i^* \text{ for } 1 \leq i \leq l-1 \text{ (see [1], [2])}.$$

The optimum division points $t_1^*, \dots, t_{l-1}^*$ may be obtained making use of a simple recursive algorithm if the density function $f(\underline{x})$ is given.

However $f(\underline{x})$ is unknown to us in advance, so we have to use the estimated density function $\hat{f}(\underline{x})$ based on the first observation $z_1, \dots, z_m$ as stated in section 2. Then we obtain

approximately optimum division points $\hat{t}_1^*, \dots, \hat{t}_{l-1}^*$ using $\hat{f}(\underline{x})$, i.e. the empiric distribution function of the sample $z_1, \dots, z_m$.

By our experiences $V(T_l|D^*)$ is smaller than one twentieth of $V(T)$ when $l=5$. If $a \leq f(\underline{x}) \leq b$ for any $\underline{x}$ in D , and if l is sufficiently large, we can get the approximate value $\widetilde{V}(T_l|D^*)$ of $V(T_l|D^*)$ as

$$(3.8) \quad \widetilde{V}(T_l|D^*) = \frac{A^2(b-a)^2}{12l^2n}.$$

References

- [1] Taga, Y (1967): On optimum stratification for the objective variable based on concomitant variables using prior information; *Ann. Inst. Statist. Math.* 19, 101-129.
- [2] Isii, K. and Taga, Y. (1969): On optimal stratifications for multivariate distributions; *Skand. Aktuarietidskr.*, 52, 24-38.