American Journal of Theoretical and Applied Statistics
2013; 2(4): 102-109
Published online August 10, 2013 (http://www.sciencepublishinggroup.com/j/ajtas)
doi: 10.11648/j.ajtas.20130204.12

A method for topographical estimation of lake bottoms by *B*-spline surface

H. Bao[1], K. Fueda[2]

[1]Graduate School of Environmental Science, Okayama University, Okayama, Japan
[2]Graduate School of Environmental and Life Science, Okayama University, Okayama, Japan

Email address:
gev421256@s.okayama-u.ac.jp(H. Bao), fueda@ems.okayama-u.ac.jp(K. Fueda)

To cite this article:
H. Bao, K. Fueda. A Method for Topographical Estimation of Lake Bottoms by *B*-Spline Surface. *American Journal of Theoretical and Applied Statistics.* Vol. 2, No. 4, 2013; pp. 102-109. doi: 10.11648/j.ajtas.20130204.12

Abstract: The application of *B*-spline (Basis spline) surface to the estimation of the lake bottom topography is described.By using the analysis of a bivariate *B*-spline, the shape of the lake bottom is approximated.According to the validity of the estimation by the bivariate *B*-spline function the method is applied to the actual data of the lake depth.Surveys over the water area have more difficulties than those on land, and the measurement data are distributed quite irregularly. The locations of the measured data donot exist regularly over the lake.Those locations were distributed along with the wake of the boat on which the sample data were collected. The density of the data is quite high in some small regions and quite low in other wide regions.Based on such irregular data, we tried a statistical estimation.The regularized term with a penalty coefficient makesa proper approximation of the parameters of the *B*-spline functions. There are many factors, such that the number of knots, the locations of those knots, the number of *B*-spline functions and the coefficient of penalized term.Appropriate information criterion which has sufficient accuracy and a small amount of computation is applied for determination of the optimal model.

Keywords: *B*-spline surface, Cross-validation, Influence function, Generalized cross-validation, Surface model selection, Numerical computation, Topography of lake bottom

1. Introduction

In this study we approximate the topography of a lake bottom with our statistical method.The lake is Kojima Lake which is located in Okayama prefecture in Japan.KojimaLake is separated by the bank from Kojima bay and turns into a freshwater lake. The water quality of the closed water area like this lake tends to worsen because of sedimentations or the pollutants from the upper stream. For the improvement of the water quality, the dredging must be tried and that requires a detailed depth data of the lake.Comparedto land, a detailed survey of the lake depth is difficult,so we applied a statistic method.Based on the data measured from September 2010 to January 2011, we made the estimation by using *B*-splines. To make an optimal model selection, various information criteria are devised. When there is a large number of models, CV (cross validation) is difficult to use for its computational cost. The GCV_{IF}(generalized cross validation with influence function) is adopted because we can obtain almost the same information as CV and it has a smaller computational cost. Furthermore, in order to obtain a high accuracy, the technique of using the influence function, which we have proposed recently [1], is applied.We are able to obtain an optimal model by this method which can approximate the smooth topography of the lake bottom.At first we applied CV and GCV_{IF} for the estimation of the selected two subdomains of the lake.Afterevaluation of the methods, we approximated the topography of the whole domain by GCV_{IF} and selected values ofβ.

2. Method of Surface Approximation

2.1. Introduction to B-splines

The spline function is a piecewise-defined polynomial function.The combinations of spline functions must have a sufficient required smoothness at the places where those functions connect.The connection points are named knots.We can consist the *B*-spline function $M_{m,i}(x)$ of the required degree *r*-1 (order *r*) by the algorithm of de Boor-Cox [2-4].This calculation can be started by the first

step

$$M_{1,j}(x) = \begin{cases} (\xi_j - \xi_{j-1})^{-1} & (\xi_{j-1} \le x \le \xi_j), \\ 0 & (otherwise) \end{cases} \quad (1)$$

and the successive recurrence formula listed below

$$M_{r,j}(x) = \frac{(x-\xi_{j-r})M_{r-1,j-1}(x)+(\xi_j - x)M_{r-1,j}(x)}{\xi_j - \xi_{j-r}}, \quad (2)$$

where $\{\xi_k\}$,k=1-r,…,n+r are the knots and n is the total number of intervals for the approximation.

Usually,B-splines with order four (degree three) are used in the calculation. Along the x direction we set the knots $x_1, x_2, \ldots, x_p$, and the knots at both ends are four-folded.So, the total number of basis B-splines will be $p-4$.The univariate spline functions are shown in Fig. 1,where$\xi_{-3} = \xi_{-2} = \xi_{-1} = \xi_0 = 0, \xi_1 = 1, \xi_2 = 2, \xi_3 = 3, \xi_4 = \xi_5 = \xi_6 = \xi_7 = 4$.

We set the approximation for the three dimensional surface as

$$u(x,y) = \sum_{i=1}^{p_1}\sum_{j=1}^{p_2} w_{ij} M_i(x) N_j(y), \quad (3)$$

wherep_1, p_2 is the total number of basis B-splines $\{M_i(x)\}$,$\{N_j(y)\}$,respectively.In addition, these functions have the support $[\xi_{i-r}, \xi_i)$, $[\eta_{j-r}, \eta_j)$ for the $x$and$y$ directionsrespectively.The shape of the three dimensional B-splines are shown in Fig. 2 where p_1=p_2=2,$\xi_i = (i-1)\times 10$,$\eta_i = (i-1)\times 10$,i=1,2,…,6.

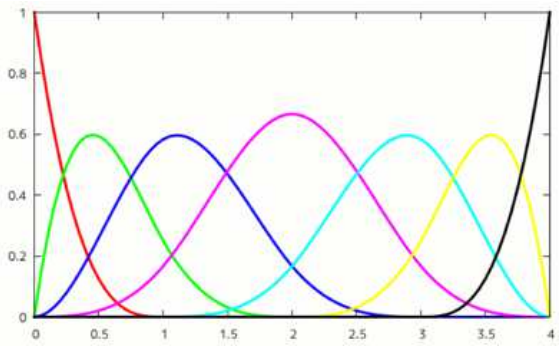


Figure1. *Spline Functions (Order Four)*

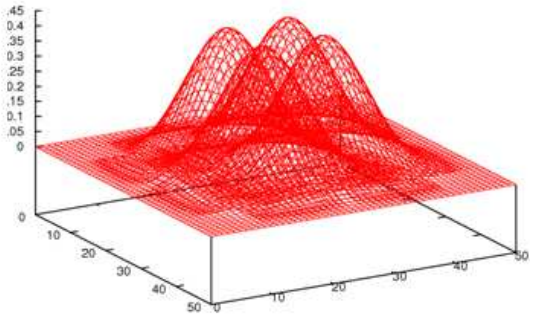


Figure2. *Three Dimensional SplineFunctions(Order Four)*

We have to satisfy the Schoenberg-Whitney condition [5] to determine the parameters w_{ij}.If there is no sample point in the domain $\{(x,y)|\xi_{i-r} \le x < \xi_i, \eta_{j-r} \le y < \eta\_j\}$, then we cannot solve the equations of the parameters.

2.2. Method of Parameter Estimation

For the nonlinear statistical modeling the maximum penalized likelihood methods are often used [6-8].Suppose that we have n observations $\{(z_\alpha, x_\alpha); \alpha = 1, \ldots, n\}$, where $z_\alpha$is the response variables generated from the unknown true distribution $G(z|x)$having a probability densityofg$(z|x)$and x_α is the vectors of explanatory variables.

We estimate w, which is a vector consisting of the unknown parameters and determines the model z=$u(x|w)$.Let$f(z_\alpha|x_\alpha;\theta)$be a specified parametric model, where θ is a vector of unknown parameters included in the model.The regression model with Gaussian noise is denoted as

$$z_\alpha = u(x_\alpha|w) + \varepsilon_\alpha, \varepsilon_\alpha \sim N(0,\sigma^2), \alpha = 1, \cdots, n \quad (4)$$

$$f(z_\alpha|x_\alpha;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{\{z_\alpha - u(x_\alpha;w)\}^2}{2\sigma^2}\right], \quad (5)$$

where$\theta = (w', \sigma^2)'$.The parameter will be determined by the maximization of the penalized log-likelihood function expressed as

$$\ell_\lambda(\theta) = \sum_{\alpha=1}^{n} \log f(z_\alpha|x_\alpha;\theta) - \frac{n}{2}\lambda H(w). \quad (6)$$

As the regularized term or penalized terms $H(w)$ with an m-dimensional parameter vector w, various types are used depending on the dimension of explanatory variables or the purpose of the analysis.For the three dimensional approximation[9], we use

$$H(w) = \iint \left\{\left(\frac{\partial^2 u}{\partial x^2}\right)^2 + \left(\frac{\partial^2 u}{\partial y^2}\right)^2\right\} dxdy, \quad (7)$$

$H(w)$ can be represented in the quadratic form by the $m \times m$ nonnegative matrix K as follows

$$H(w) = w'Kw. \quad (8)$$

Therefore, when we set $b_k(x_\alpha, y_\alpha) = M_i(x_\alpha)N_j(y_\alpha)$, k is determined by i and j,(6) will be

$$\ell_\lambda(\theta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(z-Bw)'(z-Bw) - \frac{n}{2}\lambda w'Kw, \quad (9)$$

where$z$=$(z_1,\ldots,z_n)$' and B is an $n\times m$ matrix composed of the basis functions as

$$B = \begin{bmatrix} b(x_1)' \\ b(x_2)' \\ \vdots \\ b(x_n)' \end{bmatrix} = \begin{bmatrix} b_1(x_1) & b_2(x_1) & \ldots & b_m(x_1) \\ b_1(x_2) & b_2(x_2) & \cdots & b_m(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ b_1(x_n) & b_2(x_n) & \cdots & b_m(x_n) \end{bmatrix} \quad (10)$$

By differentiating $\ell_\lambda(\theta)$ with respect to θ and setting the result equal to zero, we obtain the equations below

$$\frac{\partial \ell_\lambda}{\partial \sigma^2} = -\frac{n}{2\sigma^2} \pm \frac{1}{2\sigma^4}(z-Bw)'(z-Bw) = 0,$$

$$\frac{\partial \ell_\lambda}{\partial w} = \frac{1}{\sigma^2} B'(z - Bw) - n\lambda K w = 0 \text{ ,(11)}$$

By solving these equations, we have the estimation of the parameters by

$$\hat{w} = (B'B + n\lambda\hat{\sigma}^2 K)^{-1} B'z \text{ ,} \quad (12)$$

$$\hat{\sigma}^2 = \frac{1}{n}(z - B\hat{w})'(z - B\hat{w}) \text{ .} \quad (13)$$

For the predictive value$\hat{z}_\alpha = \hat{w}'b(x_\alpha)$, at each point of x_α, we obtain the vector of predicted values

$$\hat{z} = B\hat{w} = B(B'B + \lambda K)^{-1} B'z \text{ ,(14)}$$

where$\hat{z} = (\hat{z}_1, \hat{z}_2, \dots, \hat{z}_n)'$.At first, we set the constant value of $\beta = \lambda\hat{\sigma}^2$ and determine $\hat{w}$ for a given value of β.After that, we obtain the variance estimator $\hat{\sigma}^2$, and then we can obtain the smoothing parameter $\lambda = \beta/\hat{\sigma}^2$.

3. Model Selection

3.1. Cross Validation Criterion

From *n* observations the α-th data point (z_α, x_α) is removed and the parameter vector $\theta = (w', \sigma^2)'$ is estimated based on the remaining$n - 1$observations. We denote the parameter as $\hat{\theta}^{(-\alpha)} = (\hat{w}^{(-\alpha)\prime}, \hat{\sigma}^{2(-\alpha)})'$ .The corresponding estimated regression function is denoted as $\hat{u}^{(-\alpha)}(x)$. Weuse the log-likelihood forCross-Validation(CV) as

$$CV = -2\sum_{\alpha=1}^{n} \log\left(f\left(x_\alpha, \theta^{(-\alpha)}\right)\right)$$

$$= \sum_{\alpha=1}^{n}\left\{\log\left(2\pi\hat{\sigma}^{2(-\alpha)}\right) + \frac{\left(z_\alpha - \hat{u}^{(-\alpha)}\right)^2}{\hat{\sigma}^{2(-\alpha)}}\right\} \text{.(15)}$$

This is asymptotically equivalent to AIC (Akaike Information criterion)-type criteria such as AIC or BIC (Bayesian Information criterion) and so on [10-12].

Minimizing the (15) is a method for selecting an optimal model.Various alternative schemes are considered for the reduction of its computational costs.

3.2. Generalized Cross Validation with Influence Function

If the predicted value $\hat{z}$ is given in the form of $\hat{z} = Hz$, where *H* is a matrix that does not depend on the data *z*,then in cross-validation, the estimation process performed *n* times by removing observations one-by-one is not needed, and thus the amount of computation required can be reduced substantially.Because the matrix *H* transforms observed data z to predicted values $\hat{z}$, it is referred to as a hat matrix or is called a smoother matrix.The alternative scheme is called generalized CV (GCV)[13] which estimates the value of $\hat{u}^{(-\alpha)}(x_\alpha)$ as follows

$$GCV_{IF} = \sum_{\alpha=1}^{n}\left\{\log\left(2\pi\hat{\sigma}^{2(-\alpha)}\right) + \left[\frac{z_\alpha - \hat{u}(x_\alpha)}{\hat{\sigma}^{(-\alpha)}(1 - \frac{1}{n}\mathrm{tr}H)}\right]^2\right\} \text{,(16)}$$

where the matrix *H* is denoted in (14)as follows

$$H = B(B'B + \lambda K)^{-1} B'. \quad (17)$$

In (16) the value of $\hat{\sigma}^{2(-\alpha)}$ is also estimated by the influence function as follows

$$\hat{\theta}^{(-\alpha)} \approx \hat{\theta} - \frac{1}{n} T^{(1)}\left(z_\alpha; \hat{G}\right), \quad (18)$$

where$T^{(1)}(z_\alpha; \hat{G})$ is the influence function of $\hat{G}$ atz_α.
The general definition of the influence function is as follows.Its suitably normed limitinginfluence on the value of an estimate or test statistic T(ˆG) can be expressed as

$$T^{(1)}(x, G) = \lim_{\epsilon \to 0} \frac{T^{(1)}((1-\epsilon)G + \epsilon\delta_x) - T(G)}{\epsilon}, \quad (19)$$

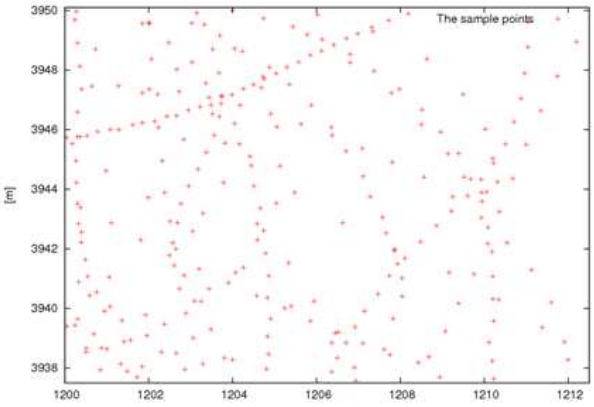
whereδ_x denotes the pointmass 1 at *x*.The above quantity, considered as a function of *x*,was introduced [14-15] under the name influence function, and is arguably the mostuseful heuristic tool of robust statistics.

4. Numerical Calculation

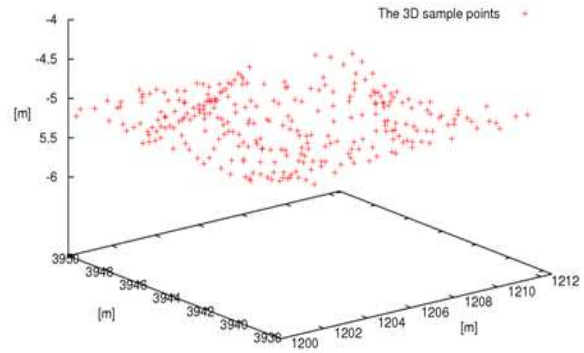
4.1. Computation for the Selected Areas

For the topographical estimation, we have acquired the experimental data of the lake.Thearea of the data is 2.79 km^2, and the total number of data is 1,178.At first, we tested the two subdomains of the whole data.After verifying the scheme to these two subdomains, we applied the scheme to the whole domain. We chose two subdomains where the distributions of the data are relatively uniform compared with other domains.Area I is an eastern area and Area II is a northern area.The distributions of data are shown in Fig. 3,4.

We used a small number of knots for the *x* and *y* directions, respectively, because of the irregularity of locations of samples.We tested10 to 18 knots for the *x* and *y* directions,respectively.For every combination of the number of knots, we generated 100 sets of the uniformly randomized *x*- and *y*-coordinates according to the density ofsamples,respectively.However some of them didnot satisfy the Schoenberg-Whitney condition,sowe generated another set of coordinates.Furthermore, if the equations of matrices made from ill-conditioned coordinates could not be solved properly, then we also generated other coordinates of knots.After solving (12) and (13), we have appliedgeneralized cross-validation with the influence function GCV_{IF}[1] as the information criterion.

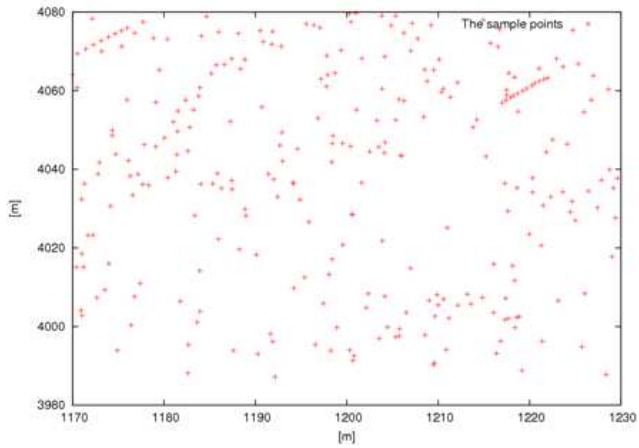


(a) Planar Distribution

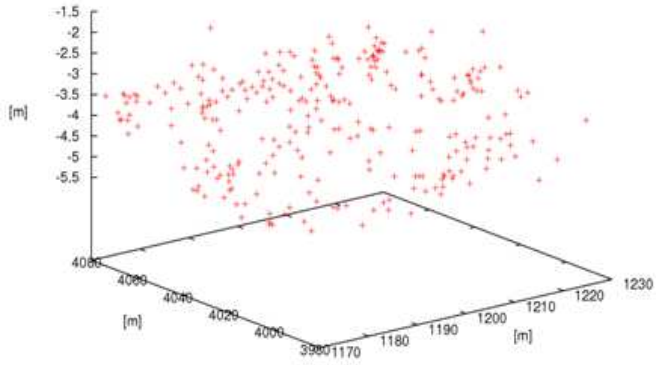


(b) ThreeDimensional Distribution

Figure3. *Distribution of Data over Area I*



(a) Planar Distribution



(b) Three Dimensional Distribution

Figure4. *Distribution of Data over Area II*

5. Conclusion

In the actual measurement data, some of the information criteria could not determine the optimal parameters.Inparticular, the solutions to small β's did not determine the shape of surfaces adequately. The major reason of those difficulties is considered as the irregularity of the distribution of data.If there is little data near the boundary of the domain, even if the surface changes sharply, the value of criterion will seldom be influenced. In spite of those difficulties, CV and the generalized CV with influence function (GCV_{IF}) can determine the optimal values to the various sets of data.The computational cost of GCV_{IF} is 1/50 of cross-validation.Furthermore, the selected optimal model by GCV_{IF} (Fig. 7) is better than that by CV(Fig. 5).We can assert that GCV_{IF} is just a practical method. This approximation method is able to contribute to the improvement of the water quality of Kojima Lake.

Table1. *CV Results for Area I*

total number of knots *x*-axis*y*-axis		β	σ^2	λ	CV
17	16	1.000E-00	0.003458	2.892E+02	-766.4
11	**10**	**1.000E-01**	**0.003292**	**3.038E+01**	**-776.8**
10	10	1.000E-02	0.003166	3.159E+00	-776.5
10	10	1.000E-03	0.003118	3.207E-01	-768.3
11	10	1.000E-04	0.002872	3.482E-02	-768.2
11	10	1.000E-05	0.002665	3.753E-03	-766.0
11	10	1.000E-06	0.002661	3.758E-04	-761.9
10	10	1.000E-07	0.002878	3.475E-05	-747.8
10	10	1.000E-08	0.002878	3.475E-06	-746.2
10	10	1.000E-09	0.002878	3.475E-07	-746.0
10	10	1.000E-10	0.002878	3.475E-08	-746.0

Table2. *CV Results for Area II*

total number of knots *x*-axis	*y*-axis	β	σ^2	λ	CV
16	18	1.000E+00	0.09638	1.038E+01	199.9
17	10	1.000E-01	0.09081	1.101E+00	203.6
11	10	1.000E-02	0.08468	1.181E-01	215.3
12	12	1.000E-03	0.07261	1.377E-02	219.9
12	11	1.000E-04	0.07048	1.419E-03	195.3
12	**11**	**1.000E-05**	**0.07015**	**1.426E-04**	**194.7**
12	11	1.000E-06	0.07012	1.426E-05	196.0
12	11	1.000E-07	0.07012	1.426E-06	197.0
12	11	1.000E-08	0.07012	1.426E-07	197.1
12	11	1.000E-09	0.07012	1.426E-08	197.1
12	11	1.000E-10	0.07012	1.426E-09	197.1

Table3. *GCV_{IF}ResultsforArea I*

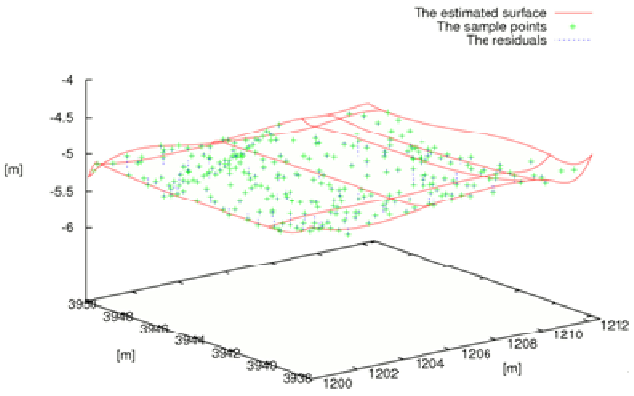
total number of knots *x*-axis	*y*-axis	β	σ^2	λ	GCV_{IF}
17	16	1.000E+00	0.003297	3.030E+02	-768.5
10	16	1.000E-01	0.003206	3.120E+01	-778.2
16	**10**	**1.000E-02**	**0.003125**	**3.200E+00**	**-779.1**
10	10	1.000E-03	0.003118	3.207E-01	-772.0
10	10	1.000E-04	0.003063	3.265E-02	-772.0
10	10	1.000E-05	0.002968	3.369E-03	-762.1
11	10	1.000E-06	0.002678	3.734E-04	-761.3
10	10	1.000E-07	0.002726	3.668E-05	-760.3
10	10	1.000E-08	0.002726	3.668E-06	-760.2
10	10	1.000E-09	0.002726	3.668E-07	-760.2
10	10	1.000E-10	0.002726	3.668E-08	-760.2

Table4. *GCV_{IF}Resultsfor Area II*

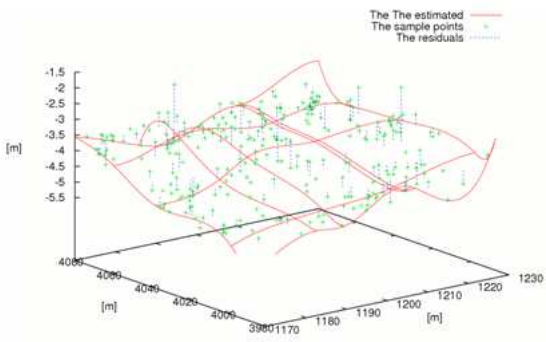
total number of knots *x*-axis	*y*-axis	β	σ^2	λ	GCV_{IF}
16	18	1.000E+00	0.09638	1.038E+01	190.6
18	18	1.000E-01	0.07264	1.377E+00	167.4
15	**14**	**1.000E-02**	**0.06267**	**1.596E-01**	**158.6**
11	15	1.000E-03	0.06199	1.613E-02	165.8
11	11	1.000E-04	0.07025	1.424E-03	173.7
11	10	1.000E-05	0.07728	1.294E-04	177.4
10	10	1.000E-06	0.07725	1.294E-05	177.7
10	10	1.000E-07	0.07725	1.294E-06	177.7
10	10	1.000E-08	0.07725	1.294E-07	177.7
10	10	1.000E-09	0.07725	1.294E-08	177.7
10	10	1.000E-10	0.07725	1.294E-09	177.7

Table 5. *GCV_{IF}Results over the Whole Area*

total number of knots *x*-axis	*y*-axis	β	σ^2	λ	GCV_{IF}
22	**19**	**1.000E-01**	**0.04969**	**2.013E+00**	**292.6**
21	18	1.000E-02	0.05146	1.943E-01	333.0
22	18	1.000E-03	0.06605	1.514E-02	347.7
22	18	1.000E-04	0.05017	1.993E-03	337.4
22	21	1.000E-05	0.09521	1.050E-04	318.6
22	18	1.000E-06	0.04231	2.364E-05	303.2
22	18	1.000E-07	0.04175	2.395E-06	299.2

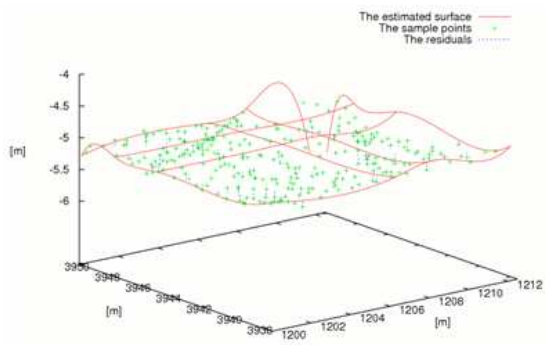


(a) Area I($\beta = 10^{-1}$)

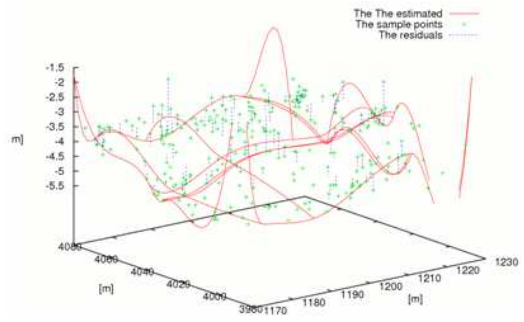


(b) AreaII($\beta = 10^{-5}$)

Figure5. *Selected Model (CV)*

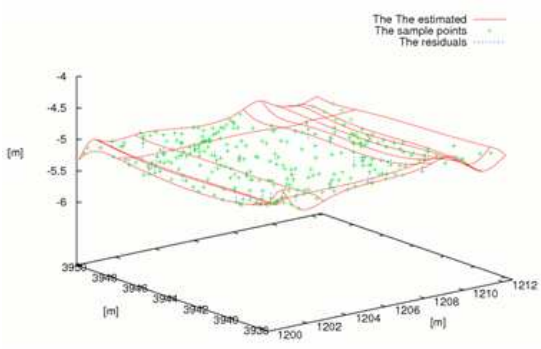


(a) AreaI($\beta = 10^{-7}$)

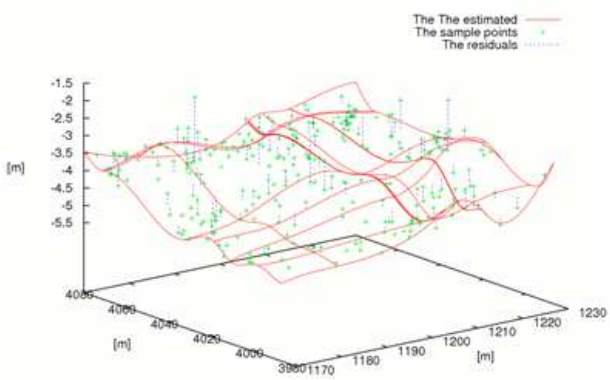


(b) AreaII($\beta = 10^{-7}$)

Figure6. *Selected Model (CV)*

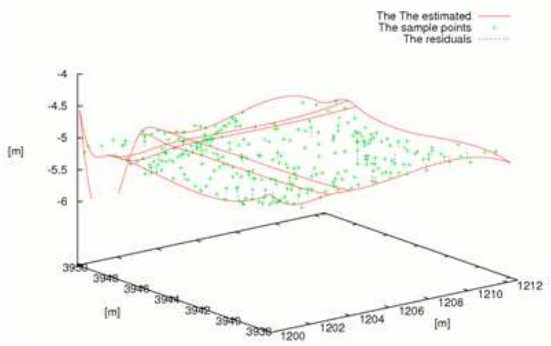


(a) Area I($\beta = 10^{-2}$)

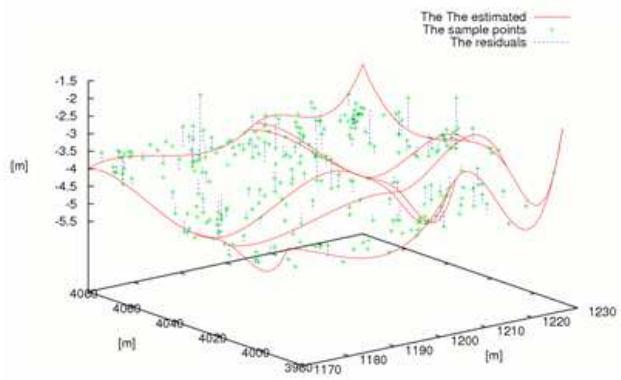


(b) AreaII($\beta = 10^{-2}$)

Figure7. *Selected Model (GCV_{IF})*

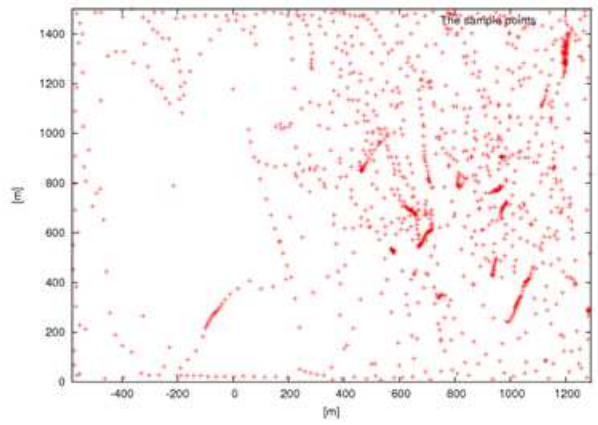


(a) Area I($\beta = 10^{-7}$)

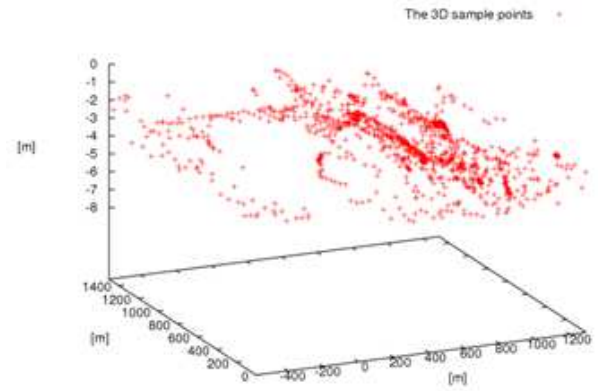


(b) AreaII($\beta = 10^{-7}$)

Figure8. *Selected Model (GCV_{IF})*



(a) Planar distribution



(b) Three Dimensional Distribution

Figure9. *Distribution of Data over Whole Area*

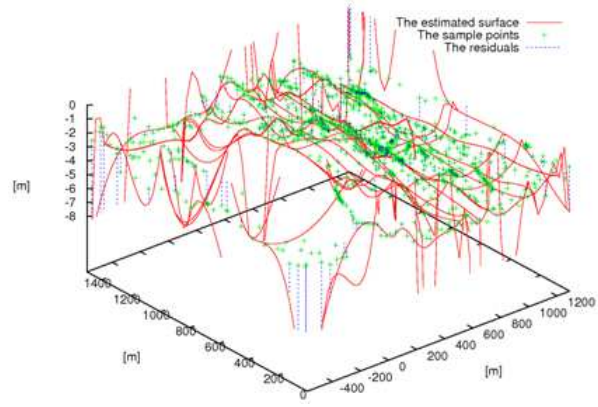
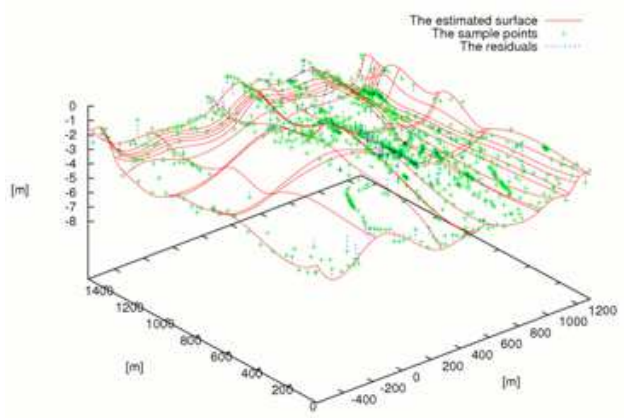
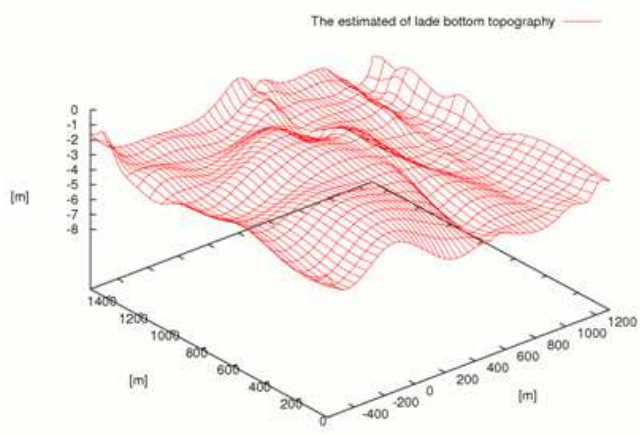


Figure11. *GCV_{IF} Results over the Whole Area* $(\beta = 10^{-7})$



(a) Samples and Estimated Surface



(b)Estimated Surface

Figure10. *Estimation of Lake Bottom Topography*

Acknowledgements

The authors are grateful for the help of Watanabe Laboratory of the graduate school of Environmental and Life Science Okayama University who provided measurement data of Kojima Lake.

References

[1] Bao, H., Fueda, K.(2013). "A new method with influence function for model selection in *B*-spline surface approximation",*ISRN Probability and Statistics*,(submitted to).

[2] Cox, M.G.(1972). "The numerical evaluation of *B*-splines", *J. Inst. Math. Appl.*, 10, pp.134-149.

[3] Cox, M.G.(1975). "An algorythm for spline interpolation", *J. Inst. Math. Appl.*,15, pp.95-108.

[4] de Boor, C.(1972). "On calculation with *B*-splines", *J. Approx. Theory*, 6, pp.50-62.

[5] Schoenberg, I. J., Whitney, A.(1953). "On Pólya frequency functions III", *Trans. Amer. Math. Soc*, Vol. 74. pp. 246-259, pp. 246-259.

[6] Good, I. J. and Gaskins, R.A.(1971). "Non parametric roughness penalties for probability densities", *Biometrika*, Vol. 58. pp. 255-277.

[7] Good, I. J. and Gaskins, R.A.(1980). "Density estimation and bump hunting by the penalized likelihood method exemplified by scattering and meteorite data", *Journal of American Standard Association*, Vol. 75. pp. 42-56.

[8] Green, P. J., Silverman, B. W.(1994). "*Nonparametric Regression and Generalized Linear Models*", Chapman and Hall, London.

[9] Umeyama, S. (1996). "Discontinuity extraction in regularization using robust statistics", *Technical report of IEICE.*,PRU95-217 (1996). pp. 9-16.

[10] Konishi, S., Kitagawa, G. (2008). "*Information Criteria and Statistical Modeling*", Springer Science+Business Media, LLC.

[11] Akaike H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automatic Control* 19(6), 716-723.

[12] Schwarz G. (1978). Estimating the dimension of a model. *Annals of Statistics*6(2), 461-464.

[13] Yuen K.V. (2010). *Bayesian methods for structural dynamics and civil engineering*. John Wiley and Sons, NJ.

[14] Hampel, F.R.(1968).“*Contributions to the theory of robust estimation*”, Ph.D. Thesis,University of California, Barkeley.

[15] Hampel, F.R.(1974).“The influence curve and its role in robust estimation”, *J. Amer.Statist. Assoc*., 62, 1179-1186.