Journal of the Operations Research Society of Japan
Vol. 53, No. 4, December 2010, pp. 289–304

SAMPLING SITE LOCATION PROBLEM IN LAKE MONITORING HAVING MULTIPLE PURPOSES AND CONSTRAINTS

Kazuya Haraguchi
Ishinomaki Senshu University

Yuichi Sato
Lake Biwa Environmental Research Institute

(Received June 16, 2010; Revised October 14, 2010)

Abstract *Monitoring* is a task of actual water collection in water quality assessment from lakes. In this paper, we consider the problem of locating water sampling sites which is a significant issue in the design of lake monitoring. We formulate the location problem so that it can handle multiple purposes and constraints arising in monitoring tasks, and design an algorithm based on *iterated local search* (*ILS*). In our experiments, we apply our formulation to the real situation of Lake Biwa, the largest freshwater body in Japan. The ILS based algorithm finds such a location of sampling sites from which we can achieve better estimation of water quality distribution over the entire lake than the existing location, where we assume the existence of the true distribution and generate it by Lake Biwa Basin Hydrological and Material Cycle Model. Also, we observe that some points are newly selected in the output solutions more frequently than others; such points can be interpreted as potential sampling sites.

Keywords: Optimization, mathematical modeling, lake monitoring, iterated local search

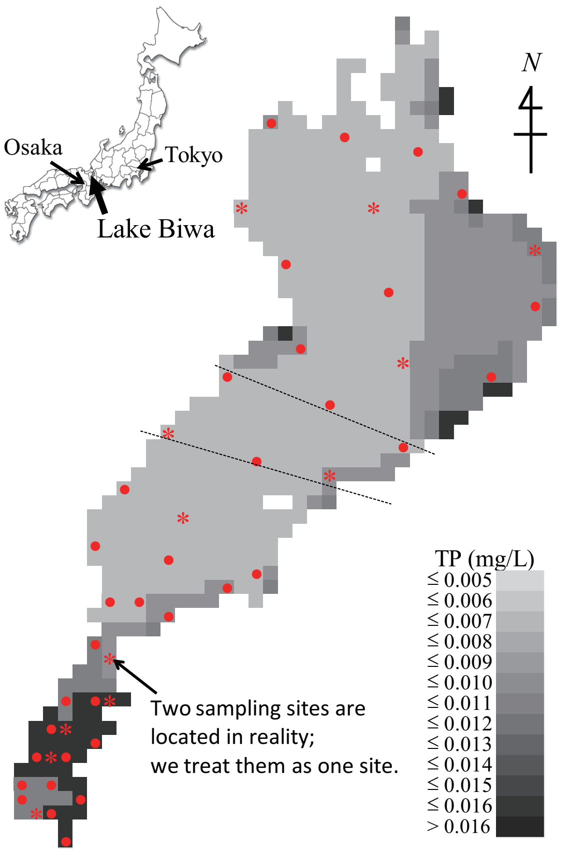
1. Introduction

1.1. Background

Monitoring is one of the significant tasks in water quality assessment from lakes, and refers to a task of actual collection of water. It is hard to design monitoring in general, and the strategies should be decided by lake use, lake problems being addressed, and the availability of resources for undertaking the assessment program [2].

In this paper, we mainly consider the monitoring task in Lake Biwa, the largest freshwater body in Japan. Lake Biwa is about 670km^2 in area, and its maximum depth is more than 100m. Located in Shiga Prefecture, it gives numerous benefits to the 14 million people in the Kinki region (which contains such big cities as Osaka, Kyoto and Kobe) by supplying vital water to their households and industries, by providing an abundant source of fishery products, and by offering tourists and residents a venue for rest and relaxation [16]. In Lake Biwa, water is sampled at 47 *sampling sites* once or twice a month. We show the existing location of the sampling sites (indicated by $*$ or $\bullet$) in Figure 1. We will explain details of the figure in later sections.

Location of sampling sites is one of the main concerns in designing lake monitoring. The existing sampling sites for Lake Biwa were decided at the end of 1970's; the surface of the lake is cut simply by approximately parallel lines from the west to the east (see two dashed lines in Figure 1 for example), and a couple of sampling sites are located on each line, with an approximately equal distance. However, it has been suggested that the data collected from these sampling sites do not necessarily contribute to clarification of the water quality system [4]. Also, if the budget were reduced by the organizer due to financial reasons, we



Two sampling sites are located in reality; we treat them as one site.

Figure 1: The existing location of the sampling sites for Lake Biwa

might need to prune some sampling sites and even to move the remaining ones to other points for balance. For risk management, we should discuss rearrangement of sampling sites under various conditions. These motivate us to establish a mathematical model of computing an "effective" location of sampling sites.

1.2. New contribution

There are some studies that attempt to compute the optimal location of sampling sites in the literature of water environmental research. In the previous studies, a location is evaluated only by goodness of water quality estimation. In fact, it is an essential goal to achieve good estimation of the water quality distribution over the entire lake only from the data collected at the sampling sites. However, it is not the unique purpose in general, and lake monitoring is usually conducted for multipurposes (e.g., examination of supplied water quality) under various constraints (e.g., polluted areas have to be sampled to some degree).

In this paper, we propose a new formulation of the problem of deciding the optimal location of sampling sites. Our formulation is so general that it can handle multiple purposes and constraints which often arise in a monitoring task. We refer to the formulated problem as the *sampling site location problem* (*SLP*). In SLP, a solution (i.e., a location of sampling sites) is represented as a subset S of the candidate point set P. We define the objective function as the weighted sum of several *value functions* on S, by which we can deal with multipurposes. The budget is a simplest constraint of a monitoring task; we restrict the cardinality $|S|$ to a positive constant. To represent other constraints, we require S to have overlap with specific point sets to some extent. We design an algorithm based on *iterated local search* (*ILS*) [7] to solve SLP, which is a general algorithmic framework for hard combinatorial optimization problems.

Another contribution of this paper is a case study of SLP on Lake Biwa. We take 3 value functions, each of which evaluates goodness of water quality estimation, continuity of existing sites, and the number of intakes, respectively. Among these, we define the first one as the difference (i.e., the sum of the squared error) between the "true" water quality distribution and the distribution estimated by an appropriate interpolating method where we use the data on S as the input. The difference appears hard to be evaluated since the "true" distribution is not available in general. Then we regard the distribution computed by *Lake Biwa Basin Hydrological and Material Cycle Model* (*LB-Model*) [14] as the true one, and employ *2-dimensional spline method under tension* (*SPT*) [12] as the interpolating method. In the computational experiments, our ILS based algorithm can find such a solution that achieves smaller difference than the existing solution although the number of sampling sites is smaller. Also, we find that some points are selected as the solution more frequently than others even though they are not in the existing solution; such points can be interpreted as potential sampling sites.

The paper is organized as follows. We mention related works in Section 2. In Section 3, we formulate the problem SLP and also describe LB-Model and SPT. We present the ILS based algorithm for SLP in Section 4, and then report experimental results in Section 5. We discuss other models to estimate the true distribution in Section 6, and give concluding remarks in Section 7.

2. Related Works

We consider lake monitoring in this paper, while *river monitoring* has been studied since 1970's. Sanders et al. published a standard book [13] of river monitoring, and discussed how to design its procedures. They dealt with not only location of sampling sites but also sampling frequency and water quality indices to be sampled. They stated that, among these issues, location of sampling sites is the most significant in general. There are several papers that attempt to compute the optimal location of sampling sites under their own definitions of optimality. For example, Dixon et al. [5] defined the cost of assigning a sampling site to a river section as the additional expense needed to find pollution source when pollution has been detected in the sampling site. To find the optimal location, they applied simulated annealing to minimize the total cost over all river sections. Alvarez-Vázquez et al. [1] assumed that the water quality is distributed by a differential equation system, and proposed an algorithm to find the location that minimizes the squared error sum between the model distribution and the estimated distribution; their idea is similar to ours in the sense that the "true" distribution is assumed.

In general, river monitoring problems are not so hard as lake monitoring ones. In the former, solution spaces can be modeled by such simple structures as 1D line or rooted tree, and it is not too academic even to assume monotonicity on water flow, by which the problems become more tractable; e.g., downstream pollution does not affect the water quality of the upper parts of the river. However, the water pollution mechanism in a lake is more complex. Lakes have three fundamental characteristics in common; integrating nature, long retention time, and complex response dynamics [10]. These characteristics make predicting the water quality changes harder.

For lake monitoring, there are several studies on optimal location of sampling sites. Matsuoka et al. [11] analyzed the optimal location problem using stochastic estimation. Fujiwara et al. [6] examined how to prune the existing sampling sites so that the prediction ability is not diminished to a large extent. More recently, Hedger et al. [8] investigated to calibrate the location by the spatial dynamics of the lake observed by remote sensing data.

Most of the previous studies deal with optimization of a single criterion, that is, goodness of water quality estimation. Our formulation is more general, where we can handle multiple purposes and constraints arising in lake monitoring. This is the crucial difference between the previous studies and our work.

3. Sampling Site Location Problem (SLP)

In this section, we formulate the problem SLP and show its ability to handle multiple purposes and constraints of lake monitoring, by discussing its application to the real situation of Lake Biwa. We then describe LB-Model and SPT, which we will use in the computational experiments in Section 5. We generate the true water quality distribution by LB-Model, and use SPT as the interpolating method to estimate the distribution over the entire lake.

3.1. Formulation

For simplicity, we do not consider the water depth but only the surface of the given lake. We may partition the surface by 2D grid, and approximate it by a set $P = \{p_1, p_2, \ldots, p_n\} \subseteq \mathbb{R}^2$ of n grid points on 2D plane. We specify each $p_i \in P$ by its coordinate values, and denote it by $p_i = (x_i, y_i)$. We assume that at most one sampling site is allocated to each $p_i \in P$. Then we refer to any subset $S \subseteq P$ as a *solution*.

We design SLP as a maximization problem, and define the objective function $f : 2^P \to \mathbb{R}$ as the weighted sum of several value functions on S. Let us denote by $\mathcal{V}$ the set of the value functions. Each $v \in \mathcal{V}$ is a function $v : 2^P \to [0,1]$, where the value $v(S)$ indicates how S accomplishes the criterion associated with v. We give a constant weight $w_v \in [0,1]$ to each $v \in \mathcal{V}$ so that $\sum_{v \in \mathcal{V}} w_v = 1$ is satisfied, showing the relative significance of v.

More sampling sites would deliver more useful information to us, but the monitoring budget is usually limited. The budget is mainly used for fuel charge of the ships, manpower cost and reagent cost, which are proportional to the number of sampling sites in general. We assert that the budget determines the number of sampling sites, and hence we restrict the cardinality $|S|$ to a positive constant m. To represent other constraints, we introduce a family $\mathcal{C} \subseteq 2^P$ of *constraint subsets*. We require S to have overlap with each $C \in \mathcal{C}$ to a certain degree, i.e., $b_C^- \leq |S \cap C| \leq b_C^+$ for constants b_C^- and b_C^+. We formulate the problem SLP as follows.

Sampling Site Location Problem (SLP)

$$\text{maximize} \quad f(S) = \sum_{v \in \mathcal{V}} w_v \cdot v(S) \tag{3.1}$$

$$\text{subject to} \quad b_C^- \leq |S \cap C| \leq b_C^+ \quad (\forall C \in \mathcal{C}) \tag{3.2}$$

$$|S| = m, \; S \subseteq P \tag{3.3}$$

An *SLP instance* is then defined by a 2D point set P, a set $\mathcal{V}$ of value functions, weights w_v's for each $v \in \mathcal{V}$, the number m of sampling sites, a family $\mathcal{C}$ of constraint subsets, and lower bounds b_C^-'s and upper bounds b_C^+'s for each $C \in \mathcal{C}$. We call a solution $S \subseteq P$ *feasible* (resp., *infeasible*) if it satisfies (resp., does not satisfy) both Equations (3.2) and (3.3).

It is possible that an SLP instance has no feasible solution. It must be hard to decide the feasibility of a given SLP instance efficiently (i.e., in polynomial time); even in the restricted case where $b_C^+ = +\infty$ for each $C \in \mathcal{C}$, the decision problem becomes *constrained set multicover problem* [17], which is a generalization of *set cover problem* (*SCP*), a well-known NP-hard problem. Then it must be also computationally hard to output a feasible solution or Null according to whether a given SLP instance is feasible or not.

3.2. Application to Lake Biwa

We consider constructing an SLP instance by following the real situation of Lake Biwa. We take north-south and east-west grid lines on the surface of Lake Biwa at 1km intervals. Then the number n of grid points amounts to $n = 677$, and the bounding box has 37 points in width and 59 points in height. Taking the point where the lowermost and the leftmost grid lines meet as the origin $(0, 0)$, we assume that each point $p_i = (x_i, y_i) \in P$ is integral, i.e., $x_i, y_i \in \mathbb{Z}$. Among the 47 existing sampling sites, some 2 sites are close and belong to the same grid point in this setting (as indicated in Figure 1). In the sequel, we regard them as one sampling site and the number of the existing sampling sites as 46. We denote by $S_{\text{exist}} \subseteq P$ the set of existing sampling sites (and thus we have $|S_{\text{exist}}| = 46$). We may call S_{exist} the *existing solution*, and each $p_i \in S_{\text{exist}}$ an *existing sampling site* (or an *existing site* for short).

We assert that the following criteria are essential and thus should be taken into account in evaluating a solution.

- Goodness of water quality estimation.
- Sampling at specific points: existing sampling sites, intakes of supplied water, coastal areas, and polluted areas.

Based on the interview with technical staffs in Shiga Prefectural Government, we define the set $\mathcal{V}$ of value functions as $\mathcal{V} = \{v_{\text{est}}, v_{\text{exist}}, v_{\text{intake}}\}$, and the family $\mathcal{C}$ of constraint subsets as $\mathcal{C} = \{S^*_{\text{exist}}, S_{\text{coast}}, S_{\text{pol}}\}$. For $\mathcal{V}$, the 3 value functions $v_{\text{est}}, v_{\text{exist}}, v_{\text{intake}}$ evaluate goodness of water quality estimation, the number of existing sites in a solution, and the number of intakes in a solution, respectively. For $\mathcal{C}$, the set $S^*_{\text{exist}} \subsetneq S_{\text{exist}}$ denotes the subset of the existing sites which we have to include in a solution, and S_{coast} and S_{pol} ($S_{\text{coast}}, S_{\text{pol}} \subseteq P$) are the sets of coastal points and polluted points, respectively.

We asked the interviewees to give nonnegative weights to the value functions so that the total weight amounts to 1. Taking the averages over the interviewees, we use $w_{\text{est}} = 0.324$ for v_{est}, $w_{\text{exist}} = 0.581$ for v_{exist}, and $w_{\text{intake}} = 0.095$ for v_{intake}; they consider that continuity of the existing sampling sites is more important than others. In the experiments, we will try several values for m to observe the change of obtained solutions. We describe the details of the value functions and the constraint subsets below.

Goodness of water quality estimation

For each sampling site, the sampled water is analyzed to examine its quality indices; e.g., COD (Chemical Oxygen Demand), TN (Total Nitrogen), TP (Total Phosphorus). The real values of these quality indices are available only for the sampling sites. To understand the water quality condition in the entire lake, it is desirable to achieve good estimation of the water quality distribution over the entire lake by using an appropriate interpolating method.

Focusing on one quality index, we assume the existence of a *true distribution* which we denote by a mapping $d : \mathbb{R}^2 \to \mathbb{R}$. Given a solution $S \subseteq P$, we assume that we can access the value $d(x_i, y_i)$ of any $p_i = (x_i, y_i) \in S$, but cannot access the value $d(x_j, y_j)$ of any $p_j = (x_j, y_j) \notin S$; this assumption comes from the fact that the real values are available only for S. We estimate d by using $d(x_i, y_i)$'s for each $p_i = (x_i, y_i) \in S$ and an interpolating method. Precisely, an interpolating method is an algorithm that outputs a mapping from $\mathbb{R}^2$

to $\mathbb{R}$ by using the set $\bigcup_{p_i=(x_i,y_i)\in S}\{(x_i, y_i, d(x_i, y_i))\}$ as the sample. For a given interpolating method and a solution S, we denote the *estimated distribution* by a mapping $\hat{d}_S : \mathbb{R}^2 \to \mathbb{R}$. We define $\varphi_{\mathrm{est}}(S)$ as the sum of the squared errors over all points in P;

$$\varphi_{\mathrm{est}}(S) = \sum_{p_i=(x_i,y_i)\in P} \left(\hat{d}_S(x_i, y_i) - d(x_i, y_i)\right)^2. \tag{3.4}$$

We define the value function $v_{\mathrm{est}} : 2^P \to [0, 1]$ by normalizing φ_{est} as follows;

$$v_{\mathrm{est}}(S) = \max\{0, -6.81 \times 10^3 \times \varphi^2_{\mathrm{est}}(S) - 4.13 \times 10^1 \times \varphi_{\mathrm{est}}(S) + 1.0\}, \tag{3.5}$$

where we decide the coefficients based on the interview; We asked the interviewees to give scores from 0 to 1 to some instance values of φ_{est}. We compute the quadratic regression by least squares method to see the scores as a function of φ_{est}'s, which results in the second quadratic function in the braces of the righthand in Equation (3.5). A smaller error sum $\varphi_{\mathrm{est}}(S)$ would be pleasing to us, and one can readily see that $v_{\mathrm{est}}(S)$ is monotone non-increasing with respect to $\varphi_{\mathrm{est}}(S) \geq 0$.

In the experiments, we focus on TP that is considered one of the most influential water quality indices upon the ecosystem of Lake Biwa. We generate the true distribution d by LB-Model, where we compute the averaged distribution over year 2004. (An overview of this d is shown by shading in Figure 1.) We use SPT for the interpolating method to construct an estimated distribution. We describe the technical overview of LB-Model and SPT in the next subsections.

There is no model that can capture the future of the lake perfectly. For example, LB-Model cannot predict accidental pollution occurred by abrupt increase of plankton. (To understand such unpredictable phenomena, we need to conduct a monitoring task periodically.) However, it must be a realistic approach for us to assume the existence of the true distribution which is generated by a mathematical model. LB-Model can achieve good estimation of a long-term (e.g., 1 year) average of the real water quality [14], which encourages us to employ it to generate the true distribution.

Sampling at specific points

For Lake Biwa, the monitoring task started more than 30 years ago, and a large amount of data has been stored so far. For consistency of the data, we should not move too many existing sampling sites to other points. For a solution $S \subseteq P$, we define $\varphi_{\mathrm{exist}}(S)$ as the size of the intersection between S and S_{exist}, showing continuity of the existing sites;

$$\varphi_{\mathrm{exist}}(S) = |S \cap S_{\mathrm{exist}}|. \tag{3.6}$$

We then define the value function $v_{\mathrm{exist}} : 2^P \to [0, 1]$ by normalizing φ_{exist} as follows;

$$v_{\mathrm{exist}}(S) = -1.55 \times 10^{-4} \times \varphi^2_{\mathrm{exist}}(S) + 2.89 \times 10^{-2} \times \varphi_{\mathrm{exist}}(S),$$

where the coefficients are determined based on the interview, similarly to the case of v_{est}. A larger $\varphi_{\mathrm{est}}(S)$ would be pleasing to us, and $v_{\mathrm{exist}}(S)$ is monotone increasing with respect to $\varphi_{\mathrm{exist}}(S) \in [0, 46]$. ($|S_{\mathrm{exist}}| = 46$.)

We cannot move some existing sites to other points since they are established as environmental standard points by law, or are used for sampling of deep water. A constraint subset $S^*_{\mathrm{exist}} \subsetneq S_{\mathrm{exist}}$ is the set of such points. We have $|S^*_{\mathrm{exist}}| = 12$, and the existing sites in S^*_{exist} (resp., $S_{\mathrm{exist}} \setminus S^*_{\mathrm{exist}}$) are indicated by $*$ (resp., $\bullet$) in Figure 1. We set the lower and the upper bounds as $b^-_{\mathrm{exist}} = b^+_{\mathrm{exist}} = |S^*_{\mathrm{exist}}|$.

Besides existing sites, there are some points that should be selected in the solution due to their own reasons. For example, Lake Biwa serves as a water source for domestic use. To monitor how the supplied water is affected by change of the water quality in the lake, we should locate some sampling sites near the intakes. For other example, since coastal water is more relevant to the scenery than offshore water, we may need to locate some sampling sites at the coastal areas. Also, we should pay more attention to the polluted areas than to the rest parts of the lake.

Let $S_{\rm intake}, S_{\rm coast}, S_{\rm pol} \subseteq P$ denote the sets of the intake points, the coastal points and the polluted points, respectively. We have $|S_{\rm intake}| = 14$, $|S_{\rm coast}| = 214$ and $|S_{\rm pol}| = 71$ for Lake Biwa. For a solution $S \subseteq P$, we define $\varphi_{\rm exist}(S)$ as the size of the intersection between S and $S_{\rm intake}$, showing how many intake points S contains;

$$\varphi_{\rm intake}(S) = |S \cap S_{\rm intake}|. \tag{3.7}$$

Then we define the value function $v_{\rm intake} : 2^P \to [0, 1]$ as follows;

$$v_{\rm intake}(S) = 7.14 \times 10^{-2} \times \varphi_{\rm intake}(S),$$

which is monotone increasing with respect to $\varphi_{\rm intake}(S)$.

Many interviewees considered that S should have overlap with $S_{\rm coast}$ and $S_{\rm pol}$ to some extent, i.e., $|S \cap S_{\rm coast}|$ and $|S \cap S_{\rm pol}|$ should be neither too small nor too large, compared with the number m of sampling sites. Then we treat $S_{\rm coast}$ and $S_{\rm pol}$ as constraint subsets. The interviewees answered suitable bounds on $|S \cap S_{\rm coast}|$ and $|S \cap S_{\rm pol}|$, by which we set $b^-_{\rm coast} = 0.4m$ and $b^+_{\rm coast} = 0.7m$ for the coastal points, and $b^-_{\rm pol} = 0.2m$ and $b^+_{\rm pol} = 0.5m$ for the polluted points.

3.3. Lake Biwa Basin Hydrological and Material Cycle Model (LB-Model)

LB-Model consists of three component models; *the land model*, *the lake flow model*, and *the lake ecological model*. Each model simulates hydrological and/or material cycle in Lake Biwa Basin, after reading input data about climate, land use, social situation, and so on, and output data from other models [14]. In other words, LB-Model calculates water quality and quantity in Lake Biwa Basin by coupling three component models. See Figure 2.

The land model simulates water quality and quantity on the land area (e.g., river discharge, river water quality, and groundwater quality). In this model, rain water is divided into evapotranspiration, infiltration, and surface runoff by the evapotranspiration model. Infiltrating water and surface runoff go into the groundwater model and the surface runoff model respectively, and flow into the river channel model, finally the lake flow/ecological models. Through this water flow, load from point (e.g., domestic waste water) and non-point (e.g., waste water from urban area) source are aggregated in each factor model.

The lake flow model simulates flow direction, flow velocity, and water temperature in the lake. Lake Biwa is divided into 1km meshes horizontally and 8 layers vertically, that is, 3 dimensional model. The base equations used in this model are dynamic equation of each dimension, water temperature balance equation, and equation of continuity.

The lake ecological model simulates lake water quality (e.g., COD, TN and TP) by calculating biochemical process, by which we generate the true distribution d. The structure of Lake Biwa can be approximated by the lake flow model (i.e., 3 dimensional model), and in this model, food web is simulated among nutrients, plankton, fish, detritus, and so on.

3.4. 2-dimensional spline method under tension (SPT)

In this paper, we employ SPT [12] to compute an estimated distribution $\hat{d}_S$ for a given solution $S \subseteq P$. SPT is one of the interpolating methods that estimate the distribution of

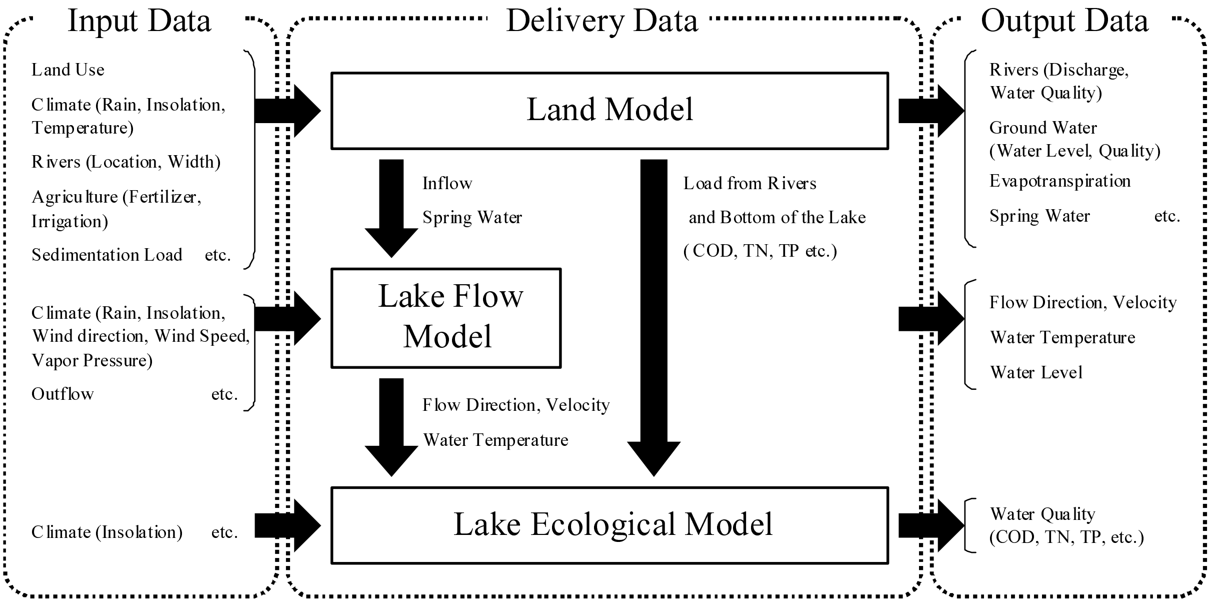


Figure 2: Overview of LB-Model

water quality, and has been applied for water pollution studies over 20 years [11]. Based on the infinitesimal displacement theory on elastic body, SPT proposes to estimate the true distribution $d : \mathbb{R}^2 \to \mathbb{R}$ by such $\hat{d}_S : \mathbb{R}^2 \to \mathbb{R}$ that minimizes the energy function $E(\hat{d}_S)$ defined as follows;

$$E(\hat{d}_S) = \int_R \big((\Delta \hat{d}_S)^2 + \sigma(\nabla \hat{d}_S)^2\big)dxdy, \tag{3.8}$$

among those satisfying $\hat{d}_S(x_i, y_i) = d(x_i, y_i)$ for any $p_i = (x_i, y_i) \in S$. In Equation (3.8), R denotes a sufficiently large 2D region containing the considered point set P, σ denotes a positive parameter controlling the function shape of $\hat{d}_S$ (where $\hat{d}_S$ becomes smoother if σ gets smaller) and Δ and ∇ are operators defined as $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ and $\nabla = \frac{\partial}{\partial x}e_x + \frac{\partial}{\partial y}e_y$, where e_x and e_y are basis vectors. The boundary condition is given as follows;

$$\frac{\partial \hat{d}_S}{\partial \mu} = \Delta \hat{d}_S = 0 \text{ on } \partial R,$$

where $\frac{\partial}{\partial \mu}$ denotes the partial differentiation in the direction of the normal from the boundary region ∂R to its outside.

The paper [12] considers how to compute $\hat{d}_S$ by difference equation system. The system updates $\hat{d}_S(p_i)$ of each point $p_i \in P$ iteratively, based on the $\hat{d}_S(p_j)$'s of neighbor points p_j's and the parameter σ. In the experiments, we set $\sigma = 0.3$, by which $\hat{d}_S$ becomes smooth enough, and the number of iteration times to 200 which is sufficient for convergence.

4. Iterated Local Search (ILS) Based Algorithm

In this section, we present our heuristic algorithm based on iterated local search (ILS) [7] for SLP. The motivation is described as follows: If the number n of grid points and the number m of sampling sites were relatively small, then we might be able to obtain the optimal solution by examining all solutions of size m. However, we have $n = 677$ for Lake

Biwa, and if we take $m = 46$ (i.e., the same size as the existing solution), the number of combinations amounts to $\binom{677}{46} > 10^{71}$. Then it is smarter to design an efficient algorithm that delivers a nearly optimal solution.

Our algorithm calls local search as a subroutine iteratively. Let $S_{\text{init}} \subseteq P$ denote an *initial solution*. Starting with $S = S_{\text{init}}$, one local search repeats searching the *neighborhood* of S for a better solution S' (in the sense of the objective) and setting $S \leftarrow S'$ if such S' exists in the neighborhood; otherwise, it returns S to the main routine.

However, as described in Section 3.1, it must be hard to generate a feasible solution efficiently for S_{init}. In order to resolve this issue, we expand the search space into all solutions of size m (including infeasible ones), by which we can take S_{init} as any set of m grid points. Taking the degree of constraint violation into account, we need to extend the objective function so that it evaluates not only a feasible solution but also an infeasible one. For each constraint subset $C \in \mathcal{C}$, we introduce the *penalty function* ρ_C defined as follows;

$$\rho_C(S) = \begin{cases} b_C^- - |S \cap C| & \text{if } |S \cap C| < b_C^-, \\ |S \cap C| - b_C^+ & \text{if } |S \cap C| > b_C^+, \\ 0 & \text{otherwise.} \end{cases}$$

Then we define an alternative objective function $\tilde{f}$ as follows, based on the original objective function f defined in Equation (3.1);

$$\tilde{f}(S) = f(S) - M \sum_{C \in \mathcal{C}} \rho_C(S), \tag{4.1}$$

where M denotes a sufficiently large constant. We see that, if S is feasible, then we have $\tilde{f}(S) = f(S)$. Otherwise, $\tilde{f}(S)$ becomes much smaller and then S is rated as a poor solution. Note that penalty function is not a new notion but is frequently used in the optimization literature (e.g., Chapter 48 of the handbook [7]).

Algorithm 1 shows a summary of our ILS based algorithm. ILS-SLP is the main routine. It calls the subroutine LS-SLP iteratively, which corresponds to one local search. The τ and δ are positive parameters, where τ represents the number of iteration times and δ represents the radius of the neighborhood of a point or a solution. For a given δ, we define the neighborhood $N_\delta(p_i) \subseteq P$ of a point $p_i \in P$ as follows;

$$N_\delta(p_i) = \{p_j \in P \mid ||p_i - p_j||_1 \leq \delta\},$$

where $||p_i - p_j||_1 = |x_i - x_j| + |y_i - y_j|$ denotes the Manhattan distance between $p_i = (x_i, y_i)$ and $p_j = (x_j, y_j)$. We then define the neighborhood $\mathcal{N}_\delta(S) \subseteq 2^P$ of a solution S as follows;

$$\mathcal{N}_\delta(S) = \big\{\{S \cup \{p_j\} \setminus \{p_i\}\} \mid \forall p_i \in S,\ \forall p_j \in N_\delta(p_i)\big\}.$$

That is, $\mathcal{N}_\delta(S)$ is the family of the solutions which can be obtained by shifting a sampling site on $p_i \in S$ to a point in its neighborhood, $p_j \in N_\delta(p_i)$. The size of the neighborhood is evaluated as $|\mathcal{N}_\delta(S)| = O(|S|\delta^2)$. Clearly we have $|S| = |S'|$ for any $S' \in \mathcal{N}_\delta(S)$. This implies that, if $|S_{\text{init}}| = m$, then LS-SLP exactly returns a solution of size m. In Line 17 in LS-SLP, we search the solutions in $\mathcal{N}_\delta(S)$ in a random order, and employ the first met better solution as S' (which is called the *first admissible move strategy* in the literature).

We denote by $S^{(t)}$ $(t = 1, 2, \ldots, \tau)$ the solution obtained by the t-th LS-SLP, and by S_{opt} the *incumbent solution*. Initially, S_{opt} is set to $S^{(1)}$ (Line 4). For each $t > 1$, if $S^{(t)}$ is better

Algorithm 1 An ILS based algorithm for SLP

```
 1: procedure ILS-SLP($\tau, \delta$)
 2:     $S_{\mathrm{init}} \leftarrow$ a solution of size $m$
          ▷ $S_{\mathrm{init}}$ needs to satisfy Equation (3.3), but does not need to satisfy Equation (3.2).
 3:     $S^{(1)} \leftarrow$ LS-SLP($S_{\mathrm{init}}, \delta$)
 4:     $S_{\mathrm{opt}} \leftarrow S^{(1)}$
 5:     for $t \leftarrow 2, 3, \dots, \tau$ do
 6:         $S_{\mathrm{init}} \leftarrow$ a solution obtained by perturbing $S_{\mathrm{opt}}$
          ▷ $S_{\mathrm{init}}$ needs to satisfy Equation (3.3), but does not need to satisfy Equation (3.2).
 7:         $S^{(t)} \leftarrow$ LS-SLP($S_{\mathrm{init}}, \delta$)
 8:         if $\tilde{f}(S^{(t)}) > \tilde{f}(S_{\mathrm{opt}})$ then
 9:             $S_{\mathrm{opt}} \leftarrow S^{(t)}$
10:         end if
11:     end for
12:     output $S_{\mathrm{opt}}$
13: end procedure

14: procedure LS-SLP($S_{\mathrm{init}}, \delta$)
15:     $S \leftarrow S_{\mathrm{init}}$
16:     while $\mathcal{N}_\delta(S)$ contains a better solution than $S$ (in terms of $\tilde{f}$) do
17:         $S' \leftarrow$ a solution in $\mathcal{N}_\delta(S)$ with $\tilde{f}(S') > \tilde{f}(S)$
18:         $S \leftarrow S'$
19:     end while
20:     return $S$
21: end procedure
```

than the incumbent solution S_{opt}, then it is set $S_{\mathrm{opt}} \leftarrow S^{(t)}$ (Line 8). After LS-SLP is called τ times, the algorithm outputs S_{opt}.

Let us describe how to select an initial solution. For the 1st local search, we use randomly chosen m grid points as S_{init} (Line 2). For the subsequent ones, we obtain S_{init} by *perturbing* the incumbent solution S_{opt} (Line 6). We perturb S_{opt} by shifting *all* m sampling sites in S_{opt} to their neighborhoods. By this perturbation, we expect the algorithm to search solutions which are "close" to S_{opt}, the best solution searched so far, but are hardly reached by a single local search from S_{opt}; recall that $\mathcal{N}_\delta(S)$ is the family of solutions obtained by shifting only *one* sampling site in S to its neighborhood.

5. Computational Experiments

In this section, we present the experimental results of applying the algorithm ILS-SLP to the SLP instance constructed for Lake Biwa. We wrote all source codes in C language, and conduct all experiments by our PC carrying 2.83GHz CPU, which is not a special computer nowadays.

Setting $M = 100$ for the coefficient in Equation (4.1) and $\tau = 10$ and $\delta = 3$ for the parameters, we perform 50 trials of ILS-SLP by taking initial solutions at random. We compare the best solution S_{ILS} (in the sense of the objective f) with the existing solution S_{exist}. We take the number $m = |S_{\mathrm{ILS}}|$ of sampling sites as $m = 46$, in order to compare S_{ILS} with S_{exist} having 46 sampling sites. We confirmed that the constructed SLP instance has feasible solutions, including S_{exist}. We show function values of the two solutions in Table

Table 1: Function values of the existing solution S_{exist} and the best solution S_{ILS} of ILS-SLP

Solution	f (3.1)	φ_{est} (3.4)	φ_{exist} (3.6)	φ_{intake} (3.7)
S_{exist}	0.904	1.66×10^{-3}	46	4
S_{ILS}	0.832	1.23×10^{-3}	36	6

1, where we display not only the objectives but also $\varphi_{\text{est}}, \varphi_{\text{exist}}, \varphi_{\text{intake}}$, instead of the value functions $v_{\text{est}}, v_{\text{exist}}, v_{\text{intake}}$. We do not discuss the value functions $v_{\text{est}}, v_{\text{exist}}, v_{\text{intake}}$ but their component functions $\varphi_{\text{est}}, \varphi_{\text{exist}}, \varphi_{\text{intake}}$ here since the latter is easier to comprehend. Recall that the value function v_{est} is monotone non-increasing with respect to the squared error sum φ_{est}, and that the ones v_{exist} and v_{intake} are monotone increasing with respect to the numbers φ_{exist} and φ_{intake}, respectively. As shown in the table, S_{ILS} is superior to S_{exist} in the squared error sum φ_{est} and in the number φ_{intake} of intake points, but is inferior in the objective f and the number φ_{exist} of existing sites. This result seems agreeable because S_{exist} itself is a good solution in our SLP instance; the number $\varphi_{\text{exist}}(S)$ takes its maximum 46 if and only if $S = S_{\text{exist}}$ (see its definition in Equation (3.6)), and the weight w_{exist} given to the associated value function v_{exist} is larger than others ($w_{\text{est}} = 0.324$, $w_{\text{exist}} = 0.581$ and $w_{\text{intake}} = 0.095$).

We show the sampling sites in S_{exist} and S_{ILS} in Figure 3, along with overviews of the estimated distributions. We have $|S_{\text{ILS}} \setminus S_{\text{exist}}| = 10$, and observe that the 10 sampling sites in $S_{\text{ILS}} \setminus S_{\text{exist}}$ are located in the northern part of the lake, as shown in Figure 3(b). They may be used for decreasing the squared error sum φ_{est} or for increasing the number φ_{intake} of intake points. For example, S_{ILS} is better than S_{exist} in estimating the TP distribution of the area Ω. (Compare Figure 3 with the true distribution shown in Figure 1.) Also, S_{ILS} has 2 intake points which S_{exist} does not have. We will observe this phenomenon in a more general setting later.

We conjecture that ILS-SLP starts from such an initial solution that is “far from” S_{exist} with high probability, which is supported by the following observation: if we disregard the feasibility, the expectation of the number of existing sampling sites in an initial solution (which is selected at random) is just $46 \times (46/677) = 3.13$. Then the output local optima can be also distant from S_{exist}. In our experiments, the algorithm ILS-SLP does not output a better solution than S_{exist} in the sense of f, mainly due to the weights assigned to the value functions. However, observing the improvement on φ_{est} and φ_{intake}, we conclude that it can provide alternative solutions by searching a solution subspace distant from S_{exist}.

We have performed 50 trials of ILS-SLP, and in Figure 4, we show the number of trials in which each point $p_i \in P$ is contained in the output solutions. We observe that points in the area Ω are selected more frequently than others although they are not existing sampling sites. On the other hand, the 4 existing sampling sites in Ω' are selected fewer times. This phenomenon comes from the characteristics of SPT. As SPT is an interpolating method of 2D functions, to decrease the squared error sum φ_{est}, it must be more effective to locate sampling sites near critical points in the true distribution (e.g., Ω) than on its slopes (e.g., Ω'). (See Figure 1 for the true distribution.) Then the points in Ω can be regarded as potential sampling sites so long as we employ SPT to compute an estimated distribution.

Next, we take different numbers for m in order to observe the change of the tendency of the obtained solutions. Here we take $m = 20$, 30 and 40 in addition to $m = 46$, and perform 50 trials of ILS-SLP for each m. (For each m, we confirmed that the constructed

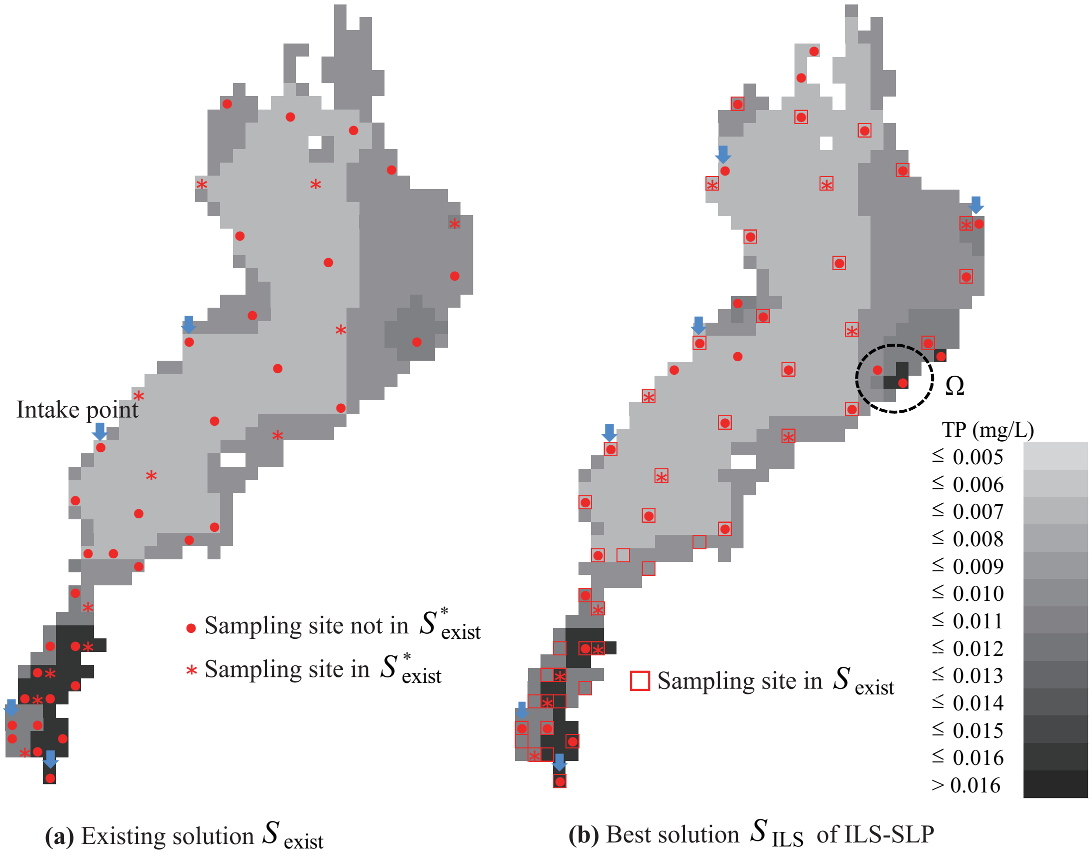

Intake point
● Sampling site not in S^*_{exist}
∗ Sampling site in S^*_{exist}
□ Sampling site in S_{exist}
Ω
TP (mg/L)
≤ 0.005
≤ 0.006
≤ 0.007
≤ 0.008
≤ 0.009
≤ 0.010
≤ 0.011
≤ 0.012
≤ 0.013
≤ 0.014
≤ 0.015
≤ 0.016
> 0.016
(a) Existing solution S_{exist}
(b) Best solution S_{ILS} of ILS-SLP

Figure 3: Sampling sites and the estimated distributions

SLP instance has feasible solutions.) We show the distributions of f, φ_{est}, φ_{exist}, φ_{intake} over the 50 trials in Figure 5. In each figure, the horizontal (resp., vertical) axis represents the function value (resp., the density). In the figures (a), (b) and (d), the vertical dotted line represents the function value of the existing solution S_{exist}; in the figure (c), the vertical dotted lines represent $\varphi_{\text{exist}}(S) = m$, which gives the upper limits on $\varphi_{\text{exist}}(S)$.

As shown in the figure (a), the existing solution S_{exist} achieves a better objective value $f(S_{\text{exist}}) = 0.904$ than all solutions obtained by ILS-SLP. As discussed before, the main reason is that the weight w_{exist} assigned to the value function v_{exist} is larger than others. Let us observe the figure (c). When $m = 20$ and 30, $\varphi_{\text{exist}}(S)$ is nearly m, meaning that an obtained solution S is almost a subset of S_{exist}. On the other hand, when $m = 40$ and 46, S contains about 5 to 10 sampling sites which are not in S_{exist}. We consider that these newly selected sites are used for improving φ_{est} and φ_{intake}, as shown in the figures (b) and (d). In the figure (b), it is interesting to see that S_{exist} is not superior even to the average case of $m = 30$.

Finally, let us mention the computation time of ILS-SLP. We show the averaged computation time over the 50 trials in Table 2. It is agreeable that the time is proportional to m and δ. We note that the solutions obtained for $\delta > 3$ are not much better than those for $\delta = 3$ (in both the objective and the value functions) in our preliminary experiments.

6. Discussion on Water Quality Estimation

To evaluate goodness of water quality estimation, one can choose other methods in place of SPT; e.g., *support vector regression* [15] and *kriging* [3]. The estimating method affects

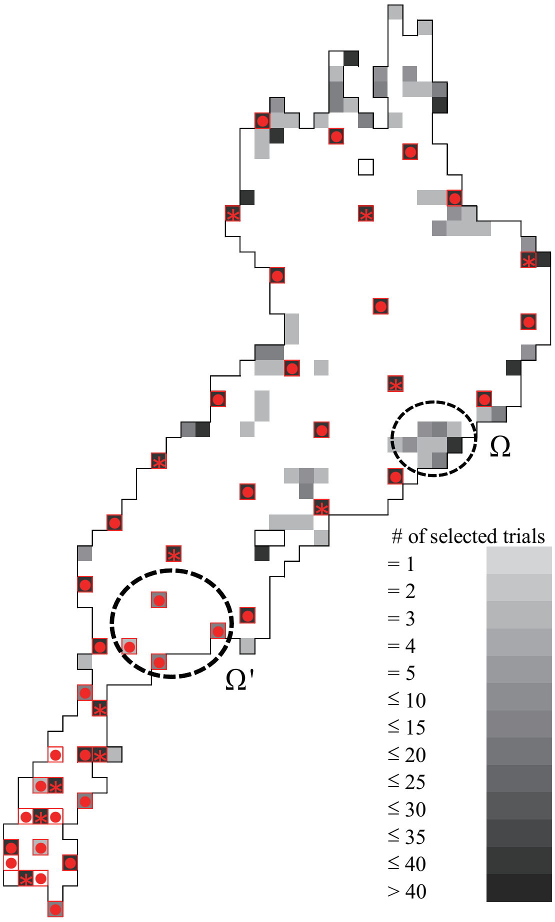


Figure 4: The number of trials in which each point is selected as the solutions by ILS-SLP

Table 2: Averaged computation time (sec.) of ILS-SLP ($\tau = 10$) over 50 trials

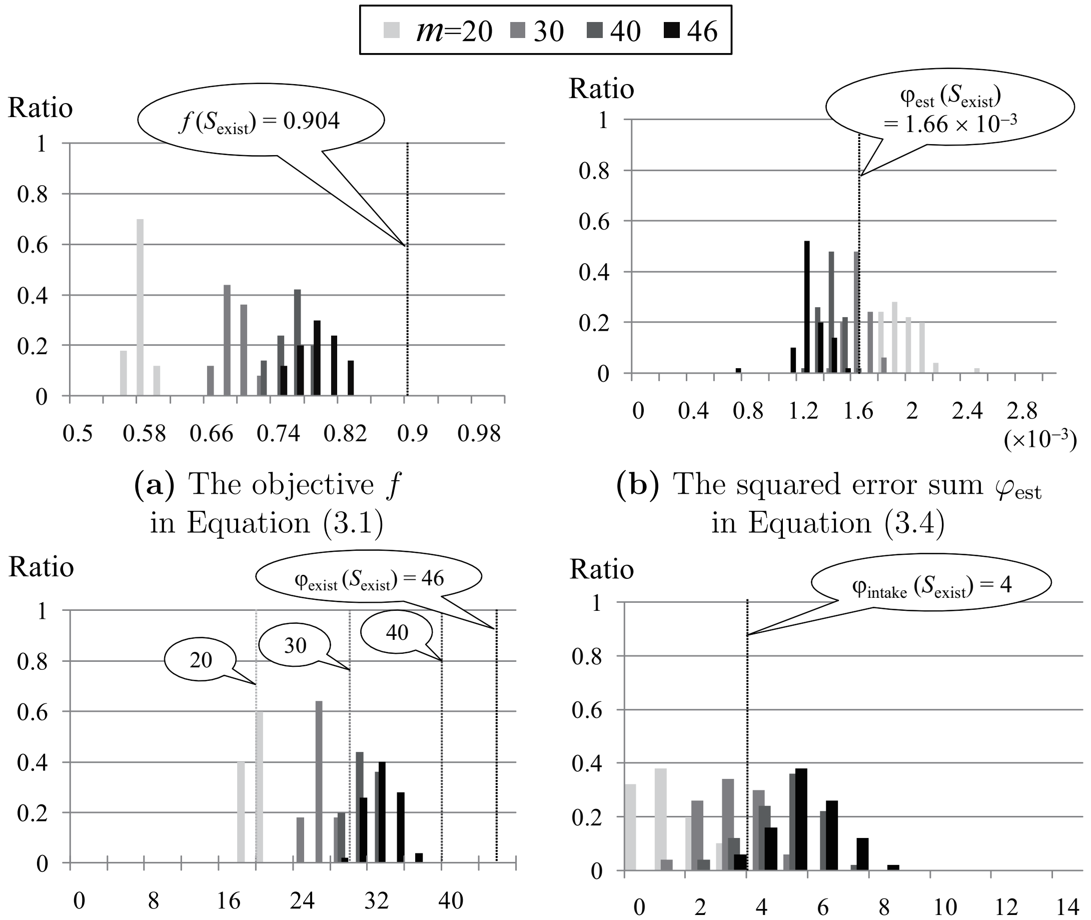
δ	$m = 20$	30	40	46
1	104.1	293.7	500.5	676.4
2	199.6	465.4	920.1	1348.4
3	301.3	655.6	1068.5	2113.3

not only the squared error sum $\varphi_{\text{est}}(S)$ but also which points are preferably selected in the solution.

In our preliminary research, we tried to estimate the true distribution based on set cover problem (SCP), which is not in the framework of iterated local search. The key idea is described as follows: Suppose that a solution $S \subseteq P$ is given. Expecting that the water quality of close points is similar, we consider estimating the water quality of neighbor points of a sampling site $p_i \in S$ by $d(p_i)$ (which we have assumed to be accessible).

For a point set P, a subset family $\mathcal{F} \subseteq 2^P$ and a cost function $c : \mathcal{F} \to \mathbb{R}$, SCP in general asks to compute such $\mathcal{F}' \subseteq \mathcal{F}$ that minimizes the total cost $\sum_{X \in \mathcal{F}'} c(X)$ among those covering P, i.e., $\bigcup_{X \in \mathcal{F}'} = P$. In our case, each subset $X \in \mathcal{F}$ is defined by a center point $p_i \in P$ and its neighbor points, and the cost $c(X)$ is defined as the squared error sum over X. For a feasible solution $\mathcal{F}'$ of SCP, we can obtain the location of sampling sites as the set of the center points in $\mathcal{F}'$.

Taking other value functions and constraint subsets into account, we formulated the



(a) The objective f in Equation (3.1)

(b) The squared error sum φ_{est} in Equation (3.4)

(c) The number φ_{exist} of existing sites in Equation (3.6)

(d) The number $\varphi_{\mathrm{intake}}$ of intakes in Equation (3.7)

Figure 5: Distribution of function values of the obtained solutions

problem of locating sampling sites as an extension of SCP. Solving the problem by IBM ILOG CPLEX 12.1 [9], we compared the obtained solutions with those computed by ILS-SLP, and found that they are competitive in both the objective and the value functions, but that sampling sites are located differently. The SCP approach tends to cover planes (i.e., the area having small difference in the true distribution) by one large subset in $\mathcal{F}$, and slopes by plural pieces of small subsets. This tendency is different from ILS-SLP with SPT, which prefers to locate sampling sites on critical points rather than slopes.

7. Concluding Remarks

In this paper, we mainly considered the monitoring task for Lake Biwa, and formulated the sampling site location problem (SLP) so that it can handle multiple purposes and constraints arising in the task. We designed an algorithm ILS-SLP to solve SLP, which is based on iterated local search (ILS). In the computational experiments, we observed that ILS-SLP can find such a solution by which we can make better estimation of the true distribution (generated by LB-Model) than the existing solution, even if the number of sampling sites is smaller. Also, some points are selected in the solution more frequently than others, which can be interpreted as potential sampling sites.

We should refine the proposed model to put it into practice, by taking other significant elements into account; e.g., time variation of the true distribution, water depths, other water quality indices. Our goal in this paper includes providing introduction of lake monitoring problems to OR researchers who are good at designing mathematical models and algorithms. It would be the authors' pleasure if some researchers are interested in our problems and even work on them. We will open the data used in this paper at some website in the near future.

Acknowledgements

We gratefully acknowledge very careful and detailed comments given by anonymous reviewers. We appreciate the cooperation of technical staffs in Shiga Prefectural Government who answered to our interview sincerely. We would also like to thank H. Hashimoto from Chuo University, M. Yagiura from Nagoya University, and N. Yamashita from Kyoto University for giving us useful comments.

References

[1] L.J. Alvarez-Vázquez, A. Martńez, M.E. Vázquez-Méndez, and A.M. Vilar: Optimal location of sampling points for river pollution control. *Mathematics and Computers in Simulation*, **71** (2006), 149–160.

[2] D.V. Chapman (ed.): *Water Quality Assessments: A Guide to the Use of Biota, Sediments and Water in Environmental Monitoring* (Chapman & Hall, 1992).

[3] J.P. Chiles and P. Delfiner: *Geostatistics: Modeling Spatial Uncertainty* (Wiley-Interscience, 1999).

[4] Department of Lake Biwa and Environment, Shiga Prefectural Government: *Strategy for Water Environmental Monitoring on Lake Biwa and Its Basin* (Shiga Prefectural Government, Japan, 2005) (in Japanese).

[5] W. Dixon, G.K. Smyth, and B. Chiswell: Optimized selection of river sampling sites. *Water Research*, **33** (1999), 971–978.

[6] M. Fujiwara, I. Sōmiya, H. Tsuno, and S. Fujii: Estimation of continuous distribution of water pollution concentration by spline technique and examination of rational allocation of monitoring stations. *Japan Journal of Water Pollution Research*, **8** (1985), 100–109 (in Japanese).

[7] T.F. Gonzalez (ed.): *Handbook of Approximation Algorithms and Metaheuristics* (Chapman & Hall/CRC, 2007).

[8] R.D. Hedger, P.M. Atkinson, and T.J. Malthus: Optimizing sampling strategies for estimating mean water quality in lakes using geostatistical techniques with remote sensing. *Lakes & Reservoirs: Research and Management*, **6** (2001), 279–288.

[9] IBM: *IBM ILOG CPLEX*. `http://www.ibm.com/` (accessed on October 1, 2010).

[10] International Lake Environment Committee: *Managing Lakes and Their Basins for Sustainable Use* (2005).
`http://www.ilec.or.jp/eg/pubs/index.html` (accessed on October 1, 2010).

[11] Y. Matsuoka and M. Naito: Estimation of two dimensional water quality profile by means of the interpolation methods. *Japan Journal of Water Pollution Research*, **7** (1984), 182–190 (in Japanese).

[12] Y. Oonishi: Surface interpolation by a spline technique. *Journal of the Oceanographical Society of Japan*, **31** (1975), 259–264 (in Japanese).

[13] T.G. Sanders, R.C. Ward, J.C. Loftis, T.D. Steele, D.D. Adrian, and V. Yevjevich: *Design of Networks for Monitoring Water Quality* (Water Resources Publication, 1983).

[14] Y. Sato, J. Kim, T. Takada, H. Nagare, E. Komatsu, T. Yuasa, and H. Uehara: Hydrological and material cycle simulation in Lake Biwa Basin coupling models about land, lake flow, and lake ecosystem. *Proceedings of the 12th World Lake Conference* (Jaipur, India, 2007), 819–823.

[15] B. Schölkopf. and A.J. Smola: *Learning with Kernels* (MIT Press, 2002).

[16] Shiga Prefectural Government: *Mother Lake 21 Plan -Lake Biwa Comprehensive Preservation and Improvement Project-* (2000).
`http://www.pref.shiga.jp/biwako/koai/mother21-e/index.html`
(accessed on October 1, 2010).

[17] V.V. Vazirani: *Approximation Algorithms* (Springer, 2001).

Kazuya Haraguchi
Faculty of Science and Engineering
Ishinomaki Senshu University
1, Shinmito, Minami-Sakai, Ishinomaki
Miyagi 986-8580, Japan
E-mail: kazuyah@isenshu-u.ac.jp