Methodology of shellfish surveys based on a microcomputer geographic information system

Dimitri A. Stolyarenko


Shellfish survey observations should be considered as spatially referenced, such that spatial mathematical models of stock density and sampling schemes are two major subjects of survey methodology. A mathematical method of spatial regression with supplementary variables is developed in order to integrate a few layers of spatial data. It allows incorporation of supplementary variables such as bathymetry, temperature, temperature gradient, current velocity, etc., into a model for survey data. The characteristic of spatial regression is the spatially indexed relationship among the variables. The approach permits a flexible sampling scheme. Spatial regression is a new powerful method for coupling layers of information in a geographic information system (GIS) for shellfishery applications. The microcomputer-based Spline Survey Designer Software System (SSDSS) was developed to implement this approach in an aquatic GIS. Employing SSDSS, rapid data analysis and mapping in real-time aboard a research vessel allows adaptive sampling during a survey as a means of improving the sampling scheme using currently available information. Two examples of surveys are given to emphasize characteristics of the methodology.

Introduction

The purpose of a shellfish survey is to assess a population or populations using observations which are only available from scattered locations, so the sampling scheme is of paramount importance in the survey methodology. Stock density can be sampled by trawl catch, acoustic measurement of concentration, or underwater TV. Major results of a survey are a stock estimate and a map of stock density. Area borders in the open sea are only approximations because, for instance, a depth range may vary with the oceanographic conditions for the current survey. In the case of fixed borders, the population may be commercially overexploited because of fishing outside the area studied. The sampling scheme and a mathematical model of stock density should answer the question for mapping: What is the density between observational points? The model is therefore the second major component of the survey methodology.

The modern concept of a survey treats observations as spatially referenced, with latitude and longitude as indices. The spatial approach was introduced in shellfish surveys in the mid-1980s for snow crab with kriging (Conan, 1985) and for shrimp with the spline technique (Stolyarenko, 1986). In conventional statistics a powerful method of randomization to eliminate the spatial index has been developed. However, the randomization leads to a rigid sampling scheme. In the case of a spatial problem it may conflict with another powerful statistical method of replication. It is easy to overestimate the stock if additional observations are carried out to study high concentrations which were not anticipated at the moment of design. Each replication of the greatest observation increases the value of the total average. But common sense suggests that high concentrations should be studied in detail because the precision of the largest observations often determines the precision of the total stock estimate. The spatial approach permits a flexible design scheme because it takes into account an observation which is associated with its position. In this case each replication increases the precision of the observation at a particular point.

Conventional stratified sampling may be based on supplementary information such as bathymetry. But this method erases the spatial index and changes it with the stratum number because location within the stratum is not taken into account. This property of the method can
be interpreted as the "step" model of stock density with a horizontal plane over a stratum and discontinuity at its borders. The greater the number of supplementary variables the stronger the temptation to create more strata by partitioning a small sample into smaller parts and then increasing the number of direct observations of the stock density. This is because stratification eliminates the relationship between observations of separated strata.

The modern way of increasing the accuracy of a survey is to integrate a few layers of relevant supplementary information rather than increase the number of direct observations. Therefore a new approach to integration of survey information has been developed. The section on spatial regression describes a mathematical method of spatial regression as a background to the integration. It allows the incorporation of supplementary variables such as bathymetry, temperature, current velocity, etc., as a means towards building an integrated model of the data. The characteristic of spatial regression is a spatially indexed relationship among variables. Spatial regression is a new and powerful method for coupling layers of information in a geographic information system (GIS). The microcomputer-based Spline Survey Designer Software System (SSDSS) was developed to implement this approach in an aquatic GIS (Stolyarenko, 1987, 1989). As a result of employing SSDSS, rapid data analysis and mapping in real-time aboard a research vessel allow adaptive sampling during a survey. The adaptiveness allows one to improve the sampling scheme during the survey using currently available information. SSDSS plays the role of a decision support system for positioning additional observations. Examples of application to shellfishery problems are presented in this article. The Discussion is devoted to generalizations and discusses relations with kriging.

Spatial regression

Conventional regression analysis often allows better prediction from observations using a supplementary independent variable than a sample average, i.e., arithmetic mean. The technique is based on spreading out the observations \( y_1, y_2, \ldots, y_n \) along the axis of the supplementary variable \( x \) and studying a relationship between the independent variable \( x \) and the dependent one \( y \) which is given with additive errors by the observations \( y_i = y(x_i) + e_i, i = 1, \ldots, n \). Of course, the approach employs extra information as values \( x = (x_1, x_2, \ldots, x_n)' \) for the observations. Thus the 0-dimensional space of the support is embedded in the one-dimensional one. The extra dimension is the supplementary variable \( x \). In order to predict the value of the dependent variable at a given point \( x^* \) of the axis \( x \), it is necessary to substitute

\[ y(x^*) = y^*(x^*) \]

the given value \( x^* \) to the function of the relationship \( y^* = y(x^*) \). The substitution returns the extended space back to the original dimension.

The same logical approach may be applied to studying a spatial problem. There are many characteristics that influence observations. The supplementary variables may allow us to explain anomalies of dependence of spatial observations. Many of them are much more precisely known than the function observed. Any random variation of such a supplementary variable is so small compared with the range of that variable observed that it is possible to ignore the random variation. It is a basic principle of regression (Draper and Smith, 1981). For example, bathymetry may be used as the supplementary independent variable for shellfishery survey data. Then catches are observations which measure stock density with the standard deviation of errors ca. 20% or greater, that is, the variation of the maximum catch is at least 20% of the range (Stolyarenko, 1986). In this case there is no reason for considering depth as subject to random variation.

However, a supplementary variable is usually a spatial one as well. The major idea of spatial regression is creation of a special space in which distance between two points reflects a potential relationship between the observations better than distance in the original space. The closer points are in the special space, the closer should be values of observations. The idea of incorporating the third dimension for one supplementary variable was suggested by the author with a technique based on three-dimensional spline approximation (Stolyarenko, 1987) and was applied to shrimp surveys (Stolyarenko and Ivanov, 1988).

Let us consider spatial regression in a two-dimensional physical space with one independent spatial supplementary variable. Generalizations to more dimensions for the physical space and for supplementary variables are considered in the Discussion. Let points of support \( (x_1, y_1), \ldots, (x_n, y_n) \) be in the two-dimensional physical space.

Using the independent supplementary variable, the area of interest of \( D \) may be embedded in Euclidean space of more dimensions \( D_{x+k} \). Distance between two points \( P_1 = (x_1, y_1, z_1) \) and \( P_2 = (x_2, y_2, z_2) \) in the multi-dimensional space is

\[ h(P_1, P_2) = [(x_1 - x_2)^2 + (y_1 - y_2)^2 + k(z_1 - z_2)^2]^{1/2}. \]

The coefficient \( k \) is analogous to a regression coefficient. Estimation of the ratio \( k \) for the supplementary variable may be treated as a problem of anisotropy. Thus, points with the same values of the supplementary variable keep the same distance as the two-dimensional ones. Distance between points with \( |z_1 - z_2| > 0 \) is increased. So, the smaller the difference of values of the supplementary variable for points with a given physical distance, the greater should be the interrelation of the observations at these points.
The three-dimensional spline is a function of the following form:

$$u(P^*) = \sum_{i=1}^{n} a_i h(P^*, P) + bx^* + cy^* + dz^* + e.$$  

The spatial variable $z = z(x,y)$ is usually assumed to be known over $D$ or its subset but not only $z^*$ pointwise. Then the substitution $u(x^*,y^*,z(x^*,y^*))$ provides the final estimate at the arbitrary point $(x^*,y^*)$ of the original space $D$. The function $z(x,y)$ may be a bathymetry map in applications and known precisely. Of course, it is not necessary to know the function $z = z(x,y)$ everywhere over $D$. But it is important to note that the prediction is possible only at the points $(x^*,y^*)$ where the value of the supplementary variable $z^*$ has been known or could be measured. Otherwise, the spatial relationship $u(x^*,y^*,z)$ as a function of $z$ is the result produced by spatial regression.

The geometric interpretation of the result function is a linear combination of figures of rotation or kernels plus a linear trend. For the three-dimensional case the function may be interpreted as density of a body. Let us consider a particular kernel. The geometric interpretation of the substitution is the following: a section with the surface $ku(x,y)$ through the reconstructed body in the space $DxkZ$ is projected to a plane. A surface of the equal influence of an observation is a sphere. If $z(x,y)$ is a one-dimensional function over $D$, the section of the sphere by a tangent plane at the point $(x^*,y^*)$ after the projection specifies an ellipse of anisotropy at the point $(x^*,y^*)$. Thus, the supplementary variable specifies anisotropy at every particular point of $D$, i.e., locally.

Cross-validation in the numerical form by Gu et al. (1989) may be applied to estimate the parameter $k$ together with the smoothing one. Observations are noisy. As in conventional regression, the observations are to be smoothed. In data-smoothing the objective may be to create a function of stock density which passes close to the observations. A basic problem is balancing the smoothness of the density reconstructed with fidelity to the observations. An interpolating function which passes through all the observations (i.e., overestimation of precision of measurements) is too unstable when replicated observations with noise are entered. At the other extreme, a linear function (i.e., underestimation of precision of observations) does not provide the desired fidelity to the observations. The balanced approach to smoothing consists in modelling the situation when the predicted value of the stock density at any point is compared with the real observation. The modelling is provided by deleting each observation in turn. The smoothness and the ratio $k$ are adjusted to minimize the standard deviation of the prediction error.

Aquatic GIS

A geographic information system (GIS) is a computer system designed to allow users to collect, manage, and analyse spatially referenced and associated attribute data. A GIS consists of a flexible user interface, database management controlled by application, spatial data manipulation and analysis, and map generation. With a powerful microcomputer a scientist is provided the opportunity to process and interrelate many more kinds of data than was previously feasible and to generate maps by analysing and modelling data in a variety of ways based on the application. The capability to interrelate layers of information is a basic feature of the GIS that differentiates it from other types of mapping systems, e.g., computer-aided design (CAD) or automated mapping. The relatively complete set of principal functions of a GIS was presented by Guptill and Nystrom (1985) (cited in Albert, 1988).

Most commercial GISs were developed as land information systems, so they operate with spatial objects that can be expressed by points, lines, areas, and sets of these primitives. The characteristic of data for the land information system is availability of lines and borders of areas in details due to remote sensing.

In contrast, an aquatic GIS has to operate mostly with continuous objects given only with sets of scattered points. Underwater borders are usually unknown and can be studied only pointwise. Therefore it is costly to get accurate data on borders if they exist; many layers vary temporarily and should be studied for each new survey. Even the most stable layer of bathymetry is to be considered as known pointwise for a detailed survey of a scallop bank. Modern multi-beam acoustic equipment allows reaching the same level of accuracy that is characteristic for a land information system. But currently it is rather the exception than the usual practice for other species. Therefore reconstruction of continuous objects on the basis of data given pointwise is a common request for aquatic applications. Points and the reconstructed continuous surfaces are main objects of the aquatic GIS.

Commercial systems place enormous emphasis on the cartographic aspects of spatial data and the creation of databases about objects, but they do not provide a wide set of operations for analysing these data. They usually allow data retrieval and analysis through Structured Query Language (SQL), overlay, and network routing. Only a few advanced systems provide methods for generating Digital Elevation Model (DEM) and interpolating continuous surfaces from point observations. The continuous surface is a final product of the system.
Figure 1. Bathymetry of the area studied off Spitsbergen as a supplementary variable. The bathymetry shows different directions of anisotropy. The small rectangles are trawl stations.

In contrast, an aquatic GIS should operate with objects given as continuous surfaces because the separate reconstruction is not very powerful operation, it is a spatial analogy with averaging. The continuous objects are to be interrelated.

Spatial regression provides the mathematical background of the new fundamental GIS operator for interrelation of the layers, which is characteristic for an aquatic GIS. Spatial regression introduces the operator connecting the upper layer given pointwise with the lower layers given by continuous objects. The result is a continuous object. Thus it allows reconstruction of the continuous surface from scattered points on the basis of supplementary layers of information given as continuous surfaces. The lower layers specify anisotropy of the space of the upper layer at every point of the area studied. They help spread information from the point to its neighbourhood and to solve the problem: What is density between the points? For instance, it is possible to use the bathymetry map $z(x,y)$ for reconstruction of the map of shrimp concentrations $u(x,y,z(x,y))$ from data of trawl stations $(x_1,y_1,z_1,u_1), \ldots, (x_n,y_n,z_n,u_n)$, where $x$ is latitude, $y$ is longitude, $z$ is depth, and $u$ is catch. The operator is an extension of Guptill and Nystrom's set.

So, the characteristic of the aquatic GIS is the following: it provides interrelation of continuous spatial objects produced from point observations. Spatial regression was developed as a method for the interrelation. The microcomputer-based Spline Survey Designer Software System (SSDSS) implements this aquatic GIS approach.

Adaptive sampling

Because the radial function for the three-dimensional spline and generally for spatial regression is linear, computational efficiency of the operation is high. As a result of employing SSDSS, rapid data analysis and mapping in real-time aboard a research vessel allow adaptive sampling during a survey.

Because accuracy of stock assessment (i.e. integral) is
Figure 2. Map of shrimp stock density as a result of spatial regression. The small rectangles are trawl stations. Designed by the Spline Survey Designer Software System.

Figure 3. Three-dimensional view of the shrimp stock density of Figure 2.
a primary goal in a shellfish survey, the theory of Monte Carlo experiments leads to the following ideal survey strategy. The richer a subarea is, the greater should be the portion of research efforts invested in studying the subarea. Otherwise, the number of stations should be proportional to the stock in the subarea of any size. Accuracy of a stock density estimate is less important for poor subareas because they contribute little to the stock estimate. However, it is only an ideal strategy for distributing research effort because the current stock distribution is unknown before a survey. Therefore approximate information from previous surveys should be used to produce a frame design before the survey. The frame design should consist of approximately 70% of the total number of stations if surveys for the area have been carried out for few years. It is based on previously accumulated knowledge about the stock distribution. The knowledge is formulated as an importance function which is a combination of the sample mean $\mu$ and standard deviation $\sigma$ of the average stock density at every point $P$ using the set of the surveys for the same season: $IF(P) = [\mu^2(P) + \sigma^2(P)]^{1/2}$. By fixing the area of interest and division by the integral of the importance function, it is easy to convert the importance function to the probability density. The density may be used for a random generator to produce points of the frame design. The above-described procedure provides some optimal properties of the integral estimate (Stolyarenko, 1987).

Adaptive sampling can adjust a sampling scheme to current stock distribution that is observed only after a survey starts. SSDSS allows one to improve the sampling scheme during a survey by using currently available information. SSDSS plays the role of a decision support system for positioning additional observations. A trawl station is costly; SSDSS allows one to test the hypothetical position of the next one or two stations to increase the increment of information value of each station.

GIS in shellfish surveys

Example 1
The first example is an application of spatial regression to 89 scattered observations of stock density in the trawl.
shrimp survey off Spitsbergen in 1987. The area of interest is two-dimensional. Bathymetry information is used as a supplementary variable. Depth is a valuable variable because it relates with water currents – which are directed along depth contours – and water temperature. Measurements of the last characteristic are fragmentary and variable when compared with stable and precise bathymetry data. The main reason for using a supplementary variable is to specify anisotropy locally over D. Therefore bathymetry is suitable. Note that the space with depth as a supplementary variable is not a real physical one. The depth has a specially adjusted scale due to the ratio coefficient $k$. In addition to depth, water temperature and its gradients may be considered as extra variables (Stolyarenko and Lobodenko, 1989).

Generally, regression does not evaluate reason and effect, only correlation between variables.

The area consists of the narrow ocean slope and relatively flat area to the south of Spitsbergen (Fig. 1); the southern fragment of the shrimp survey area off Spitsbergen is shown. The regression approach keeps approximately the same distances between observations over the flat area but increases distances to the observations in the slope. The area of interest is of the depth range from 150 to 800 m. It is interesting to note that the depth range was expanded during the survey. Before the survey it was pre-specified from 200 to 800 m for the standard stratification scheme of the area. However, because of specific oceanographic conditions the eastern concentration with the highest catch was located out of the pre-specified range (see the depth contour of 200 m in Figure 2). SSDSS aboard the research vessel assisted in modifying the sampling scheme and in delineating the concentration with adaptive design.

The resulting map of the shrimp data may be seen in Figure 2. The same result is presented by the three-dimensional view in Figure 3. The example emphasizes the feature of spatial regression to delineate the natural border of a population. It is important in fishery management to be able to estimate stock or biomass of a population but not density over a specified area because of shrimp mobility during a year. For instance, the great

Figure 5. Map of fish stock density as a result of spatial regression. The small rectangles are sites of acoustic measurements. The connections of high level measurements are made along depth contours. Designed by the Spline Survey Designer Software System.
central area of moderate catches is delineated with only a few western stations because they are at the slope and therefore of greater depth. Second, the southeastern concentration was not studied from all sides. In the three-dimensional space of the model the stations have neighbours with less value of depth. As a result, the concentration is delineated. This feature of spatial regression makes it a powerful tool for survey design because direct observations may be compensated with information on supplementary variables.

Example 2
At present, acoustics is applied for shrimp surveys as well. Another example is an acoustic survey. The author shows characteristics of the survey using fish data which were available. The survey consists of 986 measurements of fish concentration. Bathymetry is a supplementary variable. The depth range is 40 to 400 m. Figure 4 is the bathymetry of the area off Norway. Figure 5 presents the map of stock density. To connect or not to connect high level acoustic measurements of the neighbour tacks? The problem is solved using bathymetry for specifying anisotropy of the space: they are connected preferably along depth contours.

Discussion
Generalization of spatial regression to the three-dimensional physical space is easy using the distance of the analogous form $h = [(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 + k_2(w_1 - w_2)^2]^{1/2}$. The supplementary variable should be a function of three variables $w = w(x,y,z)$. Similar to conventional linear regression which should be linear only on parameters, spatial regression may be generalized to the arbitrary function of the supplementary variable. For example, if instead of $z$ the model uses the function $\sin(z)$, the distance should be $h = [(x_1 - x_2)^2 + (y_1 - y_2)^2 + k_2(\sin(z_1) - \sin(z_2))^2]^{1/2}$. It is the only change of the model for this case. This flexibility of modelling is based on the assumption of the independence of a supplementary variable.

The generalization for the case of many supplementary variables is easy as well. The distance should be $h = [(x_1 - x_2)^2 + (y_1 - y_2)^2 + k_2(z_1 - z_2)^2 + t_2(v_1 - v_2)^2 + \ldots]^{1/2}$ in the space $Dxk'Z$, where the space $Z$ is the one of the supplementary variables $z,y, \ldots$ Each supplementary variable is coupled with the coefficient, i.e., the anisotropy ratio that are $k,t, \ldots$. The type of the radial function may be the same as for the three-dimensional spline that is a straight line from the origin.

Spatial regression is formulated in terms of radial functions and is applicable to kriging (Matheron, 1971) as well.

Pure adaptive sampling is applicable for small areas when previous information is poor. In this case a frame design is developed only for the first four stations. Starting with the fifth station, SSDSS may be used for sequential positioning stations and for routing a research vessel. As a result, only 11 stations allowed mapping and assessing the Icelandic scallop bank (Stolyarenko et al., 1988). The interval between stations was 10 to 20 min, and only 2 min was spent making a decision on each next station position with trial of 2 to 3 hypothetical ones. The latest version of SSDSS provides real-time mapping with resolution up to one pixel of the graphics screen.

Acknowledgements
The author thanks Dr B. G. Ivanov for discussion of applications, and Dr K. G. Foote for permission to use the acoustic survey data set of Example 2.

References