

Statistical analysis of fisheries data using Systat

Introduction

Fishery management is broadly defined as the art and science (process) of producing sustained (annual) productivity of aquatic resources (usually fish), commensurate with the productive capacity of the environment, and according to established objectives set with consideration of user (constituent) needs. In short, it is the integration of the fish, habitat, and user dimensions of the resource to yield a given product or set of products.

Fishery management is based on imprecise estimation of population size, productivity, age structure and an incomplete knowledge of the dynamics of populations and their fisheries. The recent collapse of several fish populations world-wide and the ensuing social consequences for many fishers' demands that key sources of uncertainty in stock assessment must be identified.

Collection of basic data on catches, fishing effort, prices, values and other related information such as size at capture and length frequencies, constitute a key factor in a wide variety of applications. Sample-based fishery surveys that are conducted on a regular basis ought to be viewed not as an end in themselves but as an important source of fishery information of wide utility and scope.

It would seem appropriate to identify here a number of applications that depend directly on the availability of basic data resulting from fishery surveys. The list is not exhaustive but it essentially involves:

- (a) Estimated total production of fish, combined with data on imports and exports, constitutes the basis for calculating per capita consumption of fish, which is subsequently used in the formulation of food balance sheets;
- (b) Estimated total value of fish production is an important element in assessing the relative importance of the fishing industry within the national economy;
- (c) Prices at landing, combined with data on operational costs can provide indices of fleet performance;
- (d) Catch and fishing effort constitute the basic elements in the formulation of indices of abundance;
- (e) Catch rates by boat and gear categories, often combined with data on size at capture, permit a large number of analyses relating to gear selectivity and indices of exploitation;

- (f) Catch rates are often used for formulating indices of abundance for different fishing grounds;
- (g) Time series of prices are used in socio-economic studies;
- (h) Time series of fishing effort are indicative of declining or increasing trends of fisheries in districts and regions;
- (i) Trends regarding the human involvement in the fishing industry can be formulated.

Management studies differ from research

Managers manage according to conceptual relationships, and focus their assessment efforts on key attributes believed to be important indicators of the status of their resource. Therefore, management studies are likely to differ substantially from research studies. Rather than elucidating mechanisms, the manager is likely to assume that a fundamental model pertains to the system and focus on monitoring status of key components of the stock, environment or users. In doing so, the manager is apt to depend on indices and trends rather than getting absolute estimates (e.g., use of catch-per-unit effort of sampling gear to index stock size). Rarely will the manager rely upon measures associated with a single resource attribute, so studies commonly assess several variables concurrently, frequently during the same sampling endeavor. Internal consistencies among data sets, e.g., increases in condition indices corresponding to decreases in catch rates, are examined to reinforce confidence in the validity of individual study results.

Also in contrast to research, assessment by the manager is likely to be site- or resource-specific, of local rather than broad applicability. Unlike original research, these studies may be repetitive of other studies; in fact, agencies commonly establish standardized protocols for resource investigations to optimize comparisons across time and/or space.

However, in fisheries data modeling, the strength of research comes from the ability to develop and apply new and novel models to fisheries data, little of which is amenable to "standard" analyses. Often research includes estimating mixing proportions of mixed stock fisheries, modeling environmental effects on growth and abundance of fish using nonlinear and generalized linear mixed effects models, estimation of size-selectivity curves, and application of Bayesian state-space methodology for stock assessment and stock-recruitment modeling.

Fish behavior also plays a pivotal role in the understanding of variability in the results of resource and monitoring surveys and fishing operations, and a further role in the design of selective fishing gears to reduce bycatch and lessen ecosystem effects of fishing. Yet fish behavior studies are often qualitative and when quantified, rarely incorporated into stock assessment and forecasts of stock size. Modeling fish behavior includes mechanistic and descriptive approaches, individual-based and other simulation models of natural and fishing gear induced behavior. Usually numerical models which incorporate fish behavior into the estimation of resource abundance indices and the development of statistical models to handle large volumes of behavioral data from observational devices, such as data storage tags and other acoustics instruments are developed. Observation techniques and experimental designs are also used for the study of fish behavior, including sampling schemes, tools and methods for quantifying behavior, and the tools and methods used to observe fish in their natural environment, such as optical systems, passive and active acoustics, biotelemetry, and the development of new non-intrusive technologies.

The prediction of species distributions is of primary importance in ecology and conservation biology. This is also true for fish species distribution in particular. Statistical models play an important role in this regard. A comprehensive list includes both traditional and modern techniques for predicting species distributions, namely, logistic regression analysis, linear discriminant analysis, classification trees, etc.

However the idea of this note would not to get into the issues of management, research, behavior, and etc. with respect to fisheries and other related fields data analysis. The main purpose is to propagate the idea of statistical analysis of data as a useful means of communication while carrying out scientific investigations, building effective management strategies, quantification of fish behavior, and in different aspect of fisheries using Systat.

The schema for the rest of the note would be presenting some statistical methods required by personnel in the area of fisheries and other related areas. Each method as available would be followed by, if available, abstract of published research article where it was analyzed using Systat. No particular reference to fisheries management nor research or fish behavior, etc. is mentioned. Also, for example, if an example is used to illustrate a particular method, the same need not

be in a case study. The analytical methods do not follow any chronological sequence.

Analytical Methods

Developing an adequate design to an experiment is perhaps the trickiest and most difficult task that a fisheries biologist faces. Fisheries biologists must balance the need to control the experiment to better understand the results with the need to assure that the results do not stray far from what we would expect in nature. Many statistical techniques used by fisheries biologist to analyze the results of experiments come from disciplines like agriculture where experiments are easier to develop and manipulate. By necessity, fisheries biologists often rely on experimental units, such as lakes or fish, over which they have little control. Lack of control over experimental units is an important reason why developing a sound experimental design is critical to the success of any fisheries experiment.

Systat offers three methods for generating experimental designs: Classic DOE, the DOE Wizard, and the DESIGN command.

☛ Classic DOE provides a standard dialog interface for generating the most popular complete (full) and incomplete (fractional) factorial designs. Complete factorial designs can have two or three levels of each factor, with two-level designs limited to two to seven factors, and three-level designs limited to two to five factors. Incomplete designs include: Latin square designs with 3 to 12 levels per factor; selected two-level designs with 3 to 11 factors and from 4 to 128 runs; 13 of the most popular Taguchi designs; all of the Plackett and Burman two-level designs with 4 to 100 runs; the 6 three-, five-, and seven-level designs described by Plackett and Burman; and the set of 10 three-level designs described by Box and Behnken in both their blocked and unblocked versions. In addition, the Lattice, Centroid, Axial, and Screening mixture designs can be generated.

☛ The DOE Wizard provides an alternative interface consisting of a series of questions defining the structure of the design. The wizard offers more designs than Classic DOE, including response surface and optimal designs. Optimization methods include the Fedorov, k-exchange, and coordinate exchange algorithms with three optimally criteria available. The coordinate exchange algorithms accommodate both continuous and categorical variables. The search algorithms for fractional factorial designs allow any number

of levels for any factor and search for orthogonal, incomplete blocks if requested.

🔗 The **DESIGN** command generates all designs found in Classic DOE using Systat's command language.

Now that we have an experiment designed properly, with random pond selection and reasonable interspersions of treatments, for example, the hatchery manager wishes to determine how growth of muskellunge *Esox masquinongy* is affected by differences in stocking densities. Assume that the standard stocking density into hatchery ponds is 20 fish/acre, but the manager wishes to see if growth is reduced at two other stocking densities, one higher (30 fish/acre) and one lower (10 fish/acre) than the standard density. The hatchery manager applies the higher density to four ponds, the lower density to four ponds, and the standard density to four ponds (as controls). The hatchery manager measures weight and length of a sample of fish being stocked into each pond before the experiment begins and after the experiment ends, and uses the difference in mean weight before and after the experiment as a measure of growth during the experiment.

How does the hatchery manager determine if the differences in stocking density caused differences in mean weight? One tool available to the manager is the **general linear model**. General linear model is a term used to refer to an entire class of models that are linear in their parameters. In most of these models, we measure a response variable and then determine how response variables are influenced by predictor variables. For clarification, this means that no parameter in the model is an exponent or multiplied/divided by another parameter. The term general is used, because both continuous and categorical variables can be used as predictor variables.

General Linear Model (GLM) in Systat can estimate and test any univariate or multivariate general linear model, including those for multiple regression, analysis of variance or covariance, and other procedures such as discriminant analysis and principal components. With the general linear model, one can explore randomized block designs, incomplete block designs, fractional factorial designs, Latin square designs, split plot designs, crossover designs, nesting, and more. The model is:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{e}$$

where **Y** is a vector or matrix of dependent variables, **X** is a vector or matrix of independent variables, **B** is a vector or matrix of regression coefficients, and **e** is a vector or matrix of random errors.

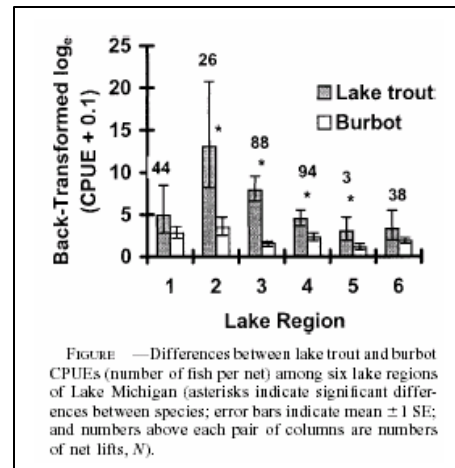
Once the parameters of a model have been estimated, they can be tested by any general linear hypothesis of the following form:

$$\mathbf{ABC}' = \mathbf{D}$$

where **A** is a matrix of linear weights on coefficients across the independent variables (the rows of **B**), **C** is a matrix of linear weights on the coefficients across dependent variables (the columns of **B**), **B** is the matrix of regression coefficients or effects, and **D** is a null hypothesis matrix (usually a null matrix). For the multivariate models described here, the **C** matrix is an identity matrix and the **D** matrix is null. The **A** matrix can have several different forms, but these are all submatrices of an identity matrix, and are easily formed using **HYPOTHESIS**.

Case Study:

Camille Ward et al. in their study [Relative Abundance of Lake Trout and Burbot in the Main Basin of Lake Michigan in the Early 1930s, Transactions of the American Fisheries Society 129:282–295, 2000] used Systat's general linear models to compare and to assess the sources of variation in CPUEs (catches per unit effort) of lake trout *Salvelinus namaycush* and burbot *Lota lota* netted in small-mesh gill nets from the main basin of Lake Michigan in the early 1930s, before populations of both species collapsed.



Fish abundance measures that are estimated repeatedly through time are typically examined for trends or patterns of change through time. The relation between observations close in time may be correlated – that is, values in a given year may be similar to values in the previous year and this similarity generally decreases with increasing time intervals. Such data are said to exhibit positive autocorrelation. The presence of autocorrelation (or serial dependence) in fish abundance data compromises statistical interpretation of correlation and regression analyses that may be undertaken to relate changes in fish abundance to environmental or biological variables. The reason is that most parametric statistical tests assume independence (correlation equals 0) of observations. Hypothesis tests on autocorrelated data require adjustments to the degrees of freedom to reflect the lack of independence among observations.

For predictive modeling or exploratory analyses, autocorrelated data must be transformed. Several transformations have been used with autocorrelated data including prewhitening (sometimes termed first-differencing), detrending, and smoothing. In fisheries, many time series exhibit low-frequency variation typical of slow, long-term changes in abundance, and prewhitening or first-differencing of the CPUE series often removes this variation. High-frequency variation is evidenced by changes occurring over short time scales (rapid changes through time) and can be removed using smoothing. An example of high-frequency variation is measurement error. It should be noted, however, that the decision to employ any of the transformations should be taken with extreme caution.

Systat's Time Series (SERIES) implements a wide variety of time series models, including linear and nonlinear filtering, Fourier analysis, seasonal decomposition, nonseasonal and seasonal exponential smoothing, and the Box-Jenkins approach to nonseasonal and seasonal ARIMA.

Case Study:

Norman D. Yan AND Trevor W. Pawson in their study [Changes in the crustacean zooplankton community of Harp Lake, Canada, following invasion by *Bythotrephes cederstroemi*, *Freshwater Biology* (1997) 37, 409-425] used Systat's SERIES to examine temporal autocorrelation functions for the logarithmically transformed abundance of each of the eighteen most common zooplankton taxa at annual time steps for the entire time series. Significant firstorder temporal autocorrelations were only observed for *Bosmina longirostris* and *Chydorus sphaericus* (O.F.M.). None of the higher order autocorrelations were significant.

Size structure (length frequency) analysis is one of the most commonly used fisheries assessment tools. The size structure of a fish population at any point in time can be considered a "snapshot" that reflects the interactions of the dynamic rates of recruitment, growth, and mortality. Thus, length frequency data are valuable tools to gain insight into the dynamics of fish populations, and to identify problems such as inconsistent year class strength, slow growth, and excessive mortality. In most cases, other population assessment tools, such as catch-per-unit-effort, age and growth analysis, recruitment analysis, mortality, and body condition, complement a thorough interpretation of size structure data.

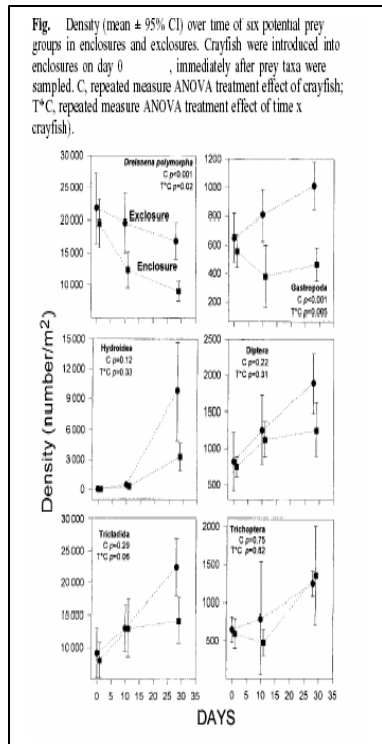
Fisheries researchers use several techniques to analyze size structure data. In the simplest case, a fisheries biologist might construct a length frequency histogram, or calculate a size structure index, and make an assessment based on a qualitative evaluation of the size structure characteristics. Oftentimes, the primary objective is to compare size structures among samples. In these cases, a fisheries biologist may be interested in answering several questions. For example, does the size structure of white crappie populations differ among water bodies? Did the size structure of a largemouth bass population change over time in response to a management action? Are the size structures obtained from an alewife population different between two or more sampling gears? What factors influence the size structure of walleye populations? There are many statistical methods available to researchers wishing to analyze structure data. Among them is the repeated measures ANOVA.

Systat handles a wide variety of balanced and unbalanced analysis of variance designs. The Analysis of Variance (ANOVA) procedure includes all interactions in the model and tests them automatically; it also provides analysis of covariance, and repeated measures designs. After you have estimated your ANOVA model, it is easy to test post hoc pairwise differences in means or to test any contrast across cell means, including simple effects.

In a repeated measures design, the same variable is measured several times for each subject (case). A paired-comparison t test is the most simple form of a repeated measures design (for example, each subject has a before and after measure).

SYSTAT derives values from your repeated measures and uses them in analysis of variance computations to test changes across the repeated measures (within subjects) as well as differences between groups of subjects (between subjects). Tests of the within-subjects values are called polynomial test of order 1, 2, ..., up to k , where k is one less than the number of repeated measures. The first polynomial is used to test linear changes: do the repeated responses increase (or decrease) around a line with a significant slope? The second polynomial tests whether the responses fall along a quadratic curve, and so on.

Case Study:



William L. Perry et al. in their study [Impact of crayfish predation on exotic zebra mussels and native invertebrates in a lake-outlet stream, Can. J. Fish. Aquat. Sci. 54: 120-125 (1997)] to test the effect crayfish had on the density of zebra mussels and cooccurring invertebrates. That is, the objective was to determine how crayfish affected the population density and size structure of zebra mussels and the abundance of associated invertebrates in a lake-outflow stream using a cage experiment. They used **Systat's Repeated Measures ANOVA** to compare the density of each of six dominant taxa between the enclosure and exclosure cages. Overall, crayfish reduced all size-classes of zebra mussels in the enclosures relative to the exclosures, including the largest size-class that laboratory results suggested were not eaten.

Mortality is a concept that describes the rate at which individuals are lost from a population. This concept is central to understanding the ecology of populations, particularly their dynamics, and is essential to managing fish stocks. Each species is adapted to its own mortality patterns, with its own distribution over life stages and age groups. Mass mortality is common at the egg or larval stages, largely due to abiotic conditions, but the lethal effects of abiotic conditions usually become minor when the larvae become mobile. In the early stages of external feeding, a limited food supply may directly influence mortality. If the fish survives, the limited food supply becomes an indirect source of mortality via retarded growth and lengthening of the time searching for food, which makes the fish more vulnerable to predation. Later in life, fishing may be an important source of mortality, and exponentially decreasing patterns suggest mortality is fairly constant from adulthood onwards. Knowledge about the patterns and causes of death helps understand inter- and intra-specific interactions, and interactions between the population and its abiotic environment.

Length-based models do not use direct age estimates; instead, they use the L (asymptotic length) and K (rate at which L is approached)

parameters from the von Bertalanffy growth equation to convert length to age. Like catch-curve models, assumptions of length-based models include

- (1) constant recruitment within the period covered by the length distribution, or at least have recruitment has varied in a random fashion,
- (2) Z is constant over ages,
- (3) only lengths fully recruited to the gear are included (equivalent to the descending portion of a catch curve),
- (4) growth is constant and adequately described by the von Bertalanffy model, and
- (5) the sampling gear adequately represents the standing length distribution.

Another assumption made by length-based models is that recruitment into the smallest length considered for analysis is constant through time each year, so that the shape of the length distribution and mean length does not vary seasonally. This assumption is violated in populations that exhibit seasonal instead of continuous recruitment, but may be avoided by taking multiple samples within the year and pooling them before analysis, or by limiting analysis to longer (i.e., older) fish for which length-at-age is generally highly variable and recruitment spread out. Given these stringent assumptions, length-based estimates should be used when only a rough approximation will do, or there is no better option. Although mortality can be estimated from a sample that only provides lengths of individual fish, these models usually require some minimum information about the relation between length and age. Nonlinear regression is used to analyze length-based models.

Nonlinear modeling estimates parameters for a variety of nonlinear models using a Gauss-Newton (Systat computes exact derivatives), Quasi-Newton, or Simplex algorithm. In addition, one can specify a loss function other than least squares, so maximum likelihood estimates can be computed. One can set lower and upper limits on individual parameters. When the parameters are highly intercorrelated, and there is concern about overfitting, one can fix the value of one or more parameters, and Nonlinear Model will test the result against the full model. If the estimates have trouble converging, or if they converge to a local minimum, Marquarding is available.

For assessing the certainty of the parameter estimates, Nonlinear Model offers Wald confidence regions and Cook-Weisberg graphical

confidence curves. The latter are useful when it is unreasonable to assume that the estimates follow a normal distribution.

When response contains outliers, one may want to downweight their residuals using one of Nonlinear Models's robust ψ functions: median, Huber, Hampel, bisquare, t , trim, or the p th power of the absolute value of the residuals.

Case Study:

Casimiro Quiñonez-Velázquez in the study [Age validation and growth of larval and juvenile haddock, *Melanogrammus aeglefinus*, and pollock, *Pollachius virens*, on the Scotian Shelf, Fish. Bull. 97:306–319 (1999)] concluded that, pelagic and early demersal growth appear to represent distinct stanzas in the growth history of these gadoids. A Laird-Gompertz growth model curve was fitted to the length-at-age data for each species in each year. This model has been shown to provide an adequate fit for length-at-age data on age 0+ fish of many different species. The parameters were derived by nonlinear least squares tests by using Systat's NONLIN.

The analysis of fish condition has become a standard practice in the management of fish populations, as a measure of both individual and cohort (e.g., age or size group) fitness. Condition has been generically described as the well-being or robustness of an individual fish and has typically been estimated by comparing individual fish weight of a given length to a standard weight and assuming heavier fish reflect a healthier physiological state (condition index, or directly measuring physiological parameters related to the energy stores such as tissue lipid content).

Condition indices are widely used to assess many facets of fish populations, including the general health of fish stocks, the effects of management actions, community structure, or environmental influences. Condition indices are intended to indirectly estimate physiological condition (e.g., lipid stores) based on the premise that a fish of a given species and length should weigh as much as a standard for its length, and variations from the standard are taken as an indication of the relative fitness of an individual. Measures of fish condition based on a standard weight have been available since the early 1900's and have since undergone an evolution in methodology, as well as a rigorous review regarding their correlation with physiological parameters and statistical merit. However, they have

remained popular tools because they are simplistic and non-invasive. Linear regression is used to analyze condition indices.

The model for simple linear regression is:

$$y = \beta_0 + \beta_1 X + \varepsilon$$

where y is the dependent variable, x is the independent variable, and the β 's are the regression parameters (the intercept and the slope of the line of best fit). The model for multiple linear regression is:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

Both Regression and General Linear Model can estimate and test simple and multiple linear regression models. Regression is easier to use than General Linear Model when you are doing simple regression, multiple regression, or stepwise regression because it has fewer options. To include interaction terms in your model or for mixture models, use General Linear Model. With Regression, all independent variables must be continuous; in General Linear Model, you can identify categorical independent variables and SYSTAT will generate a set of design variables for each. Both General Linear Model and Regression allow you to save residuals. In addition, you can test a variety of hypotheses concerning the regression coefficients using General Linear Model.

The ability to do stepwise regression is available in three ways: use the default values, specify your own selection criteria, or at each step, interactively select a variable to add or remove from the model.

For each model you fit in REGRESS, SYSTAT reports R^2 , adjusted R^2 , the standard error of the estimate, and an ANOVA table for assessing the fit of the model. For each variable in the model, the output includes the estimate of the regression coefficient, the standard error of the coefficient, the standardized coefficient, tolerance, and a t statistic for measuring the usefulness of the variable in the model.

Case Study:

Nicholas J. Gotelli¹ and Christopher M. Taylor in their study [Testing macroecology models with stream-fish assemblages, *Evolutionary Ecology Research*, 1999, 1: 847–858] measured species-level probabilities of colonization and extinction from a decade (1976–86) of stream-fish censuses at 10 sites on the Cimarron River, Oklahoma. Overall, their results confirm that species-level traits are correlated with standardized estimates of extinction and colonization probabilities within large assemblages of species. Because condition indices

indicated little collinearity among macroecological variables, entire multiple regression models were computed using Systat's REGRESS, rather than relying on stepwise procedures.

Multivariate statistical methods provide a useful tool for analysis of complex data set sets because many variables can be integrated in one analysis. Complex ecological concepts can be better understood using multivariate statistical methods. Statistical analyses have been an increasing part of the fisheries literature, and increased statistical knowledge and computing power suggest that the use of multivariate methods may have increased as well.

There are two general ways multivariate methods are used: descriptive (exploratory) and inferential (confirmatory). In descriptive uses, analyses combine variables in some "optimal" way, whereas inferential uses test a priori hypotheses. In general, multivariate methods are most often used to describe data relationships and not for hypothesis testing.

Principal components analysis is an ordination technique, which breaks down or partitions a resemblance matrix (variance-covariance or correlation) into a set of orthogonal (perpendicular) axes or PCA "components". The first few PCA components will explain the largest percentage of variation in the data set, and ordinations of sampling units on these axes provide information about the ecological relationships between them.

Systat's Factor analysis provides principal components analysis and common factor analysis (maximum likelihood and iterated principal axis). Systat has options to rotate, sort, plot and save factor loadings. With the principal components method, one can also save the scores and coefficients. Orthogonal methods of rotation include varimax, equamax, quartimax and orthomax. A direct oblimin method is also available for oblique rotation. Users can explore other rotations by interactively rotating a 3-D Quick Graph plot of the factor loadings.

Case Study:

Robert M. Goldstein et al. in their study [Development of a stream habitat index for use with an Index of Biotic Integrity in the St. Croix River Basin, Minnesota, Water-Resources Investigations Report 99-4920.] developed a habitat index for use to evaluate water quality and the effects of nonpoint-source effects not associated with habitat degradation. Core habitat variables were determined with a

concurrency analysis using Systat's Principal Components of two subsets of sites with pristine or least affected habitat. Although core habitat variables differed slightly between data sets, sufficient similarities allowed development of an index. The index (the sum of pluses or minuses dependent on the variable's correlation to biotic integrity), composed of 12 core habitat variables in 5 classification groups (hydrology, geomorphology, substrate, instream habitat, and riparian/land use), was able to distinguish sites with low Index of Biotic Integrity scores not related to habitat degradation.

Cluster analysis is a technique that sorts objects (such as sampling units) into groups or clusters based upon their overall resemblance to one another. There are several cluster analysis methods, including single-linkage, average-linkage, and complete-linkage, all having different ways in which clusters are formed; and although most of the cluster analysis strategies give similar results, several methods should be explored and the results compared. Clusters can then be determined from one of the cluster analysis methods based on the "underlying ecological knowledge of the data". Interpretation of cluster analysis results can often be highly subjective.

Systat provides a variety of cluster analysis methods on rectangular or symmetric data matrices. Cluster provides three procedures for clustering: Hierarchical Clustering, K-means and Additive Trees. The hierarchical clustering procedure splits a set of objects into a selected number of groups by maximizing between-cluster variation and minimizing within-cluster variation. The additive trees clustering procedure produces a Sattath-Tversky additive tree clustering.

Hierarchical clustering clusters cases, variables or both cases and variables simultaneously; K-means clusters cases only; and Additive trees clusters a similarity or dissimilarity matrix. Eight distance metrics are available with hierarchical clustering and k-means, including metrics for quantitative and frequency count data. Hierarchical clustering has six methods for linking clusters and displays the results as a tree (dendrogram) or a polar dendrogram. When the MATRIX option is used to cluster cases and variables, Systat uses a gray-scale or color spectrum to represent the values.

area's social, economic, and housing characteristics as reported in the 1990 census, including household type and size, ethnic composition, level of education completed, labor force participation, housing costs, and extensive income and poverty indicators. You may also request other specific data series, if available. The software does not provide geo-coding (the ability to locate a point based on an address) at this time. Therefore, to map data you have collected, you need to know either the area (for example, the census tract) within which each of your data points falls or the longitude/latitude coordinates of the area you want.

In addition, you can import ArcView files and saves the result as .SYD and .SMP files.

To plot a map, the software selects an ID for a polygon by reading the MAPNUM variable in the first record of the data file. Next, the .SMP file is read until a matching ID variable absolute value is found. The subsequent NP points of the polygon corresponding to that ID are plotted and, possibly, filled with color and fill pattern and labeled. Next, the .SMP file is read further to check for remaining ID's that match the current MAPNUM. Any polygons with ID's corresponding in absolute value to the current value of MAPNUM are similarly plotted. If an ID is negative and its absolute value corresponds to the current value of MAPNUM, its associated polygon is plotted but not filled. When no further ID's corresponding to MAPNUM's value are found, the next appropriate record in the data file is selected. Polygon(s) for this ID are plotted, and the process continues until no further appropriate records are available in the data file. The MINLAT, MAXLAT, MINLON, and MAXLON variables are used for scaling maps in the viewing window, and the LABLAT and LABLON variables are used for labeling polygons.

Case Study:

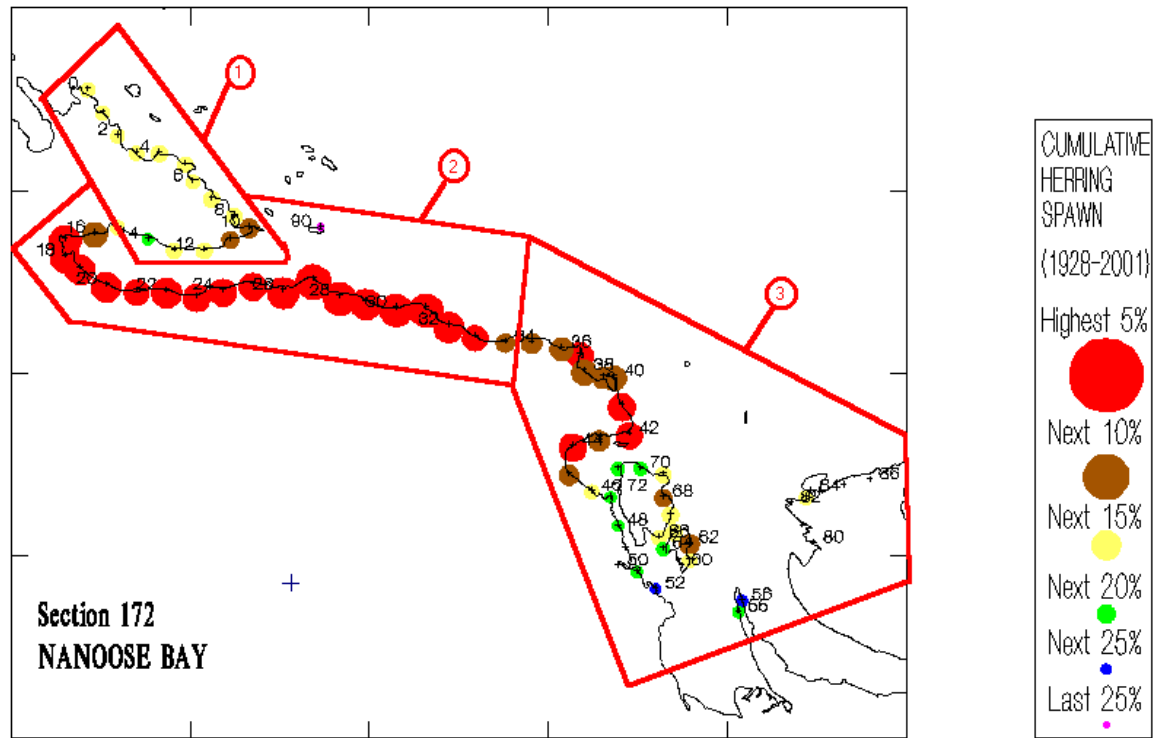
D. E. Hay and P. B. McCarter studied [Herring Spawning Areas of British Columbia: A review, geographic analysis and classification, Revised edition: 2003, Original edition: Hay, D.E., P.B. McCarter, R. Kronlund and C. Roy. 1989. Spawning areas of British Columbia herring: a review, geographical analysis and classification. Volumes 1-6. Can. MS Rep. fish. Aquat. Sci. 2019.] {Web Link: <http://www.pac.dfo-mpo.gc.ca/sci/herring/herspawn/project.htm>} the geographical distributions of Pacific herring (*Clupea pallasii*) spawning sites which have been estimated each year since 1928. The analysis was based on approximately 27,000 spawning events recorded mostly

by fishery officers and diver teams in six regions of the British Columbia coast. For each of about 100 geographical sub-areas, time series maps were constructed to delineate annual herring spawn deposition along each kilometre of shoreline from 1930 to the year 2001. Cumulative spawn deposition (since 1928) was also mapped using proportionately sized, color, bubble plots which rank and classify each kilometre of herring spawning habitat according to the long-term frequency and magnitude of spawns. Approximately 5,200 km of British Columbia's 29,500 km coastline have been ranked and classified as herring spawning habitat.

Spawn deposition polygons were also digitized on 1:20,000 scale, TRIM (Terrain Resource Inventory Management) maps for the years 1930 to 1997 and are shown on a set of 111 sub-area maps. A series of plots and histograms summarize the magnitude, frequency and timing of spawning over all years and distinguish shoreline kilometres with highly repetitive spawning activity from those with less frequent activity. A statistical summary of herring spawn and catch records, including mean, minimum and maximum spawning dates, is presented in a series of tables for each herring sub-area (section), statistical area, region and the entire British Columbia coast. The accuracy and completeness of records is discussed.

Systat Map was used to create cumulative spawn maps, times series plots, graphs and image files suitable for electronic publication. Cumulative Spawn Habitat Maps (shown below) are based on annual, herring spawn survey measurements. Cumulative spawn deposition (summed over all survey years: 1928 to the present) has been estimated for each kilometre of British Columbia (BC) coastline using the Spawn Habitat Index listed in the Cumulative Spawn Table. Cumulative spawn (since 1928) is depicted at each km position by the proportional size of each circle. The circles are colored to represent six classifications of long-term cumulative spawn. Red indicates the top 5%, brown the next 10%, yellow the next 15%, green the next 20%, blue the next 25% and violet the last 25% of ranked shoreline kilometres.

Herring Spawn Classification	Legend Color	Percentage of Ranked Km	Percent Range
VITAL	Red	top 5 %	95 - 100 %
MAJOR	Brown	next 10 %	85 - 95 %
HIGH	Yellow	next 15 %	70 - 85 %
MEDIUM	Green	next 20 %	50 - 70 %
LOW	Blue	next 25 %	25 - 50%
MINOR	Violet	last 25 %	0 - 25 %

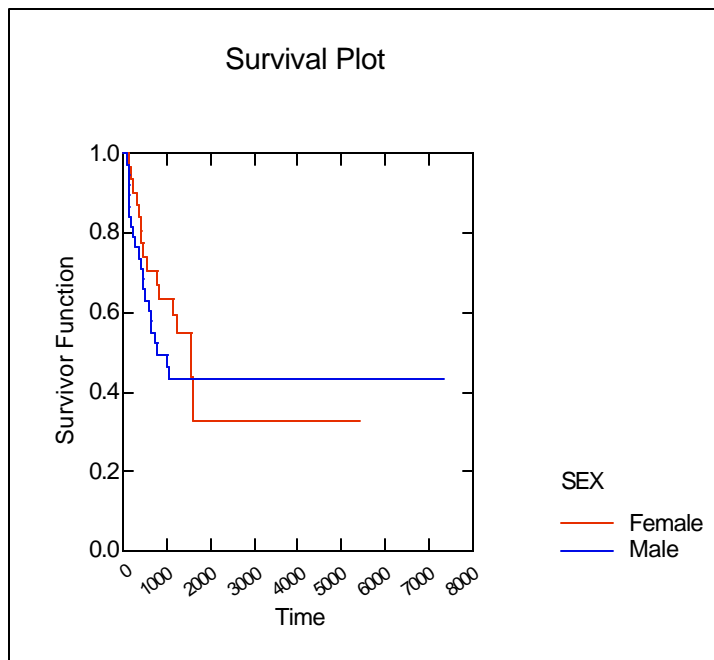


Summary

The above paragraphs just mention a bird's eye view of methods available in Systat. But Systat provides a powerful statistical and graphical analysis system in a graphical environment using descriptive menus and simple dialog boxes. The bouquet of data analytical methods runs from simple to advanced such as: Interpreting Data Distributions - Graphical Techniques, Measures of Central Tendency, Measures of Dispersion, Exploratory Data Analysis, Point Estimation and Confidence Intervals, From Estimation to Hypothesis Testing, Data Analysis With Categorical Variables, Correlations, Nonparametric or Distribution-Free Statistical Tests, Logit and Probit Models, Spatial Data Analysis, Classification and Regression Trees, Partial Order Scalogram Analysis with Coordinates and many more.

Simply pointing and clicking the mouse can accomplish most tasks. Systat's command language provides functionality not available in the dialog box interface. Matrix procedure allows you to use matrix algebra to specify statistical analyses and perform data management tasks. Systat can open data files saved in various formats.

Appendix – Survival Analysis



Systat's SURVIVAL can be used to explore grouped, right-censored, and interval-censored survival data and to estimate nonparametric, partially parametric, and fully parametric models by maximum likelihood. The SURVIVAL module's ability to handle disjoint and overlapping interval-censored data and combinations of interval censoring,

right censoring, and exact failure times is a major enhancement over other programs.

The facilities provided in SURVIVAL include the Kaplan-Meier estimator, Turnbull's generalization of the Kaplan-Meier estimator for interval-censored data, plots of failure and censoring times, quantile plots for standardized reference distributions, log-rank tests, the proportional hazards (Cox) regression, and the Weibull, log-normal, logistic, and exponential regression models. All models can be estimated with or without covariates, either directly or by stepwise regression procedures. The Kaplan-Meier estimator, quantile plots, and Cox regression all permit stratification. The survivor function, hazard function, reliabilities, and quantiles can be generated from parametric models for specific covariate values, and the baseline hazards can be derived from the Cox and stratified Cox models.

The results of most analytic techniques can be saved into SYSTAT files for further manipulation and analysis with other SYSTAT modules. Post hoc analyses, such as plotting survivor functions, computing life tables from a model, and requesting quantiles, are also available.