Data Management for Volunteer Monitors

1999 Version

This manual is a potpourri of information on data management found in various other documents and publications, organized under MassWWP's outline. Copied are writings from Janice Miller, Geoff Dates, Fred Lease, Barb Horn and Helen Rosseli.

Goals of Data Management

Who is responsible for data management?

Let's assume you are the coordinator of a successful local monitoring program. The lab sheets are in. You are faced with a lot of scribbled numbers on paper. What now? It's not your job to manage the data? Maybe, but it's your job to ensure that the hard work of the volunteers who collected the data doesn't end up permanently in a cardboard box. So, if *you* have no desire to handle data after it's been collected, recruit a volunteer whose sole job it will be to manipulate numbers; someone with computer skills AND access to a computer. This step is more than a one-person job, however. We'll go into the reasons why.

Why should we manage our data?

Data management is as important as data collection and results dissemination. Without it, your monitoring project will probably not meet all of its goals, due to data loss, undiscovered errors, untimely discovery of errors, and plain disorganization. Data management, by organizing results, detecting and correcting errors, and summarizing results, will help you get a better view of your results and provide a clean historical record for future reference. Although some people view it as lackluster, data management is a necessary step toward the actions you want to take as a result of your monitoring. In other words, it is a tool to get your story to an audience.

When is the best time to manage our data ?

Before you get to the point of managing data, you should have thought out several points. Preferably you address data management in your study design document, before anybody goes out to collect samples. The old study design question: "what are your program goals" is as important now as ever. The way you process and display your data will be different depending on how and who you want to use your data. For example, if you are trying to rally your community around upgrading sewage treatment to improve recreation opportunities in your river, you will want to manage your data on a per sampling basis, for quick turnaround of data so that you may produce simple graphs or tables to be published in your local newspaper. Your computer program should be designed to automatically produce those simple graphs or tables once the data are entered and validated. If, on the other hand, your goal is to produce a baseline of data that a state agency will use to assess or manage your lake yearly or that scientists will use in their trend analysis studies, you have more time to enter data and produce sophisticated reports. In any case, it is a good policy to keep your options open and design your process with flexibility, as your goals might change with time and so might the data users.

Get results on forms

Design clear and useful forms for field sampling and lab work

The first step, once the study design is done, is to develop good field and lab data sheets. If you don't use existing sheets (available from MassWWP, River Watch Network, or Riverways, among others), make sure to include these important items when you design your own field data forms (Miller, 1994):

- exact location of the site. (If nothing else, at least have the volunteer provide precise instructions on how to drive to the site--see appendix for site location sheet)
- site name and number (see page 17 for site numbering protocols)
- date and time of sample
- volunteer name
- recent weather conditions. A recent storm can have a major effect on the readings
- name and number of equipment
- actual readings, replicate readings, values needed to get final answer--such as drops of titrant record as well as multiplier number
- comments: anything unusual that was seen--spills, new construction, a dead animal...

Filling out forms

It is a good idea to never erase what you think is an error. Instead, cross out errors neatly and write the correct entry next to it or above it. Why? Because sometimes you were right the first time around. Also, it documents that you found and corrected an error: you may have distributed the data before the error was discovered, and if you simply erase the error on the original document, it will be very hard to reconcile old photocopies with the updated data sheeets. Ideally, you should initial and date the correction, for future reference.

Reporting results

Don't report values of zero:
For example, if the range of a test is 0-10mg/l but the smallest measureable increment is 0.02 mg/l, report a zero value as "<0.02 mg/l"

- On lab sheets, don't report values of zero. Report "less than ——" (fill in the blank with the smallest detectable value).
- Calculations

  Significant figures: Report the answer using only the appropriate number of digits. Remember the following rules:

  * Constants do not affect the number of significant figures.
  * If you perform a series of calculations, carry two extra digits inside the calculations.
  * When working with a series of values, each with a different number of significant figures:

    **for addition and subtraction, round to the least number of decimal places in the series

    **for multiplication and division, round to the last signifi cant digit.

  Rounding: Round your final answer. Use only one digit past the significant digit to determine how to round the last significant digit. Between 0-4, round down; between 6-9, round up. If the digit 5 is dropped, round off the preceding digit to the nearest even number. Thus 2.25 becomes 2.2 and 2.35 becomes 2.4.
- Units: Watch your reporting units and write down conversion factors:"1,000 ml/liter", not just "1,000".
- Be sure to report the correct chemical constituent. For example if you require results for phosphorus but the test kit you use gives results as phosphate, it is important to convert the test kit value to phosphorus.

Significant Figures:
The number 10.5 has 3 significant figures.

If you multiply it by the constant 2, you obtain
10.5 x 2=21.0
(3 significant figures).

If you add the constant 1.25, you obtain
10.5 + 1.25=11.7
(still 3 significant figures).

Atomic weight of oxygen = 15.9994
Atomic weight of phosphorus = 30.9738
Molecular weight of phosphate (PO_4) = 94.9714

A quick and dirty shortcut is to divide the PO_4 value by 3 to get the P value.

To convert phosphate results to phosphorus results:

$$\frac{\text{atomic weight of P}}{\text{molecular weight of } PO_4} \times (\text{value of } (PO_4) = \text{value of P}$$

Gather forms at a single location promptly

The field data sheets should be dropped off at a convenient location by the collectors. There, the coordinator should look over the sheets before the collectors leave. If lab analyses are necessary, the coordinator then brings the sheets to the lab if it's at a different location.

Data Screening

Collectors

Before relinquishing their data sheets, collectors should review them with the coordinator to check for missing information and calculation errors. Missing information should be added, or an explanation should be written on the sheet. Errors should be crossed out and corrected.

Lab analysts

Before giving their results to the coordinator or the data manager, the analyst should also check for completeness and correctness.

Coordinator

It is very important that the data be checked and corrected by the coordinator at this point before transferring the information to any database or spreadsheet. You would be surprised how many errors are caught at this point! The sooner after field work you check all data sheets, the more likely you will be to resolve any problems.

Storage Format

On paper

NEVER throw away field and lab data sheets. These sheets are the primary records, and anyone genuinely interested in the quality of the data may want to see them. It's even a good idea to photocopy them and file the originals in a safe place (someone needs to be willing to be the keeper of the data). Give the copy to the computer entry person. Also give back a copy of their fields data sheets to the collectors, so they can refer back to them if you have a question.

On computer

A this point, computer entry is strongly recommended. Storing the data electronically is efficient and allows its flexible view, manipulation, and retrieval. Computers also hold large amounts of information.

Software considerations:
Cost is only one factor. Know what the software can and cannot do versus what you need. Ask for what you need it to do, including printing capabilities. Other points to consider are the software's compatibility with other data users, its translatability to other forms of software and hardware, and whether there is already something out there which is already formatted for your needs. Again, keep in mind who will use the data.

The question often arises whether to use a spreadsheet type of software (such as Excel, Lotus 1-2-3, Quattro Pro, etc.) versus a database manager (Accesss, dBase, FoxPro...). These are basically different approaches to storing and retrieving data (Dates, 1996). Databases are harder to use, but are good for storing large amounts of data and to perform queries--that is, to retrieve a protion of data that satifies one or several attributes. Spreadsheets are easier to use, but can become cumbersome with large amounts of data, although this problem is diminishing these days, as newer versions of spreadsheets have many database capabilities. With most modern spreadsheets, it's easy to generate graphs of your data -a valuable feature.

Which brand of software is best? (taken from Maryland Save Our Streams' Fred Lease's article on Designing a Data Management System, in the *Volunteer Monitor* Volume 7, No 1).

"When a volunteer group considers what brand of software to buy, cost may be the first factor that comes to mind. Can you get away with a less expensive program, perhaps one that was supplied with your computer? Maybe; but be careful. Cheaper spreadsheet programs are often limited in their mathematical, statistical, and data-sorting capabilities. Inexpensive databases may... not have the wide assortment of retrieval capabilities found in more sophisticated programs. Also, less expensive programs may be weak in graphing capabilities and printing options, whereas effective data presentation is often extremely important for monitoring groups. [One statewide monitoring program coordinator] cautions: "In choosing a software package, don't rely on the expertise of one volunteer. If that volunteer leaves, you will be up a creek!" [The coordinator] recommends choosing software that is already supported by an agency or university that you work with, so that you will be able to receive training and assistance. Another important consideration when purchasing software is how easily data files can be translated by another software program. For instance, Paradox database software was specifically designed by Borland for use with Borland's Quattro Pro spreadsheet software... Take time, now, to ask other monitoring groups which software programs they are using and what problems they may be encountering with file conversion."

MassWWP has used dBase for Windows to manage our data, but plans to convert to Microsoft Access. For small databases, we use Microsoft Excel, which has useful graphing capabilities. Most groups we work with use either Excel or Access.
For a detailed look at basic system design, see the rest of Fred Lease's article on page 7 of the Volunteer Monitor Volume 7, No 1.

Computer Data Entry

This step involves taking the numbers from the lab sheets and entering them into the spreadsheet or database. These numbers are in the units that are measured in the field or lab (e.g. digits, or absorbance), not necessarily the final units (e.g. milligrams per liter). These numbers are entered by setting up the spreadsheet as follows:

	A	B	C	D	E	F
1		**Billington Sea 1995 Dissolved Oxygen**				
2						
3						
4	**Date**	**Site 1**	**Site 2**	**Site 3**	**Site 4**	**Site 5**
5	29-Apr	10.6	11.64	10.08	9.56	10.04
6	14-May	9.92	10.8	11.12	8.96	9.88
7	04-Jun	2.32	7.56	8.56	8.52	2.04
8	17-Jun	9.8	7.12	7.64	5.8	9.36
9	09-Jul	8.96	9.36	6.24	9.52	9.2
10	29-Jul	8.72	10.68	8.08	1.28	8.52
11	12-Aug	6.8	7.08	4.2	0.84	6.6
12	26-Aug	9.44	9.32	8.96	9	7.92
13	30-Sep	9.32	8.9	10.12	8	6.56
14	14-Oct	10.6	10.6	8.04	5.36	10.76

or, another example in database form:

DATE	SITE	LAKDPTH	SECDPTH	DO	PH	ANC	TP
04/29/95	1	2.4	0.8	10.60	7.17	13.1	0
04/29/95	2	2.2	1.0	11.64	7.17	11.8	0
04/29/95	3	2.2	1.0	10.08	7.19	12.6	0
04/29/95	4	3.5	1.0	9.56	8.83	13.6	0
04/29/9	5	3.8	1.0	10.04	8.87	12.7	0
04/29/95	6	3.8	0.9	6.70	8.74	12.3	0
04/29/95	7	2.4	0.9	10.24	9.11	13.7	0
05/14/95	1	2.3	1.1	9.92	7.53	15.2	64
05/14/95	2	2.1	1.2	10.80	7.39	15.1	36
05/14/95	3	2.2	1.1	11.12	7.20	14.1	339

Note in this database file we used site as a field; This allowed us to enter more than just dissolved oxygen values: for each site there is also data on the Secchi Disk Transparency (SECDPTH), the lake depth at that site (LAKDPTH), pH, acid neutralizing capacity (ANC) and total phosphorus (TP). One limitation of this database program (dBase) is that variable names can not exceed 8 characters.

Try to enter data in the computer promptly after each data collection. The data manager may find questionable numbers and find out answers from collectors and analysts while the sampling is still fresh in their mind. Again, have a volunteer who is dedicated to computer work. Preferably this person is not also sampler or a lab analyst (for overload/burnout reasons as well as for quality control).

Enter as much information as you can into a database or spreadsheet (Miller, 1994). Set up comment fields for text and, if needed, flag fields for data. For example you would have one field (column) for pH and another field next to it for pH-flag. In the pH flag, you enter a code devised by you to alert the database user of potential problems with the data. If you suspect a certain result to be erroneous but can't prove the error, you put a code in the flag field, as follows:

Flags

Date	Site	pH	pH-Flag
04/15/97	1	7.2	
05/13/97	1	7.4	
06/14/97	1	4.5	Q

Where Q stands for Questionable: the reported pH of 4.5 is questionable although there is no proof that there was an error. When the data is fine, don't put anything in the flag field. Then when you summarize the data, you can have your computer program use only the data for which 'pH-Flag not equal to Q'.

Missing Values and Values Below Detection Limits
Enter missing values as ND for NO DATA. That way the proof-reader will know you didn't just forget to enter the data.
Enter values that were below detection limits as BDL, not as zero.

Always organize the database so it is self-explanatory; if that is not possible, for example if you can't fit self-explanatory field names, document the database well, maybe with a 'Read Me' file that explains what each column and row represents. In Excel, you can attach comments to individual cells.

After the data is entered, it should be printed and then, if possible, checked against the original paperwork by another person. Basically, they are looking for keypunch errors. If no one else is available, the data entry

person can do the proofing, that's better than no checking at all. It's worth taking the time to ensure accuracy, because once data gets into a database, it has a life of its own: People **believe** data that's in a database.

Data validation

Obviously, the next step is to correct any error encountered in the data entry. Once that is done, reprint the data and give the printout to the coordinator or another person who is familiar with water quality data. This person will do a final check for accuracy (or data validation), as follows:

Screen for outliers: An outlier is a data point that is way outside the range of the other data points. Screen the data for outliers by comparing it to past data from the same site or similar sites. Visually scan the data, looking for values that are off by a factor of 10 or 100. Look at the minimums and maximums and see if anything jumps out at you, such as a Dissolved Oxygen value of 19 mg/l or pH of 15 (impossible values). Calculate a mean and median. If they are very different from one another, you probably have an outlier. Graphing data points is an excellent way to spot outliers. If you find an outlier, check the paper sheets and ask the collector or analyst for input so you can decide whether the entry is erroneous or if the data point is valid. Some validation can be automated in your computer program.

Screen for consistency: You should check for consistency between similar parameters. For example:

- Total solids should be greater than suspended solids.
- Total phosphorus should be greater than orthophosphorus.
- Total dissolved solids and conductivity should track together (go up or down together)
- Total solids and turbidity should track together.
- The chemistry should match what was visually seen in the field (construction, recent storm, etc.)

If the data shows inconsistencies or doesn't make sense, follow up on the problem. If you can track down and correct the error, fix it and note that you fixed it on the data sheet. If you really suspect that the number is incorrect but you can't figure out the correct result, remove it from the database and note on the data sheet that you did so. But don't take data out simply because you don't like it. Remember to use flags for data that is questionable but not demonstrably wrong. If you correct an error, alert others to whom you have already given the data so they can fix it too. Again, remember that data in a database has enhanced credibility and "a life of its own".

Finally, the data should be backed up (files copied onto disk or tape) regularly for safekeeping.

Data Conversion

This section uses examples and text from River Watch Network's Geoff Dates.

Make sure your computer person has all the conversion factors/formulas to produce data in the proper units

For data that must be converted to final reporting units (e.g. converting bacteria filter counts to colonies per 100 ml), there are two ways to proceed: one is to add another field or column in your spreadsheet/database, where the raw data is automatically converted with a formula. Another is to set up another table that is linked to the first one, but will convert the lab data (filter counts) to reporting units (colonies per 100 ml):

From this raw data file:

	E. coli Bacteria (Filter Counts, 50 ml Subsample)						
	1992						
Sites	27-May	16-Jun	3-Aug	20-Aug	23-Sep	7-Oct	21-Oct
1	100	45	770	155	3550	124	4
13	100	1200		6100	5000	900	650
14	100	3750		6900	5000	1450	350
15	4	215	10000	3100	600	300	700
16	100	12000	10000	1250	350	100	250
17		0			250	162	150

we convert all the data by multiplying them by 2:

	E. coli Bacteria (colonies per 100 ml)						
Sites	27-May	16-Jun	3-Aug	20-Aug	23-Sep	7-Oct	21-Oct
1	200	90	1540	310	7100	248	8
13	200	2400		12200	10000	1800	1300
14	200	7500		13800	10000	2900	700
15	8	430	20000	6200	1200	600	1400
16	200	24000	20000	2500	700	200	500
17		0			500	324	300

Summarizing Data

Besides converting data, the data manager is also responsible for summarizing the numbers into figures more easily grasped by data interpreters. There are several ways to reduce data:

Means (averages) and Medians: If you have a large data set or data from several years, you may want to summarize it by some sort of averaging. There are three types that are commonly used:

- Mean (Average): a set of data is averaged by adding all the individual data points and dividing by the number of data points. Below is our example where the results for each site are averaged for the entire season:

Mean

	E. coli Bacteria (colonies per 100 ml)							Mean
Sites	27-May	16-Jun	3-Aug	20-Aug	23-Sep	7-Oct	21-Oct	1992
1	200	90	1540	310	7100	248	8	1357
13	200	2400		12200	10000	1800	1300	4650
14	200	7500		13800	10000	2900	700	5850
15	8	430	20000	6200	1200	600	1400	4263
16	200	24000	20000	2500	700	200	500	6871
17		0			500	324	300	281

The problem with averaging these numbers is that the very high and very low numbers can skew the results that distorts the picture. For example, the results for site 1 are below 1000 on all but one of the samples dates. Yet the average is above 1000. So the mean is not a very good way to summarize this data set. In general, the mean is not a good way to summarize environmental data.

- Geometric Mean: a set of data is transformed to the logarithmic values of each data point; these are averaged, and then transformed back to the original units. The geometric mean reduces the influence of the very high and very low numbers on the data set. Below is the same example we've been using, with a column added for geometric mean calculated for each site for the entire season:

Geometric Mean

	E. coli Bacteria (colonies per 100 ml)							Geo. Mear
Sites	27-May	16-Jun	3-Aug	20-Aug	23-Sep	7-Oct	21-Oct	1992
1	200	90	1540	310	7100	248	8	276
13	200	2400		12200	10000	1800	1300	2271
14	200	7500		13800	10000	2900	700	2737
15	8	430	20000	6200	1200	600	1400	886
16	200	24000	20000	2500	700	200	500	1496
17		1			500	324	300	83

Notice that in all cases, the geometric mean is lower than the regular mean. For example, for site 1, the influence of the very high result on 23-Sep is reduced on the geometric mean. For Site 17, we replaced the zero value on 16-Jun with a 1. That is because you can't take the logarithm of zero, but the log of 1 IS zero. This creates a minor error in the result, but allows the use of the geometric mean calculation.

TIP
Note that when you calculate statistics, you remove the 'ND' and 'BDL' markers from your spreadsheet because they interfere with the calculations

- Median: This is the central value of a set of data ranked lowest to highest. Like the geometric mean, the median reduces the influence of the very high and very low numbers on the data set. See below the new column for median calculated for each site for the entire season:

Median

	E. coli Bacteria (colonies per 100 ml)							Median
Sites	27-May	16-Jun	3-Aug	20-Aug	23-Sep	7-Oct	21-Oct	1992
1	200	90	1540	310	7100	248	8	248
13	200	2400		12200	10000	1800	1300	2100
14	200	7500		13800	10000	2900	700	5200
15	8	430	20000	6200	1200	600	1400	1200
16	200	24000	20000	2500	700	200	500	700
17		0			500	324	300	312

TIP
If there are BDL values, replace them with zeroes while calculating a median

Notice that the median is closer to the geometric mean of the data set than the mean. Note that for site 17, the median is 312, which is not one of the data points. Since there is an even number of data points, the median is the average of the two central points.

Quartiles and the Interquartile Range

Used less frequently than the means and medians described previously, quartiles and the interquartile range are very useful to characterize a set of data. Quartiles are the three values below which lie 25%, 50%, and 75% of the values in a set of numbers. The median is the 50% quartile. But, while the median shows you the typical value in your data set, the other two show you the spread of the data. Looked at together, the 25% and the 75% quartile show you the values between which 50% of your data lies. This is referred to as the interquartile range. If these values are close together (a narrow range), it means that your data set is relatively consistent and clustered around the median. If they are far apart (a wide range), it means that there is a lot of variation in your data. This is useful when you're trying to determine if there is a trend over time or space.

TIP
Microsoft Excel has median, geometric mean, and quartiles functions ready to use in your spreadsheets

This summarizing tool is not usually used with bacteria data, but for the sake of staying with the same example, let's look how the quartiles look for the bacteria data:

Quartiles

	E. coli Bacteria (colonies per 100 ml)							1st Quart.	Median	3rd Quart.
Sites	27-May	16-Jun	3-Aug	20-Aug	23-Sep	7-Oct	21-Oct		1992	
1	200	90	1540	310	7100	248	8	145	248	925
13	200	2400		12200	10000	1800	1300	1425	2100	8100
14	200	7500		13800	10000	2900	700	1250	5200	9375
15	8	430	20000	6200	1200	600	1400	515	1200	3800
16	200	24000	20000	2500	700	200	500	350	700	11250
17		0			500	324	300	225	312	368

What these quartiles show us is that sites 1 and especially 17 have much more consistent bacteria counts than the other sites. This information is actually much easier to see on a graph, as we'll describe later.

More Data Manipulations

There are some extra, easy steps the data manager can take to make the data interpreters' job easier. One is to add a column in the spreadsheet or database with the water quality standard for each parameter. Take it a step further and add a column which automatically compares a result with the water quality standard and puts out a comment such as "violation" or "meets standard". In the bacteria example, the monitors chose to compare their results to EPA's standard for "lightly used full body contact recreation" of a gemetric mean of 406 colonies per 100 ml:

Compare with standard

	E. coli Bacteria (colonies per 100 ml)			
	Geo. Mean	WQ	Meets	
Sites	1992	Standard	Standard?	
1	276	406	OK	
13	2271	406	Violation	
14	2737	406	Violation	
15	886	406	Violation	
16	1496	406	Violation	
17	83	406	OK	

In the above example, we used the "if" function in Excel.

Data Display

Finally, after all manipulations are completed, the data should be displayed on paper for distribution to the people who are charged with data interpretation.

From the spreadsheets or databases described above, the data are printed in tables or in the form of various charts and graphs. This allows one to view the information as a whole.

The bacteria example we used previously just showed one parameter for one year. Usually, the data interpretation crew will be looking at several parameters, perhaps for several years. How then can you display all those data clearly? You can view your data per site or per date, to get a different perspective and help your interpretation. Or you can use only one ordering scheme, but graph the data in a variety of ways, with date on the x-axis, or site on the x-axis.

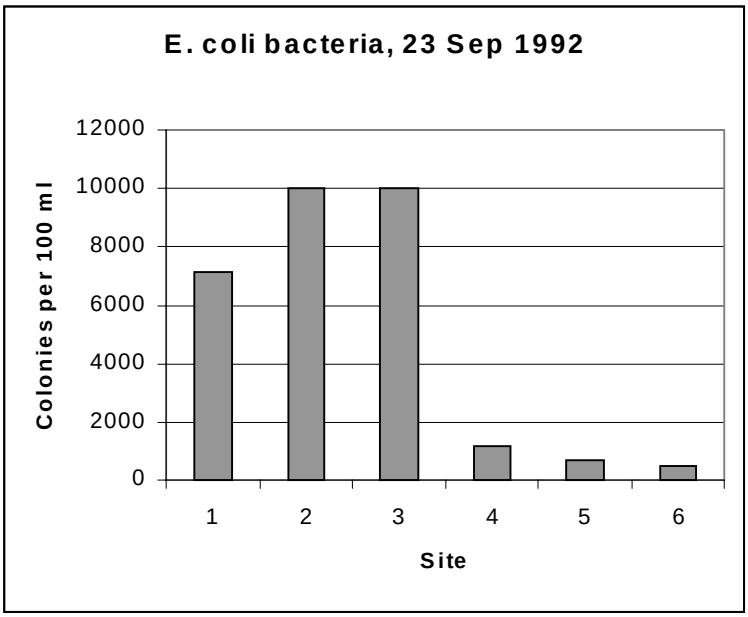
You can also query your data in a variety of ways: an example would be to ask your program to show you data only for one site, or for a specific year (simply printing the spreadsheet we used before), or to show only two parameters, side-by-side, such as temperature and oxygen. This again can help focus on one issue at a time by hiding irrelevant data, and rearranges data to make interpretation clearer.

Charts

A chart or graph is a way of presenting the data in a visual form as lines, bars, pieces of a pie, points, etc. Graphs show relationships between two or more sets of numbers and elements. Graphs may reveal trends or results that exceed or don't meet a standard better than tables would. Decide what combination of graphs and tables is the best way to summarize your data. Most spreadsheet programs have fairly extensive graphing capability. Databases such as dBase do not. Examples of graphs follow (again, from River Watch Network):

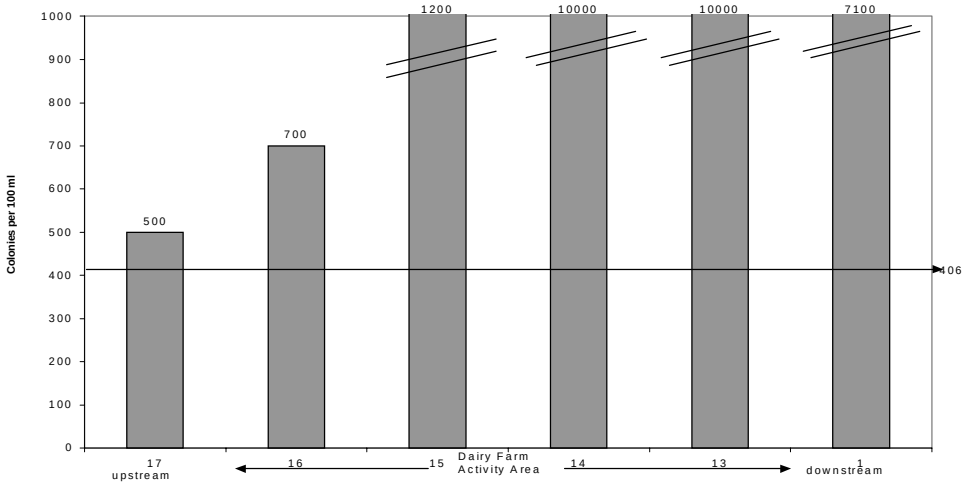
The most common form of data display is a simple column or bar chart:

Bar Charts



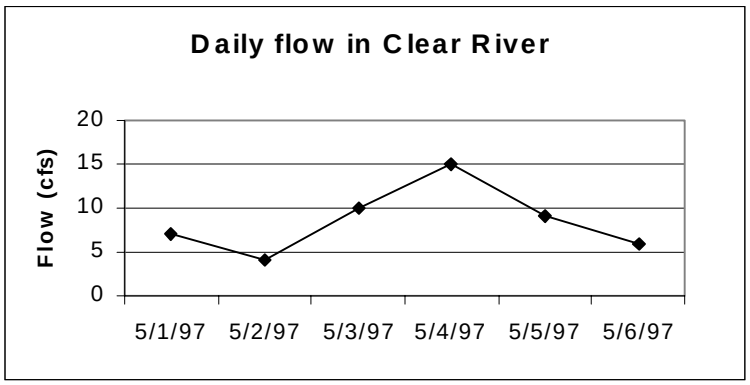
Sample sites are listed along the horizontal axis, upstream right to left. A range of E. coli (from 500 to 10,000) is displayed on the vertical axis. Results for each site are shown by the bars. The height of the bars corresponds to the level of bacteria found at each site.

In the graph below, we've added a few graphic elements to the chart in order to reveal how the data compares with the water quality standard and activity between sampling locations: Note the addition of the water quality standard which shows that all sites are in violation; the location of a dairy farm, and upstream vs downstream. We also changed the order of the sites so the upstream site would be on the left. Finally we changed the scale of the y-axis to show the details near the water quality standards. This caused the high counts to be cut off at the top, so we added parallel lines to show the cutoff, and added labels on each bar so the value could be seen.



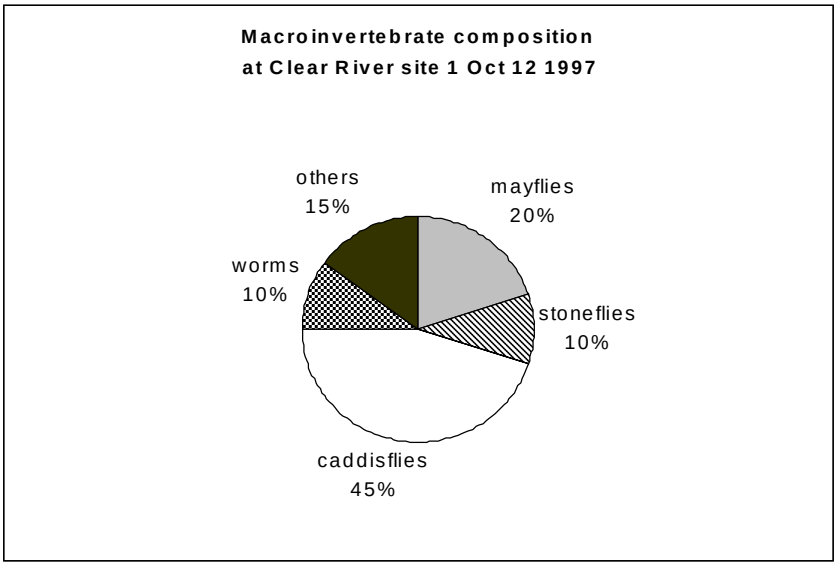
Line Graphs

In line graphs, data points are joined by a line, which implies that there is a continuous link between points. You therefore don't want to use a line graph to represent the bacteria data we've been using, because there is no garantee that the bacteria amount between sites 16 and 17 fall mid-way between the amount of bacteria at those sites. A line graph is fine to use for precipitation amounts or stream flow, however:



Pie Charts

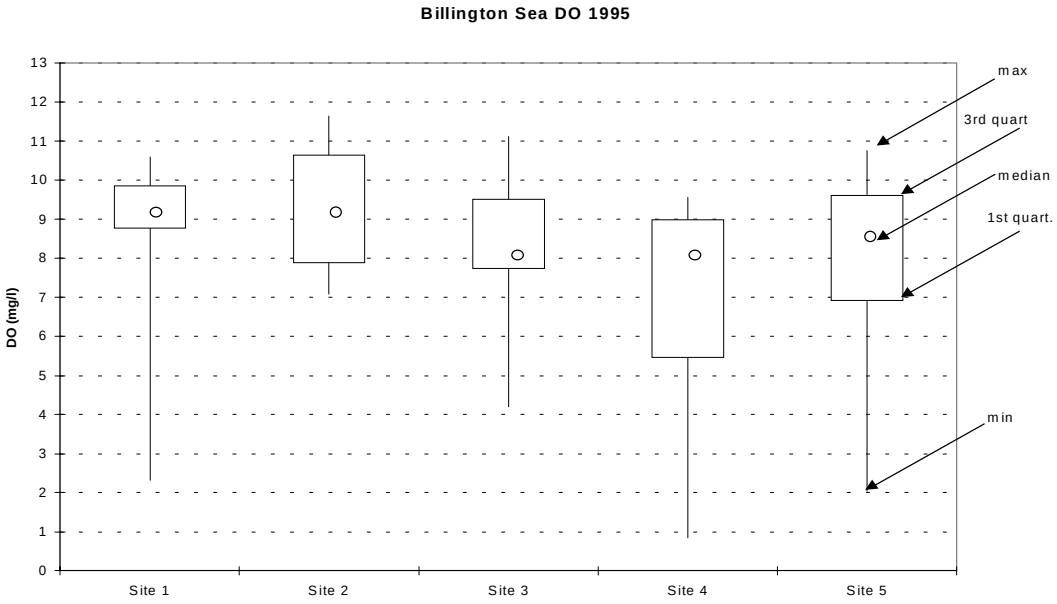
Use pie charts when you want to compare parts of a whole. For example, it is good to depict the composition of a macroinvertebrate (bugs) sample: If out of 100 bugs in a sample, you had 20 mayflies, 10 stoneflies, 45 caddisflies, 10 worms, and 15 others, you would chart it like this:



Box and Whisker Plots

To compare data among sites or from year to year, a look at quartiles can be most enlightening : a graph of the data distribution will show you the spread of data (is there a lot of variability or not?) and whether one site or one year is obviously different from another. In a box and whisker plot, you show the minimum, maximum, median , first quartile and third quartile of the data at that point. The box represents the quartiles, the 'o' represents the median, and the lines or 'whiskers' represent the min and max values.

Tip
To do this type of plot in Excel, use the stock chart type and plot series in this order: 3rd quart., min, max, 1st quart. Then add median by drawing circles with the draw tool



Other Topics

Making your data GIS compatible

A Geographic Information System (GIS), where maps are digitized (entered into a computer format) is a powerful tool to analyze geographic data. In the environmental field, people use GIS to visualize the land use in a watershed, or to see soil variability, slope steepness, location of farms or industries-- the possibilities are endless. But GIS is more than a tool to display fancy maps. It is used for analyses that would be very difficult otherwise. For instance, you could find out whether total phosphorus in a river or lake is positively correlated to the proximity of a potential nutrient source such as a manure holding area.

To make your data useable in a GIS, sort them by a special code which refers to the sampling site. Find out if there is an official code for lakes and streams in your state already assigned to hydrology coverages. For example, in Massachusetts, there is an official code for lakes called PALIS, and one for streams called SARIS. See the next section for site numbering protocols. In any event, it's a great idea to note the exact location of your sampling site using latitude and longitude. In your database or spreadsheet, just add two fields or columns, one for longitude and one for latitude.

When your data and the GIS coverages share the same code, they can be linked and your data can be attached to a map coverage. Massachusetts uses the software Arc/Info for its GIS. It is a fairly unwieldy program, and volunteer groups tend to use the 'little brother' version ArcView. ArcInfo and ArcView can use databases in the software dBase directly, so storing your data in a dBase file makes it very easy to use it with GIS. However, Access and Excel can export data into a dBase format, so in any case your data can be linked to geographical information. This gives you the ability to present your data on a map, and also to link your data to an entire drainage system. For easiest use, obtain ArcView and the Spatial Analyst module. Then you can add your sites on a coverage as well as your monitoring data.

Site numbering system protocol

MassWWP recommends a rigid site numbering coding system, which eases computer entry and data analysis and standardizes codes statewide. Codes were developed in Massachusetts for streams and rivers (SARIS) and for ponds and lakes (PALIS). The first two digits in the codes stand for the

watershed. Almost all surface waters have been assigned a code. PALIS codes have 5 digits, with the first two referring to the major watershed; for example, Black Pond in Taunton has the code 62016, 62 being the code for the Taunton watershed. SARIS codes have 7 digits, with the first two also reserved for the watershed; for example, the Taunton River SARIS is 6235000. MassWWP uses a field in its data base called PALSITE, which is the SARIS or PALIS code for the water body, plus 4 decimals to indicate the sampling site. For example, a site near the mouth of the Taunton River would have a PALSITE of 6235000.0010.

Call MassWWP or Office of Watershed Management at DEP if you want to find out your stream's or lake's code. You will need to know the name(s) of the water body and the town it's in. Once you have your PALIS or SARIS code, you need to make it a PALSITE code by adding a decimal point and four decimals.

Streams

For streams, the best way to number your sites is to measure on the map the distance between the sampling point and the stream's mouth (MassGIS uses miles), and use the distance as the decimal. For example, if your site is 29 miles upstream of the mouth, the PALSITE would be SARIS.0029. This doesn't address a tributary's tributary situation, if small streams don't have a SARIS number; you can come up with your own protocol there, as long as it is consistent. The important point is to get accustomed to using SARIS and PALIS numbers, as they are used in the MassGIS system.

Ponds

For ponds, start with your PALIS code, a five digit number such as 34675. Use the second decimal as your site number: 34675.0100 is site number 1. By the way, do use numbers and not characters such as 'Deep Hole', which is difficult to enter in the computer. MassWWP's advice is to reserve number 1 for your deepest spot site on the lake. There is no rule for subsequent numbers. Use the third and fourth decimals for the depth of sampling, in meters. So, if your sample is taken at the surface at the deep spot, the PALSITE would be 34675.0100 (PALIS.0100). If you also take a sample at a depth of 9 meters at that site, the sample code will be 34675.0109. If you take a sample at 21 meters at site 15, your PALSITE will be 34675.1521.

Data management is thus more than a one-person job: it all starts at the

Conclusion

beginning of a project, during study design, when field sheets are developed and responsibilities are assigned. It involves the following people and their responsibilities:

Data collectors: they fill out field sheets according to protocol and check their sheets before relinquishing them to the coordinator or the lab analyst

Lab analysts: they fill out the data sheets, or complete the field sheets if that is where the analyses results are reported, and check for completeness and accuracy before relinquishing the sheets to the coordinator or the computer operator.

Project coordinator: who collects all sheets and checks them for completeness, and calls collectors and lab analysts when information is missing or questionable. The coordinator also validates the data after computer entry by looking for outliers and inconsistencies. If the coordinator is not familiar with water quality data, a science-literate person must assume this responsibility.

Computer operator: who enters all data in a computer program, converts data into proper reporting units, and produces summaries such as tables and graphs for ease of data interpretation.

Proof reader: who checks the computer data entry against field and data sheets.

With a reliable crew of volunteers and a clear plan for the custody of data from the collection step to the display step, managing data becomes relatively painless and definitely productive. Once your data are transformed to intelligible information, the next step will be data interpretation.

Bibliography

Dates, Geoff, 1996
Data to Information, in Testing the Waters, Chemical and Physical Vital Signs of a River by Sharon Behar, River Watch Network 153 State Street Montpelier, VT 05602

Horn, Barbara, 1996
Dealing with your Data - The Basics. *Proceedings for the 5th National Volunteer Monitoring Conference* (to be printed)

Lease, Fred, 1995
Designing a Data Management System. *The Volunteer Monitor*, Volume 7. No. 1, pp 6-8 & 23

Miller, Janice K., 1995
Data Screening and Common Sense. *The Volunteer Monitor*, Volume 7. No. 1, pp 4-5

Pelto, Karen, 1994
The Massachusetts Water Watch Partnership Manual for Volunteer Water Quality Monitors. MassWWP Blaisdell House Univ. of Mass. Amherst MA 01003-0820

Rector, Julie, 1995
"Variability Happens" Basic Descriptie Statistics for Volunteer Programs. *The Volunteer Monitor*, Volume 7. No. 1, pp 14-16

Appendix

Worksheet # 2

Sample Site Evaluation Sheet
River Watch Network
Revised 6/28/91

Date: _____ Sampling Station #: _____

Sample For: Chemistry ___ Macroinvertebrates ___	*Reason for Sampling:*

Site Description, Location, and Accessibility

River:		*Watershed:*
River Mile (approx.):	*U.S.G.S. Quad:*	*Park At:*
Town:	*County:*	
Site Location (Include important landmarks):		
Directions To Site From Nearest Major Highway Intersection		

Sampling

Describe Where At Site To Take Sample (best access to main current):

Sample From:	*Getting From the Parking Area To the Water:*
Boat ___	Easy ___ Explain (e.g. steep bank, poison ivy):
Wade: ___	Moderate ___
Bank: ___	Difficult ___
	Access Across Private Property ___ Across Public Right-of-Way ___

River/Stream Characteristics

Stream Habitat: (check if present) Riffle (small waves on surface) Glide (smooth moving water) Pool (still water)

Stream Bottom (estimate % in each category, see size of particles in parentheses):

____ % Bedrock	____ % Boulder (>10")	____ %Cobble (2-10")	____ % Gravel (0.1-2")	____ % Sand < 0.1"	____ % Silt

Depth (estimate) ____ ft.	Current: Fast ___ Moderate ___
Width (estimate) ____ ft.	Slow ___ Still ___

Stream Shading: 100% ___ 75-100% ___ 50-75% ___ 25-50% ___ <25% ___

River Use and Pollution

List Any Known or Observed Recreation! Uses	*Describe Any Evidence of Pollution or Sources*

Additional Comments On Back?