GAH6.6: Temporal and geographical trends in infectious diseases: analysing mumps incidence for U.S. States

Objectives:

- to be able to present basic characteristics of infectious disease transmission through time and space
- to become familiar with different formats of infectious disease data

The data:

The data relate to reported monthly cases of mumps in U.S. States between 1968 and 1988 and comes in the form of a text file:

- **mumps.csv**: mumps cases by U.S. state for 1968-1998. The first column is a unique numeric ID code for each U.S. state, the second column is the year, the third column is the month, and the fourth column is the number of reported mumps cases in that month.

Acknowledgements: These data were made available to the general public as part of an American Statistical Association (ASA) Section on Statistical Graphics entitled "Statistics in Public Health Surveillance" at the August 1991 Joint Statistical Meetings in Atlanta. The data were prepared by the authors of the following paper (Note: it is **not** necessary to read this paper to complete the practical):

E. Freund; P. J. Seligman; T. L. Chorba; S. K. Safford; J. G. Drachman; and H. F. Hull (1989): 'Mandatory Reporting of Infectious Diseases by Clinicians'. *Journal of the American Medical Association* **262**:3041-3044

The work of the ASA and the authors of this paper in making the data available are both gratefully acknowledged here.

Activity:

We will use ArcMap to calculate the various summary tables for the number of mumps cases between 1968-88.

Importing data into ArcMap:

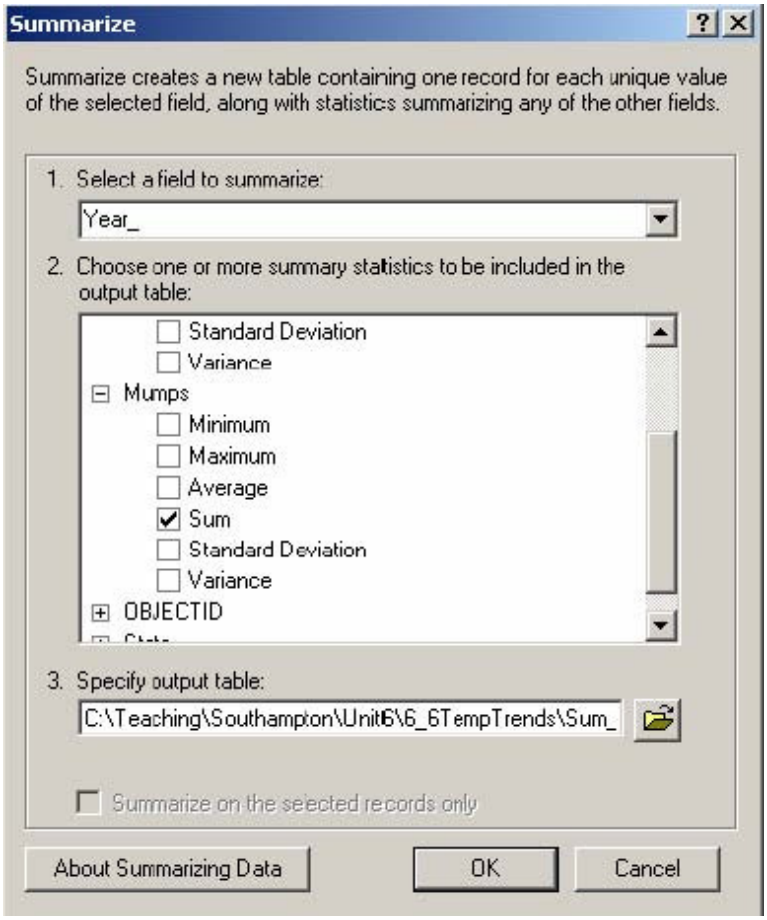
First of all, we need to open up our data in ArcMap:

- Choose *add data* and navigate to the folder where you stored the mumps.csv file
- Right-click on the data file and choose *open*.
- Note the asterix next to the number of records – this means that the file hasn't been loaded into memory completely. ArcView loads up the information in batches of 2,000 records.
- Scroll down to the very bottom of the list until the asterix disappears – there should be around 10,342 records in total.

Summarising attributes in ArcView

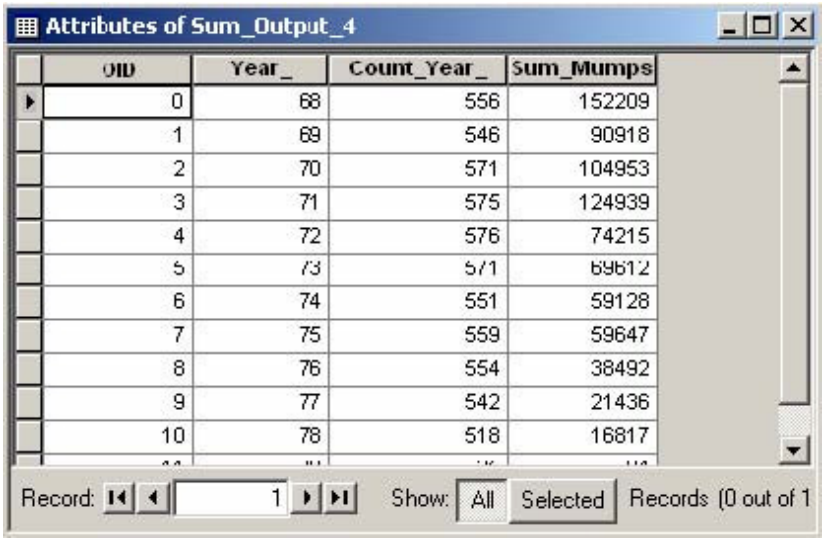
We can answer a lot of questions about trends in disease using the ArcView *summarize* facility. For example, one thing we can do is to work out the number of mumps cases in the whole of the USA in each year:

- Right-click on the *year* field name at the top of your table grid
- From the menu that pops up, choose *summarise*
- In the *summarize* dialog box, click on the **mumps** box and choose *sum*



- At the bottom of the dialog box, enter in a name for your new table under *specify output table*
- You will be asked if you want to add this to your map, so choose *yes*.
- Now right-click on the name of your new table and choose *open*.

You should now see a table that contains the year and a field called **sum_mumps**, the total number of mumps cases in the whole country in that year:



A couple of other fields are also created automatically – an object ID to uniquely identify each row (**OID**) and a count of the year (**count_year_**), which in this example is not very meaningful.

Your turn – some tasks:

Now you have seen the *summarise* facility, try the following:

1. Use the *summarise* facility to calculate the average number of mumps cases over the whole study period for each month. Save your results with the name **question1**.
2. For each state, use *summarise* to calculate the maximum number of monthly mumps cases. Save this table as **question2**.

More complex analysis – finding the month and year of peak measles incidence

We can answer many questions about geographical trends in mumps using fairly straightforward table manipulation. However, sometimes we need to undertake more complex analyses.

For example, suppose that instead of knowing the maximum monthly number of measles cases in each state, we wish to know the month and year when the

peak in mumps cases occurred. This is a more complex question that cannot be answered by summarising the values in one field.

To answer this question, we will need to join our summary data (**question2**) about the maximum number of measles cases per state back to our original data set. The diagram below shows why we need to do this. To find out when the peak mumps incidence occurred, we need to pull together the *year_* and *month_* information in the original data set and *maximum_mumps* from our summary data. The diagram below illustrates this:

Mumps table
ObjectID
State
*Year_*
*Month_*
Mumps

+

Question2 table
ObjectID
State
CountState
Maximum_mumps

=

New table
ObjectID
State
Maximum_mumps
*Year_*
*Month_*

If you wanted to join together these two tables, which field(s) would you use to join the tables together? Take a moment to think about this before reading on.

Mumps table
ObjectID
State
*Year_*
*Month_*
Mumps

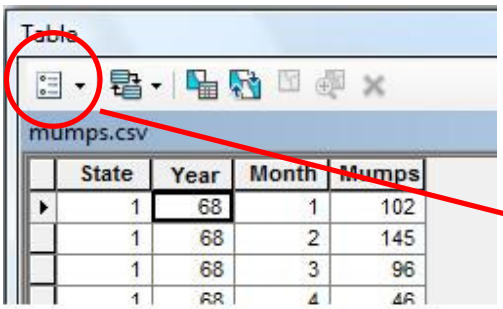
Question2 table
ObjectID
State
CountState
Maximum_mumps

New table
ObjectID
State
Maximum_mumps
*Year_*
*Month_*

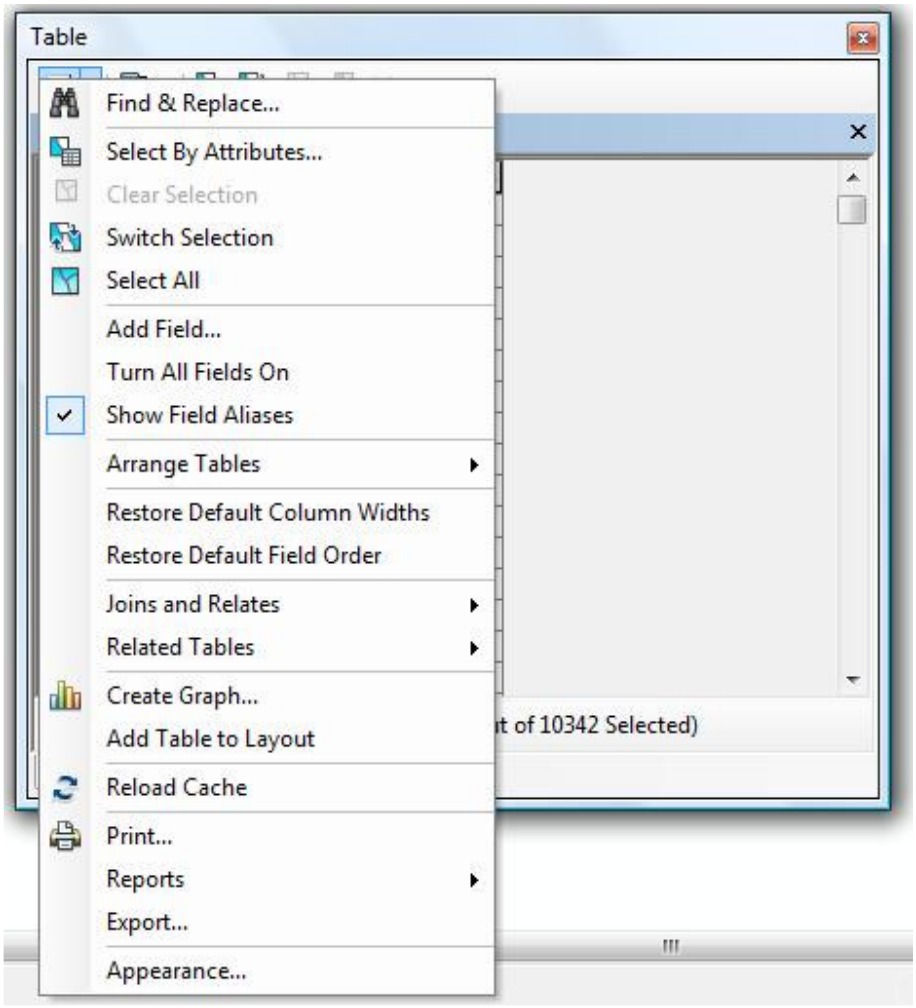
What we need to do is to link the two tables using the *State* ID number and the number of mumps cases. That way, we avoid mixing up data from different states and we make sure that we just pull out the year and month for the maximum number of mumps cases.

Unfortunately, ArcView only allows us to join or relate tables using just one field – not two. If we were wanting to join our tables using *State* only, we could do this quite easily, but because we want to use two fields (*State* and *Maximum_mumps*), we have to do some work first. First, we need to create a new database field that contains the information from both *State* and *Maximum_mumps*.

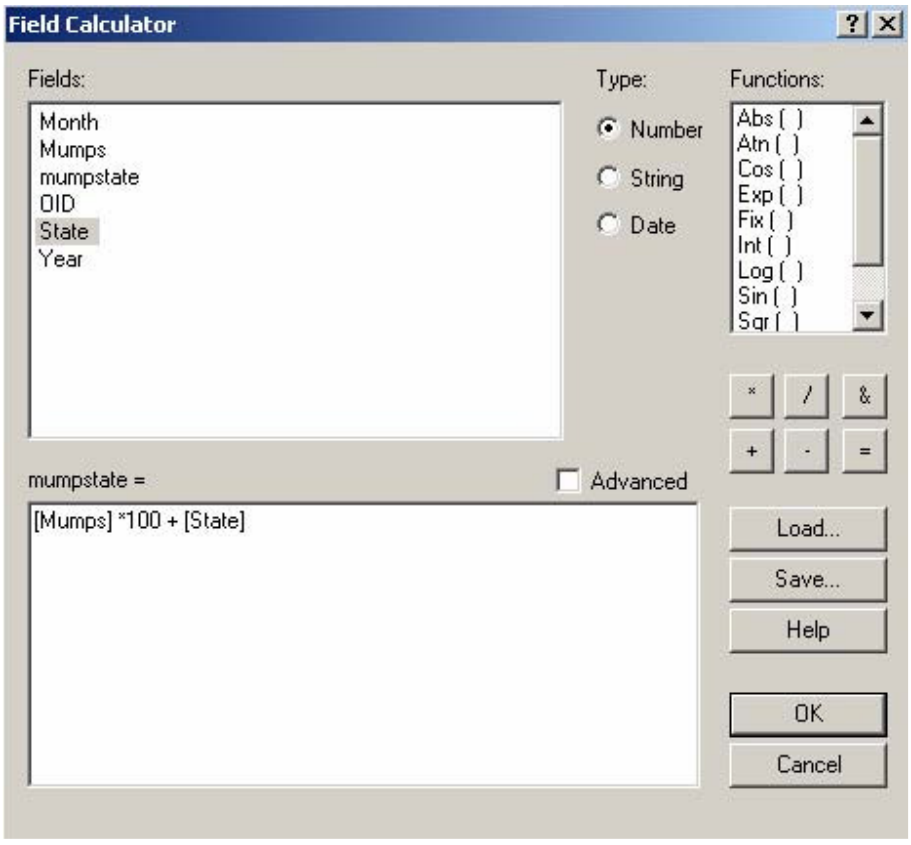
Creating a new field to join the tables together



Right click on the *mumps*.csv file and choose *open*. If you click on the *table options* button, you'll see that the option *add field* appears in gray and is disabled. In ArcView, we can only add fields to GeoDatabases or dBase tables – not to data like this, which is stored as a text file.



- We therefore have to convert our data to a table or geodatabase first: Click on the *table options* button again and select *export…* and create a dBase table called **mumpsdb**
- You will be asked if you want to *add your table in the current map* – choose *yes*
- Now right-click your new **mumpsdb** table and choose *open*
- This time, you should be able to *add field…*, so create a new field called **mumpstate** with a *long integer* data type.
- Right-click on the field name **mumpstate** and select *field calculator…* Click *yes,* when asked whether you want to proceed.
- In the new field **mumpstate**, we are now going to create a series of numbers that contain information about both the state ID and mumps cases. We will make the last two digits the state ID and the first 2-4 digits the numbers of mumps cases. Can you think how you could do this in a calculation? Take a moment to think about this before turning the page.



- Here's how: multiply the number of mumps cases by 100 and then add on the State ID (see above). That way we end up with a new number that contains information about both these things. You can click on the

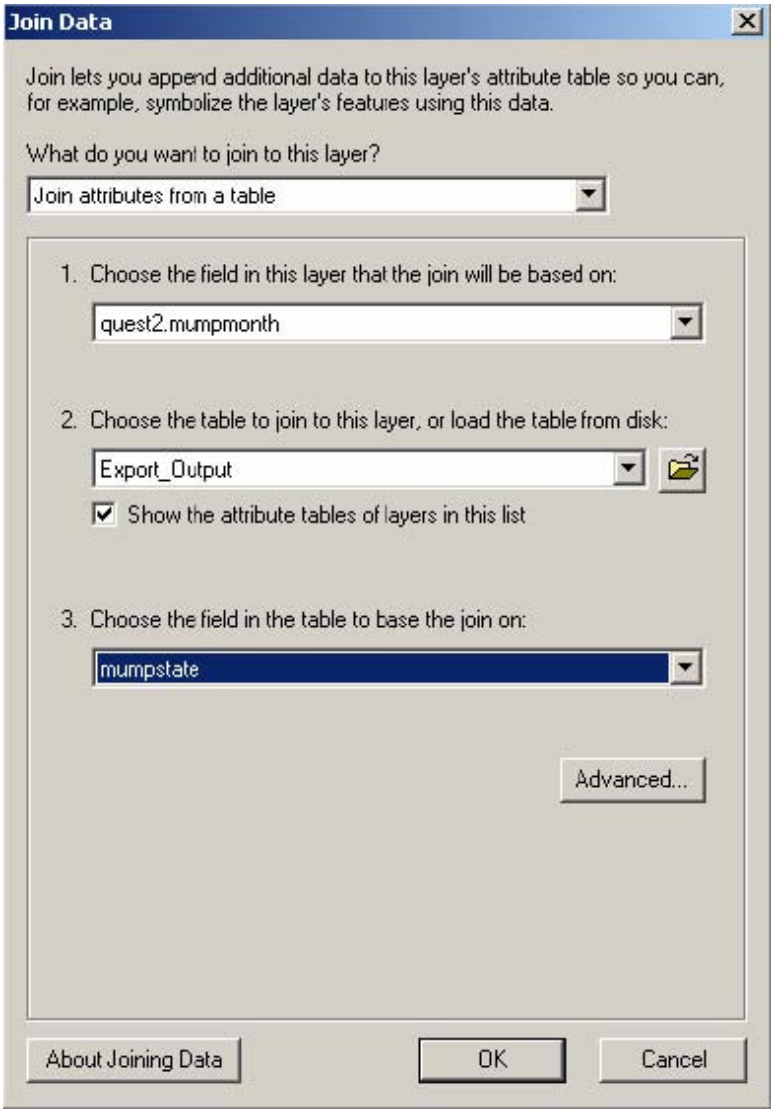
names of particular fields to bring them down into the bottom space for the formula and you'll need to type in the '100' (it's probably best to leave the *parser* setting to *vb script*).

Task:
Now try creating a new field called **mumpstate** in your **question2** table. See if you can make up composite numbers representing both **maximum_mumps** and **state** in the way we did here.

Join your tables together

We are now in a position to join our tables together:

- Right-click on the **question2** table and choose *joins and relates*, then choose *join…*



- Choose **mumpstate** as *1. the field in this layer that the join will be based on*
- Next to *2. Choose the table to join to this layer*, choose **mumpsdb**
- Finally, next to *3. choose the field in the table to base the join* on, choose **mumpstate** again
- Hit *OK* and see what happens! Take a look at the resulting data file by right-clicking on **question2** and choosing *open*.

Interpreting the results

We now know the year and month when the peak incidence of mumps cases occurred and we could produce a map of this if we had a file of state

boundaries available.

Do you notice any patterns in the peak month of mumps incidence? What about the year? Again, take a moment to think about this before reading on.

If you try sorting your table by **month_**, you'll probably notice that none of the states had a peak mumps incidence between July and September. The peak tends to be in the winter.

Try sorting your table by **year_**, and you'll probably notice that many states saw peak mumps incidence in 1968, the first year of the time series. This implies that mumps control is improving, but notice that there are a few states where the peak month occurred in the 1980s – more worrying!