*SEER*CMapper / Class Separability Classification Tool (CSCT) v.4.2*

USER MANUAL

By

David Wong

George Mason University

dwong2@gmu.edu

August, 2021

Background

Mapping data provided by the Surveillance, Epidemiology, and End Results (SEER) Program (https://seer.cancer.gov/) has several types of challenges. One methodological challenge is to create more truthful maps. Similar to many health datasets, each SEER estimate include its standard error, indicating its reliability. Typically, mapping of these health statistics and many statistical estimates using choropleth maps ignores the reliability of these estimates, treating the estimates as if they are perfectly accurate without error. The resultant maps likely put observations (areal units) with estimates not statistically different into different map classes. Similarly, observations with estimates that are statistically different may be put into the same class (Sun and Wong, 2010). Consequentially, resultant maps may be misleading, showing spatial patterns that may not exist, or concealing patterns buried in the data.

*SEER*CMapper*, also known as the *Class Separability Classification Tool* (*CSCT*), is an implementation of the class separability classification method (CSCM) for choropleth map to address this mapping issue (Sun, Wong and Kronenfeld, 2015). Different from a typical GIS or choropleth mapping package, *SEER*CMapper* can generate a choropleth map with a confidence level (CL) associated with each class break value. This CL level can be 1) approximately the minimum probability that any pair of values above and below the class break value are statistically different, or 2) the average probability of all pairwise comparisons of values above and below the class break value. In other word, this CL indicates how effective this class break in separating estimates that are statistically different into different classes. If this CL is low, estimates above and below that specific class breaks are not highly separable, and vice versa. In other words, that class break is not a desirable class break. However, CL is entirely determined by the quality of the estimates. If estimates are highly unreliable (with large standard errors), then the CL cannot be relatively high. Please refer to Sun et al. (2015) for the detail of the CSCM.

*SEER*CMapper* computes a confidence level for each potential class break value in the background process. Logically, class break values with relatively high probabilities (CLs) should be selected. If only a small number of classes are needed, the CL values for the small number of class breaks may be relatively high (again, the quality of estimates determines the CL values). When more classes are needed, the CL values of subsequent classes become lower. Therefore, in *SEER*CMapper*, users will use the minimum acceptable CL level indicated by a slider bar to determine the number of classes (and subsequent class break values). Unfortunately, choosing class break values with the highest probabilities often results in highly uneven map classes, very much due to the general empirical distributions of data. Therefore, users may want to adjust the class breaks to obtain more balanced classes. *SEER*CMapper* allow users to manually and heuristically adjust the class break values to derive more desirable classification (Sun et al., 2017). The CL value provided for each class break value will aid users to evaluate and manipulate class breaks.

*SEER*CMapper* can be used to map SEER estimates and many other datasets for estimates with error or data quality information (e.g., ACS), as long as the data include the additional column-field indicating either the standard error (SE) or the margin of error (MOE) of the estimates. However, *SEER*CMapper* has the specific functions to handle estimates with standard errors downloaded from *SEER*Stat* (https://seer.cancer.gov/seerstat/). Data downloaded from *SEER*Stat* can be ingested directly into *SEER*CMapper* to create a choropleth map. New in version 4, *SEER*CMapper* can accept tabular data other than the formats downloaded from *SEER*Stat* and ESRI shapefiles. Supported data formats include csv, xls (only Excel 97-2003), dbf, and Access. Users can manipulate map classification with the CL indicated for each class break. The resultant map can be exported into a graphic file to be used for reporting.

To summarize, *SEER*CMapper*:

- Is a stand-alone mapping program that does not require GIS software or additional GIS data.
- Can directly ingest data downloaded from *SEER*Stat*, including the *National Program of Cancer Registries* (NPCR) datasets.
- Can create maps using the CSCM, modifying class breaks determined by CSCM, and using other popular classification methods (natural breaks, quantiles, and equal intervals).
- Provides a confidence level for each class break, regardless how the class break is determined.
- Produces maps at the state or county level, depending on the input SEER data.

- Can perform all the functions above, using shapefiles as the input, and state- and county-level estimates in csv, xls (not xlsx), dbf and access formats without user providing state and county boundary data.

Requirements of Using *SEER*CMapper*

There are several system and user requirements to use *SEER*CMapper*:

1) This is a stand-alone tool developed in Java. The previous version (3, released in 2018) required Java Virtual Machine (JVM) or Java Runtime Environment (JRE) to be installed. In the current version, JRE is "packed" with the program files and therefore no separate installation of JRE is required.
2) The zipfile downloaded from this website is about 150 MB. When the file is unzipped into a folder, the folder will be approximately 300 MB. Therefore, it is recommended to have at least 450-500 MB of disk space to accommodate SEER*CMapper.
3) Displaying the resultant choropleth map and the associated interactive graphics user interface (GUI) elements require relatively large graphics memory for large datasets (county-level). Therefore, it is advisable to use a computer-laptop that has a separate graphics card with 4MB or larger memory. Graphics memory below this level may still work for less complicated maps (state-level maps), but the display may be slow or even incomplete.
4) *SEER*CMapper* has several windows. Therefore, larger monitor will be preferable to provide enough space to display multiple windows simultaneously.
5) Users intended to create choropleth maps of SEER data should be a (registered) user of *SEER*Stat*. It means that the user should have a *SEER*Stat* account and already has access to the SEER database (and have the *SEER*Stat* program already installed).
6) Users should have a basic operational knowledge to access and download data from the SEER database. If you are **new** to *SEER*Stat* or do not know which dataset(s) you may want to download from SEER, please visit the SEER website for tutorials. (https://seer.cancer.gov/seerstat/ and https://seer.cancer.gov/seerstat/tutorials/)

Download SEER data from *SEER*Stat*

The *SEER*CMapper* tool expects the data downloaded from SEER meet certain format requirements. Therefore, it is important to follow the procedure below to download and format data from the SEER website accordingly. After the data are downloaded, users

can launched the *SEER*CMapper* to map the downloaded data. Note that the mapping program uses only rates data with standard error (SE) at the county and state levels. This document and procedure are prepared under several assumptions:

- Data to be mapped are only at the state **or** county (county-equivalents: parishes for Louisiana and boroughs in Alaska) level. Other administrative and statistical units (e.g., incorporated cities, metropolitan, SEER registry areas) are not supported.
- County-level boundary data are based on the county definition adopted in the 2010 Decennial Census.

1) After launching *SEER*Stat* on the desktop, under the "File" manual, choose "New" and "Rate Session." (You may be asked to login at this point – please login to proceed).

2) Click the "Data" tab:
   - Select a dataset
   - You can choose either an "Incidence" or "Mortality" dataset.

3) Then click the "Statistic" tab:
   - Inside the "Statistic" box on the left, choose either "Rates (Crude)" or "Rate (Age-Adjusted)."
   - If "Rate (Age-Adjusted)" is chosen, be sure the relevant parameters are provided.
   - Check the box for "Show Standard Errors and Confidence Intervals"
   - Checking the box for "Use Tiwari et al., 2006 modification for CIs" will not affect the mapping program.

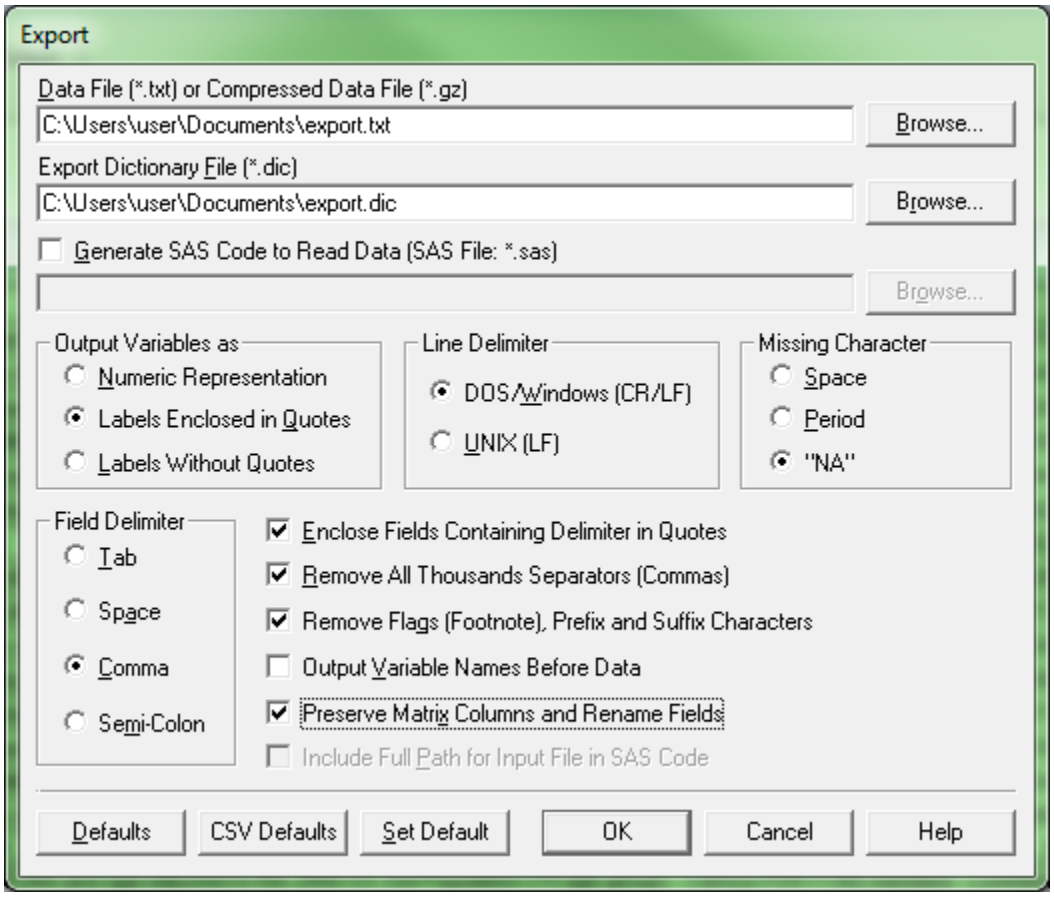
4) Then click the "Table" tab:
   - Under "Available Variables" window (lower panel), expand the tree with "… State, Cnty" or "… Registry, County". Note that "State" and "County" are often listed with other variables such as "Race, Sex, Year Dth/Dx".
   - Select one of the geographical levels (**do not include more than one level**):
     - Select "State" for tabulation at **state** level.
     - Select "State-county" for tabulation at **county** level. (Do not select just "county")
     - Select "Registry/county" for tabulation at the **county** level if "State-county" is not available.
   - Click "Row" on the right to add the selected geography to the Row variable.

- You may add other **non-geographical** variables to the columns to be mapped.
- Note that this program is designed to map rates. Therefore, it is **not recommended** to add many non-rate variables or a large number of variables not intended to be mapped.

5) Click "Execute" (the lightning icon) to generate the table.

6) After the table is displayed on screen, click the "Matrix" menu tab, choose "Export" and "Results as Text File." A form similar to the one below (Figure 1) will appear.

Figure 1: Export window in *SEER*Stat*.



- Provide a filename for the Data File (e.g., output1.txt).

- Use the SAME name for the Dictionary File (e.g., output1.dic). Make sure that both files are saved in the same folder.

- Then choose the following export parameters:
  - Output Variables as – select "Labels Enclosed in Quotes"
  - Line Delimiter – "DOS/Windows"
  - Missing Character – "NA"
  - Field Delimiter – "Comma"

- Check the following boxes:
  - Enclosed Fields Containing Delimiter in Quotes
  - Remove All Thousands Separators (Commas)
  - Remove Flags (Footnote), Prefix and Suffix Characters
  - Preserve Matrix Columns and Rename Fields

- Then click "OK"

* Because "Preserve Matrix Columns and Rename Fields" is checked, field names will be replaced (truncated) in the shapefiles created by *SEER*CMapper*. This is necessary because many field names from *SEER*Stat* are too long to be accommodated by shapefiles. However, when the .dic file is used to create choropleth map in *SEER*CMapper*, the original field names are shown in the Variable Picker. *

Download and Launch *SEER*CMapper*

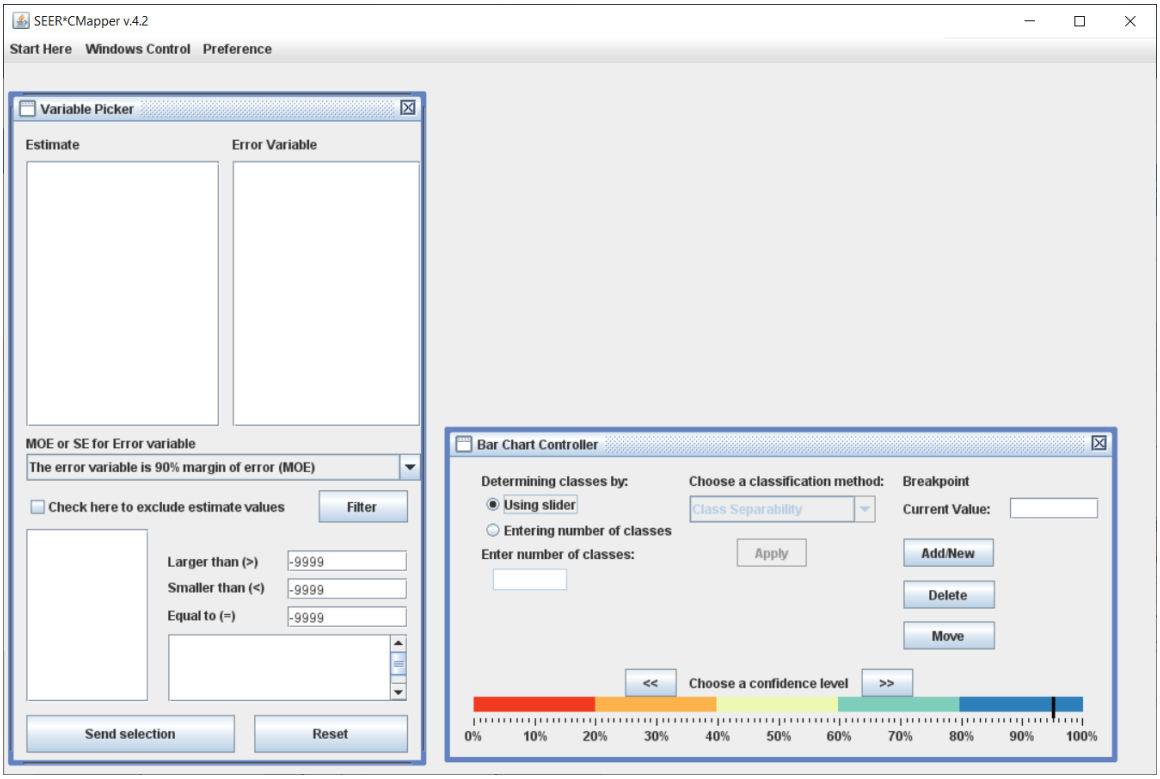
*SEER*CMapper* can be downloaded from http://geospatial.gmu.edu/seer-cmapper. Note two versions, v.3 and v.4 are available and this user manual, which is for v.4 can also be obtained from the above website. Click the link to download the compressed (zip) file to the folder where the user would like to store and access the mapping tool (such as "C:\…\SEERCMapper" where "SEERCMapper" is the folder.)

* Note that this folder SHOULD NOT be the folder where you store your typical download files or installation programs, as this downloaded file is not an installation program. *

After the file is downloaded, unzip (or "extract") the content of the file to a folder. The folder should have two sub-folders: "jre" and "mappingtool_lib." Besides these two folders, there should have two files: mappingtool.jar and run.bat. If Java (JRE or JVM) is installed, the mappingtool.jar will be an executable file and clicking it will launch SEER*CMapper. With no Java installed, double-clicking the "run.bat" file will launch the mapping program. Two default windows will appear. They are the "Variable Picker" and "Bar Chart Controller." (Figure 2)

* Note that at the bottom of the Controller is a 5-color bar corresponding to different percentages and a movable cursor (a black vertical marker) is on the right. The percent refers to the confidence level of class separability. *

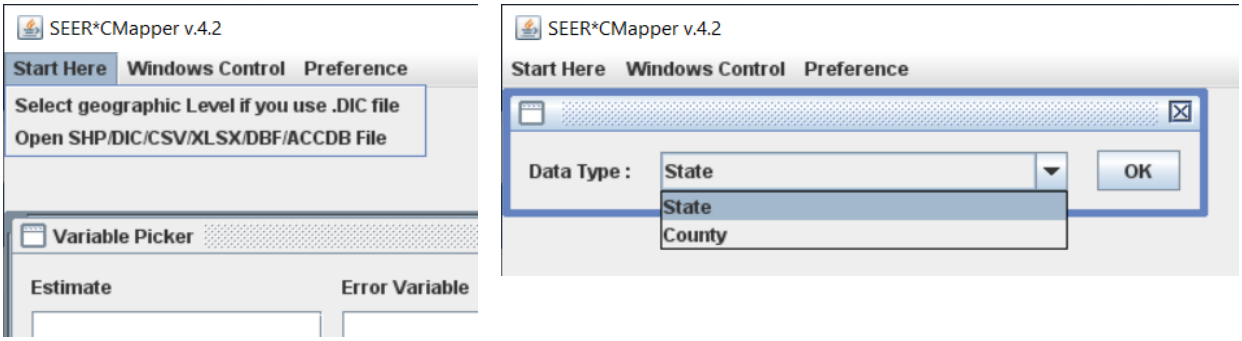
Figure 2: The interface of SEER*CMapper



Creating choropleth maps using *SEER*CMapper*

Import data

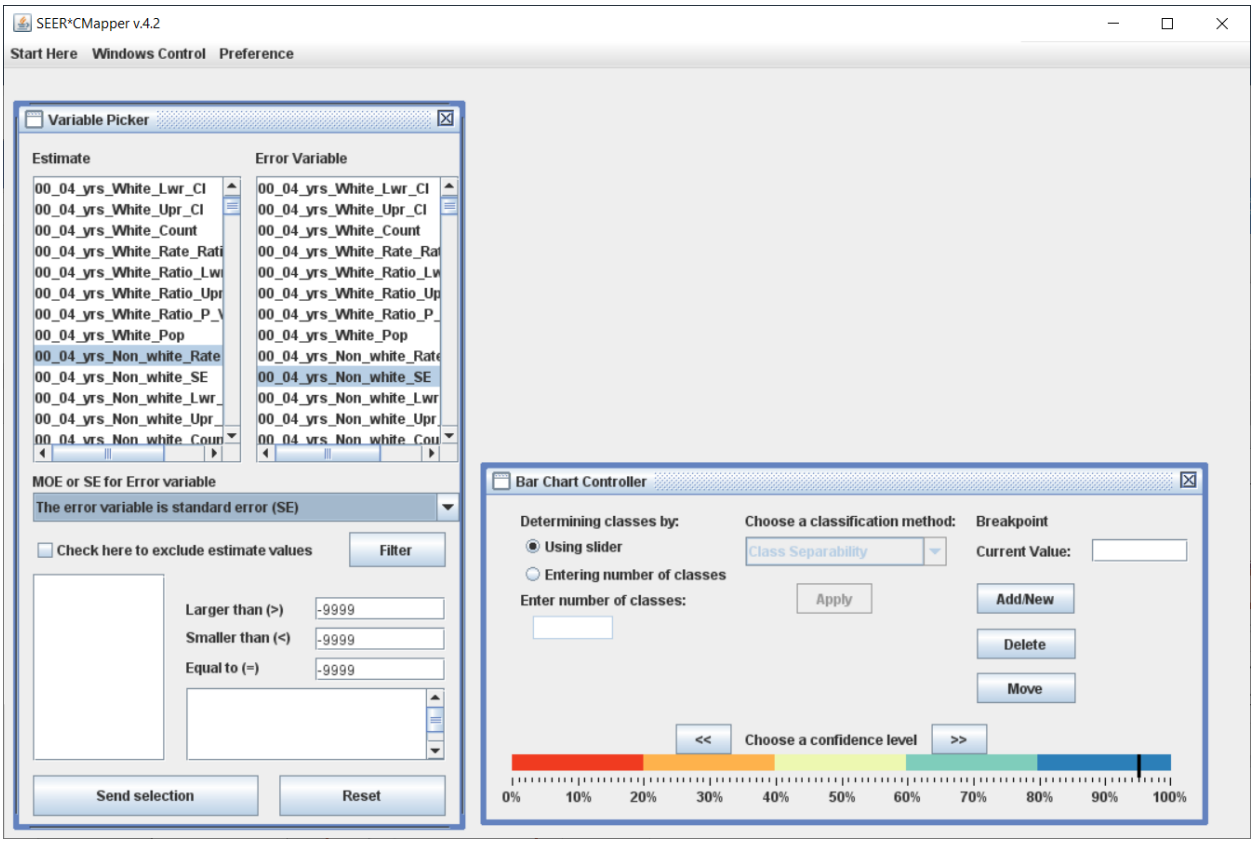
When launching *SEER*CMapper*, users should have downloaded rate data from *SEER*Stat* according to the format prescribed above in this manual. Data downloaded from *SEER*Stat* should include two files with .dic and .txt as the extensions. To bring these data files into *SEER*CMapper*, users need to indicate if the data are at the state or county level by clicking the "Start Here" menu. A dropdown menu will appear to allow users to select geographic level (figure below: left). After clicking to select geographic level, a dropdown window will appear with state and county as the choices (figure below: right). Make the appropriate section here. After selecting the geographic level, go back to the "Start Here" menu and choose "Open SHP/DIC/CSV …" to import the .dic files and files other tabular data formats without providing boundary data. If shapefiles are used, the previous step to select "geographic level" is not required.



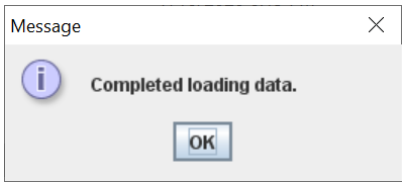
In the "Open" file window, make sure that the "Files of Type" is "Dic files." Then navigate to the folder where the SEER .dic file is stored and select the .dic file to open. Then the data in the .dic and corresponding .txt files will be loaded into the "Variable Picker" window. The column labels in the SEER data will be displayed identically under the two columns of the Variable Picker window. (Figure 3)

Variables selection and filters

Figure 3. Column labels in SEER data are ingested into the Variable Picker window.



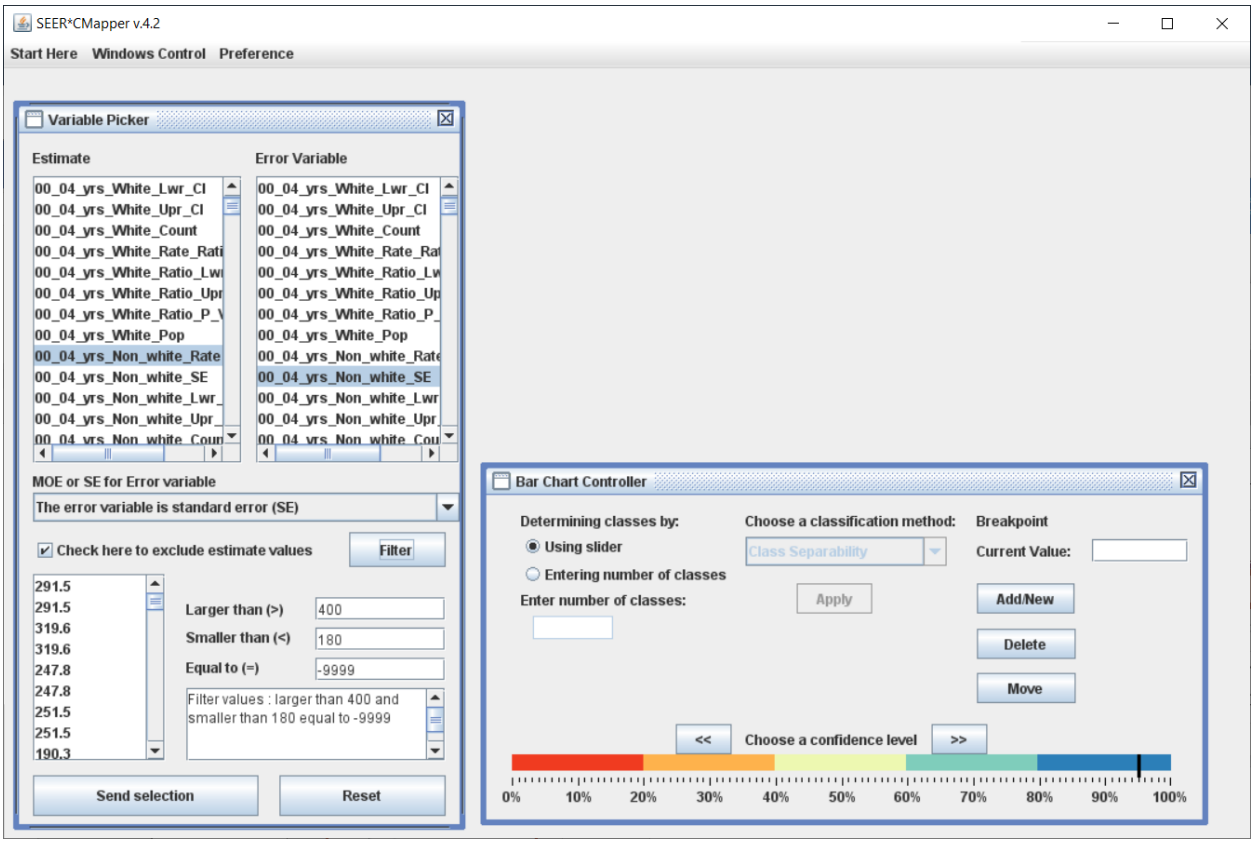
In the Variable Picker window, please choose the column label for the estimate (left column of the Picker window) and choose the column label for the error variable (right column of the Picker window). After choosing these two columns, the bottom of the picker window should specify the type of error variable. *SEER*CMapper* accepts 90% MOE, 95% MOE, and standard error (SE) as the error variable. In SEER, the error variable is standard error. Therefore, the specification of the error variable should be switched to "The error variable is standard error (SE)." Then click the "Send selection" button. After clicking "Send selection", a pop-up message "Completed loading data" (see below) will appear when the data are ingested into the system. Depending on the speed (and RAM) of the computer, slower computer and larger datasets (i.e., SEER datasets with many columns and county-level data) may take longer time.



In Figure 3, it shows that "00_04_yrs_Non_white_Rate" is the chosen estimate on the left column, and "00_04_yrs_Non_white_SE" is the chosen error variable on the right column (these are variables derived from *SEER*Stat* for "Mortality - All COD, Aggregated With State, Total U.S. for Hispanics (1990-2000) <18 Age Groups> " – these files are available as "Sample Datasets" in the tool website).

A new feature in the *SEER*CMapper* v.4 is the filter. In the lower half of the Variable Picker window, users can check the box for "Check here to exclude estimate values" to execute the filter. When this box checked, values of the selected estimate will be loaded into the window below, according to the order of the records in the dataset. Users can browse the values to decide if certain values or value ranges need to be excluded from the analysis. Assume that we do not want to include the extreme values of the "00_04_yrs_Non_white_Rate" in the analysis, we can remove values smaller than 180 and larger than 400. By putting "400" in the box next to "Larger than(>)" and "180" next to "Smaller than (<)", and then click the "Filter" button, the full query strings will appear in the box below and users can click the "Send selection" button to submit the data.

Figure 4. Variable Picker window with filter functions



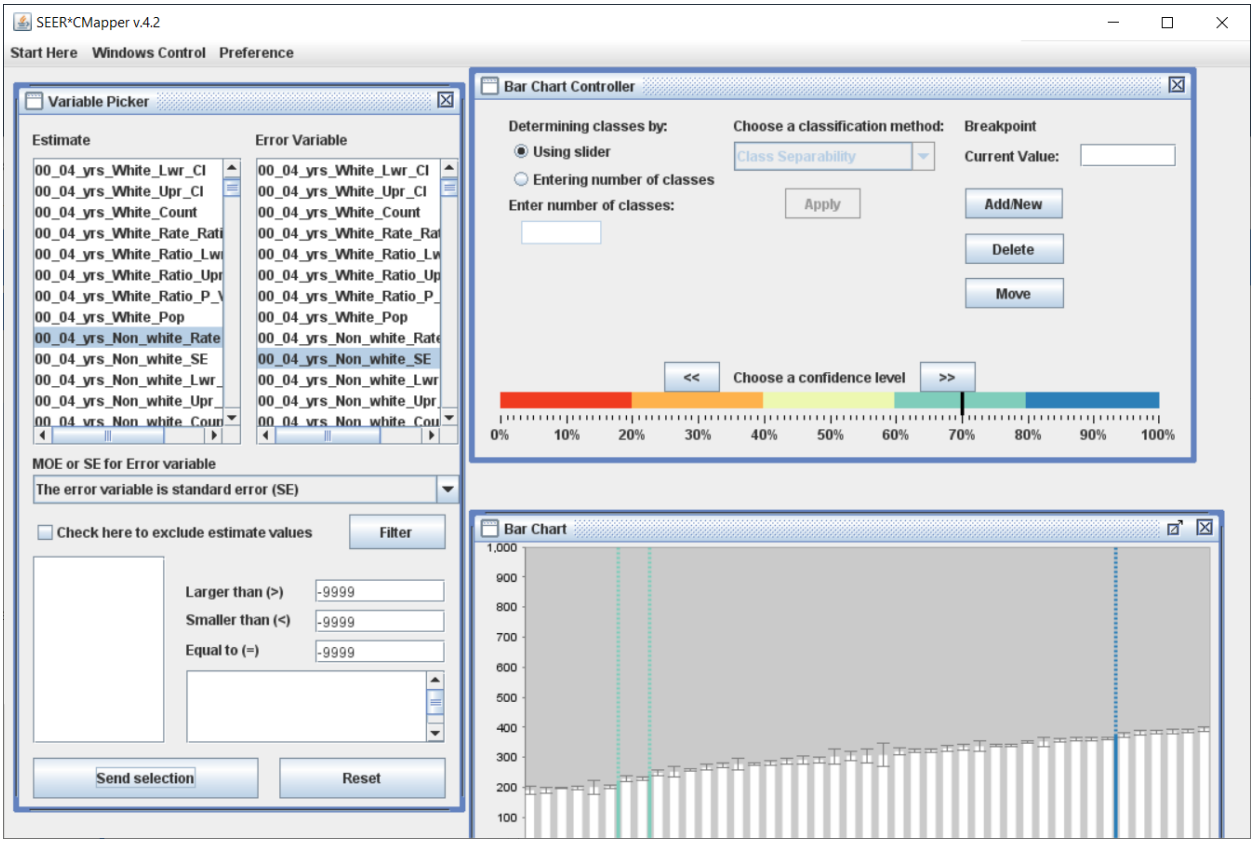
Classifications and determining class break values

After the variables are selected and data are sent, a "Bar Chart" window will appear with vertical bar plot. Alternatively, the "Bar Chart" window may be invoked by clicking the Windows Control menu and check "Chart." In the bar chart window, vertical bars indicating the estimates and the error bars at the tips of the vertical bars indicating the standard error will appear (Figure 5). A bar of different colors is displayed along the x-axis. These colors correspond to the colors on the 5-color bar of the "Bar Chart Controller" window. These colors reflect the confidence levels if class breaks are inserted respective locations. The cursor (the black vertical marker) on the 5-color bar in "Bar Chart Controller" window can be dragged by users to indicate the minimum confidence level or separability that a user is willing to accept for all class breaks.

As shown in Figure 5, by dragging the cursor to 70%, the minimum separability accepted by a user, one dark blue line and two light blue lines are inserted onto the bar chart. These lines are the resultant class breaks using the CSCM. The class break indicated by the dark blue line has a separability level of 80% or higher while the two

breaks indicated by the light blue lines have separability levels between 60% and 80%. Users can tell that these line colors correspond to the colors on horizontal color bar on the bar chart. In other words, colors on the horizontal bar indicate the separability level of each potential class break. Users can move the cursor along the color bar in the controller to experiment different levels of separability and the resultant classes. The higher the acceptable separability level (CL), fewer classes are determined. By lower the acceptable separability level, more class breaks can be identified.

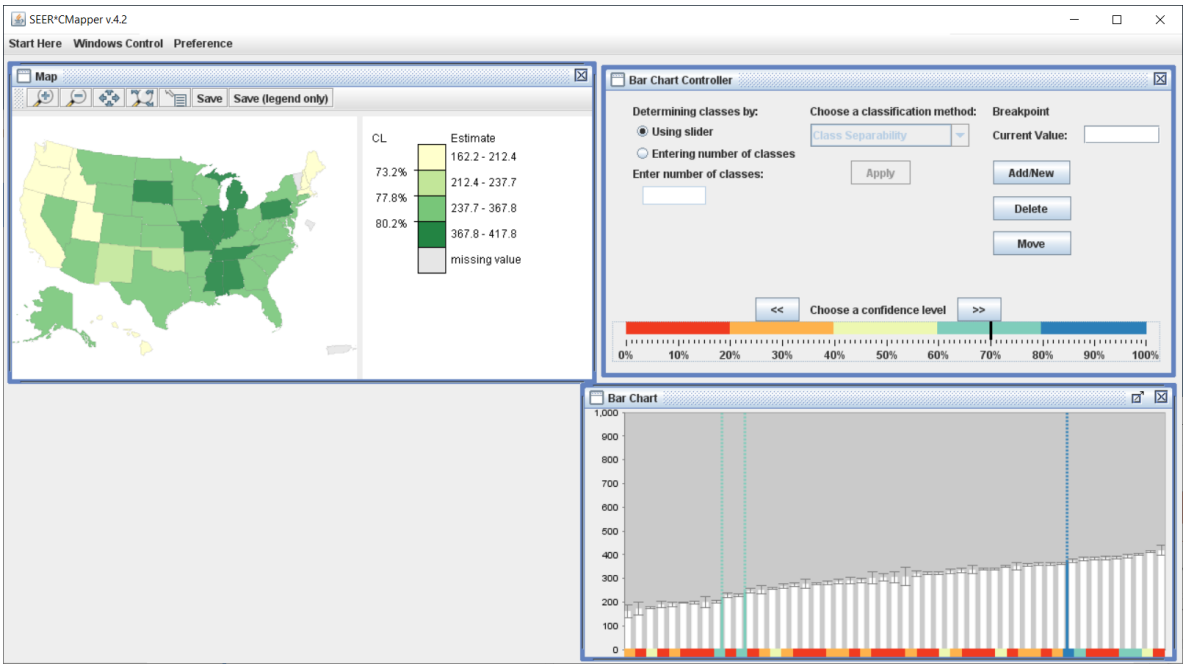
Figure 5. Bar chart with inserted class breaks by dragging the cursor to the desirable separability level.



To visualize the resultant classes on a map, under the "Classification Window" menu, check "Map" and a choropleth map using the class breaks determined via the bar chart will be displayed in the Map window (Figure 5). Note that this map has a special legend design with a probability - a confidence level (CL) attached to each class break on the left side of the legend. In this specific example, estimates between the first class [162.2 - 212.4) and estimates in all other classes are different at least 73.2% of the time. The class ranges include the lower bound values but exclude the upper bound. Similarly, estimates between the first two classes and the last two classes are different at least

77.8% of the time. These two class breaks are indicated by the two light blue lines in the bar plot. Given the data here, the highest separable classes are between the last class and all other classes (80.2%). This class break is indicated by the dark blue line in the bar plot. Because the cursor was dragged down to about 70%, only class break values with a separability level (CL) of 70% or higher will be used. If more classes are preferred, the additional classes will have separability levels lower than 70%.

Figure 5. Choropleth map with a class separability legend generated by the *SEER*CMapper*.



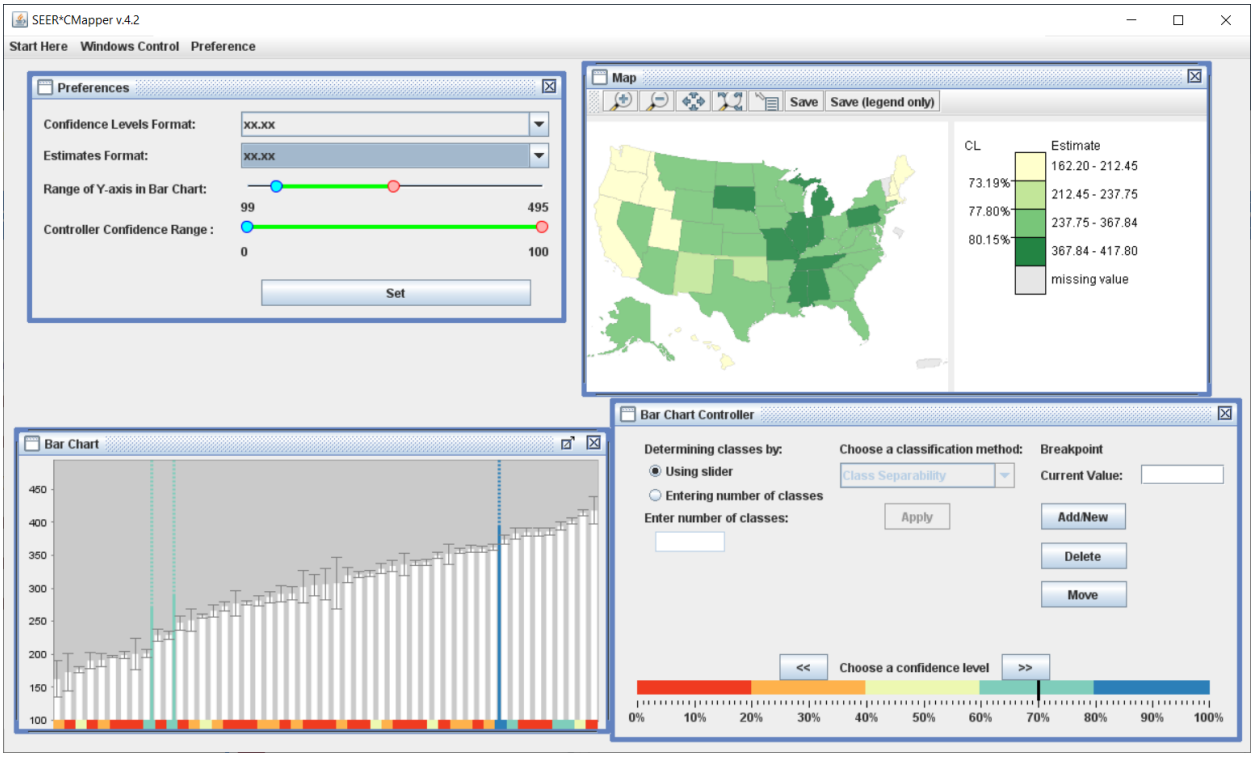
* Note that although the bar plot will show more than nine classes of class break, the map display will support up to nine classes only, as studies indicate that most people cannot handle more than nine classes in a choropleth map. When more than nine classes are determined in the bar plot, a warning message will appear in the Bar Chart Controller window and the map will not be shown. To force a map to display, the number of classes has to be reduced to nine or less. To do so, check the radio button next to "Entering number of classes" under "Determining classes by." Then enter a value of 9 or less for the number of classes. Make sure that the classification method is "Class Separability" before hitting the "Apply" button. Then a map with the highest separability levels (CLs) will be produced. *

The map window, through the menu icons, supports several standard functions: zoom-in and out, pan, maximum extent, identify, save, and save (legend only). The "Save" tool can save the map and legend as one graphics file (in .png format), and the "Save (legend only)" tool save the legend only, not the map.

Controlling display elements

In SEER*CMapper v.4, several new features were added to allow better control of display elements. A new menu item "Preference" was added to the user interface (after version 4), clicking it will allow users to access several controls. Figure 6 shows four elements under "Preference." The first one is "Confidence Levels Format." On the Map window, the legend adopts a new design with the confidence level (CL) of each class break value on the left. The CL as percent can have several decimal formats: no (xx), one (xx.x), two (xx.xx), or three (xx.xxx). Users can select a preferred format. The default is one. Figure 6 shows the CL with the two-decimal format.

Figure 6. Display control elements under the Preference menu item, SEER*CMapper



Similar to the first control on confidence level format, the second display control feature is the decimal place of estimates as displayed on the right-hand side of the legend: no (xx), one (xx.x), two (xx.xx), or three (xx.xxx). The default is one. Figure 6 shows the results of rounding to two decimal places.

The third display control feature is the y-axis of the vertical bar plot. Users can decide the range of values to be displayed on the y-axis, partly to magnify the variation across observations. Without adjustments, the y-axis has a range of 0 to 1000 despite all values

falls between 150 and 450. Figure 6 shows that by reducing the range to 99 and 490, the vertical bar plot shows the variations across observations much more clearly.

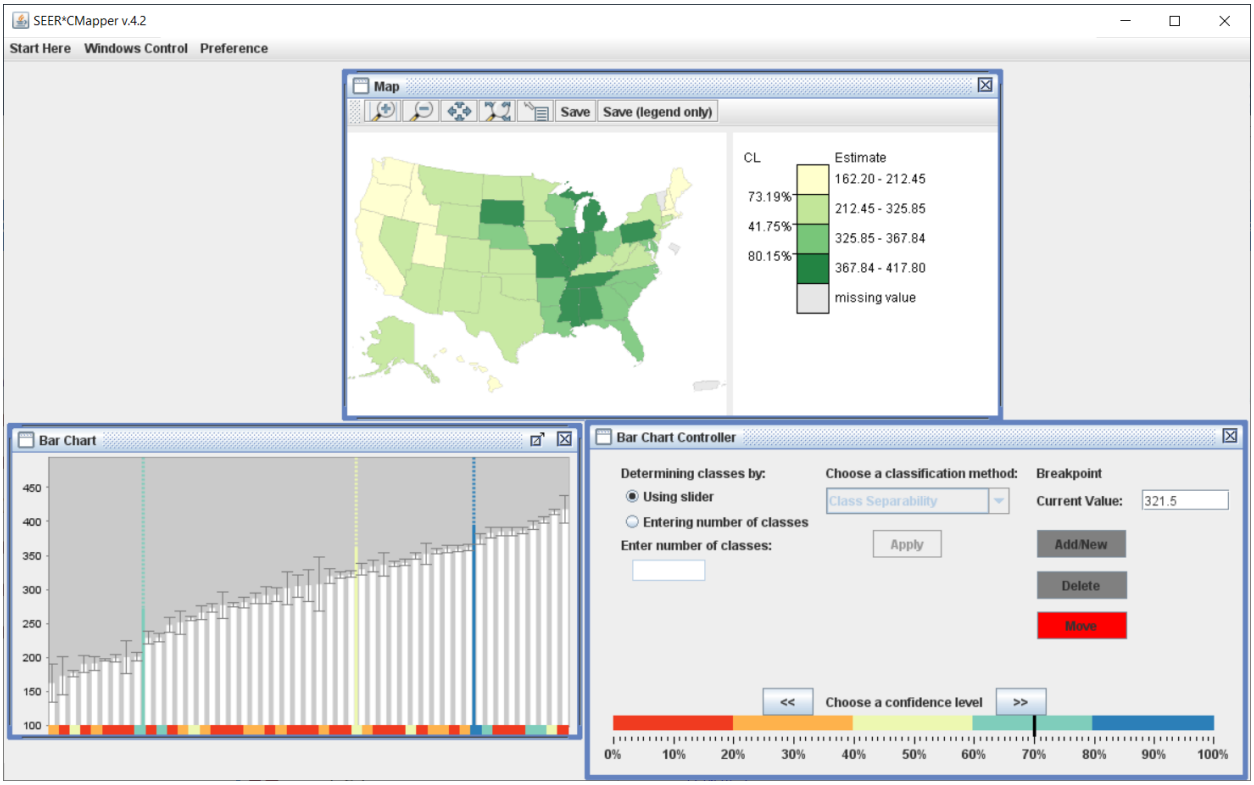
The fourth display control feature is to adjust the confidence range on the Bar Char Controller. The default for the 5-color bar covers 0% to 100%. Users may reduce this range by controlling the minimum and maximum values. Although this feature is operational, we do not recommend changing the default range unless it is really desirable, as changing the range sometimes may cause unnecessary complications.

Adjusting Class Break Values

Class breaks determined through the CSCM process are the most separable, given the minimum separability (CL) accepted by the user. In other words, these class breaks are the most reliable in assigning estimates that are statistically different into different classes. However, depending on the purposes of the mapping exercise, these classes may not be highly desirable according to other map classification criteria. For instance, due to the distribution and error levels of data, classes determined via CSCM are usually not very balanced. One or two classes will have a large number of observations and a few classes have very few observations. Therefore, adjusting the class breaks may be necessary.

In *SEER*CMapper*, there are several ways to adjust class breaks. Assuming that the class breaks determined through the CSCM procedure are not desirable, just like the example in Figure 5 that the two break values are too close (the two light blue lines in the bar plot), *SEER*CMapper* allows users to delete, add, and move class breaks through the bar plot. In the "Bar Chart Controller" window, three buttons allow users to perform such manipulations. For instance, it may be desirable to move one of the light blue lines toward the middle section of the plot as there is no class break for a large number of observations. Then, users can click the "move" button (the button will turn red), and move the mouse cursor to select the light blue line on the right (Figure 7). When the class break line is selected, it will turn to black. At this point, when the cursor moves across the plot, a line following the cursor will appear intermittently. When the mouse clicks again, the selected class break line will be moved to the new position. The question is where should the new class break be?

Figure 7. Moving a class break from a high separability level (light blue) to a moderate level (yellow) in the middle section of the bar plot.



Ideally, the new class break should have a relatively high separability level. But all potential breaks with higher separability levels have been identified already. Any other value chosen as the class break will have a lower separability level. Still, if giving up a highly separable class break is desirable, the alternate class break should still have a separability level as high as possible. To search and select the next most desirable break value, users may refer to the horizontal color bar (x-axis) in the bar plot where the colors correspond to different levels of CL in the Bar Controller window. With no light blue in the middle section, the next best choice is yellow with a moderate separability level (Figure 7). Note that when the class break is moved, the map display is also refreshed to reflect the change. The new class break has only 41.7% of CL. The other class break manipulation methods (add/new and delete) operate in a similar manner. First, select the button in the Controller window and then apply the method onto the bar plot.

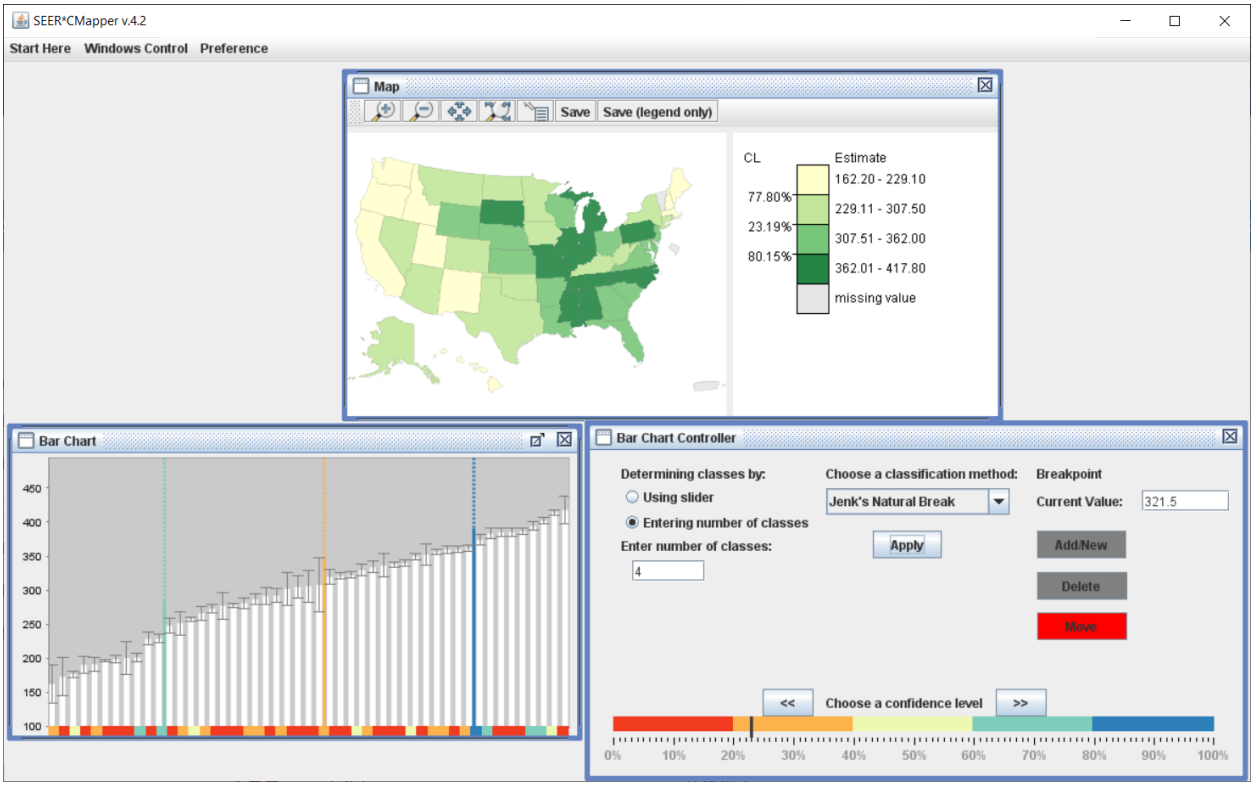
Evaluating the Reliability of Map Classes

*SEER*CMapper* not only can be used to determine class breaks using the CSCM, it can also be used to evaluate the reliability of map classes determined by other popular

classification methods. *SEER*CMapper* includes several popular map classification methods to create choropleth maps. These methods are natural breaks, equal interval and quantile. To use any of these classification methods, the button for "Entering number of classes" under "Determining classes by" should be checked (the default is "Using slider"). Then select a classification method and enter the desirable number of classes (again, it is limited to nine). Clicking the "Apply" button will refresh the displays of the bar plot and map.

In Figure 8, the natural breaks classification method was chosen with four classes. Both the bar plot and the map displays reflect the new classification result. In this specific example, the two most separable classes determined by the natural breaks method are the same as those determined by CSCM, but the remaining class break using the natural breaks classification method has a separability level of only 23.19%. Thus, *SEER*CMapper* can be used to evaluate the reliability of classes determined by popular map classification methods.

Figure 8. Resultant map using the natural break classification.



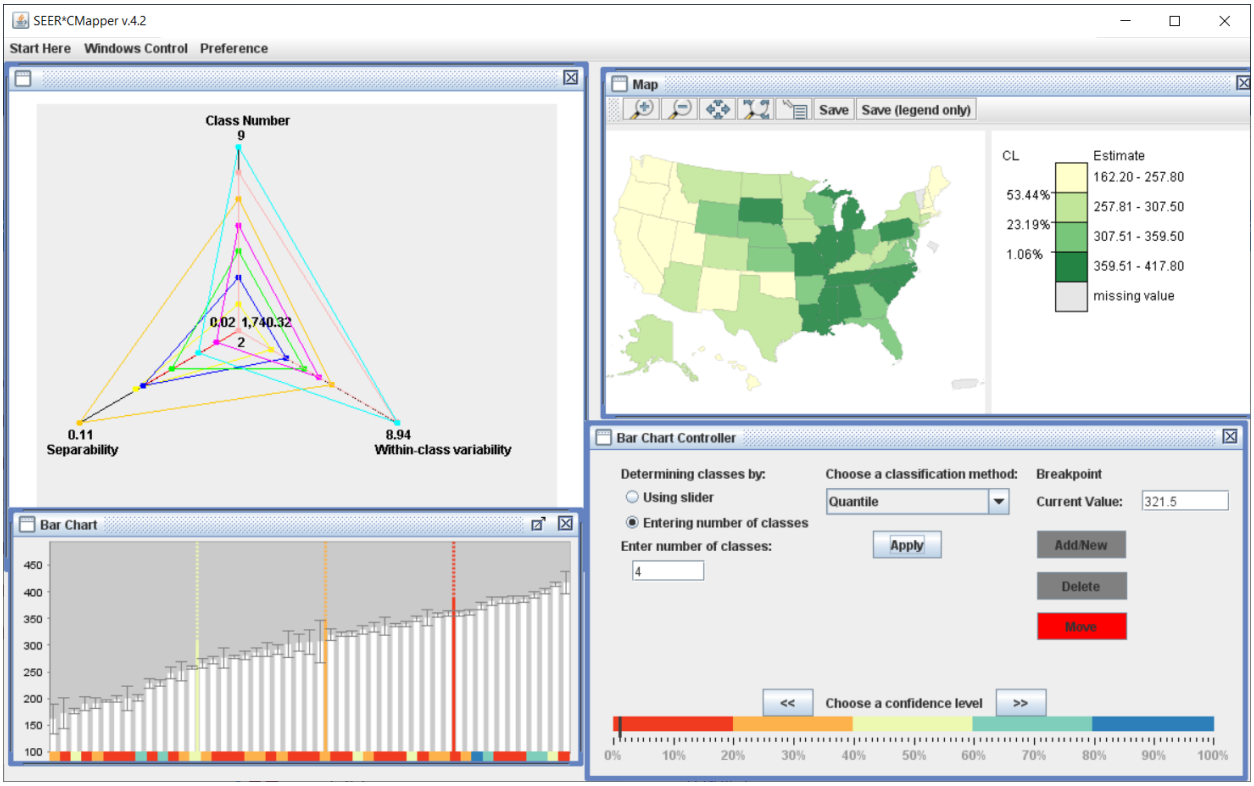
Heuristic Adjusting Class Breaks

The number of classes, separability levels of class breaks, and classification methods are interrelated. Using a specific classification method, the separability levels of class breaks cannot be adjusted freely, but are predetermined, given the number of classes.

Therefore, *SEER*CMapper* also includes a tool for users to review these relationships in a graphical manner and the graph also serves as an interface to select a specific map classification scheme.

Under the Window Control menu is an item for "Star Plot." Clicking this menu item will display a star plot showing the relationship between the number of classes, averaged separability level, and within-class variability. Within-class variability refers to the variation of estimates within a class. It is essentially the standard deviations of estimates within each class summing across all classes. The details of these criteria can be found in Sun et al. (2017). Figure 9 shows the star plot of choosing the quantile classification method of the same data set used above.

Figure 9. Star plot with three axes corresponding to three criteria to evaluate a map classification.



Note that the star plot will appear only when the classification method other than the separability method is selected under the "Choose a classification method" window. Also, as of now, the Star Plot supports only the quantile and natural breaks classification methods.

As the star plot is generated when quantile is the chosen classification method, the plot shows the averaged separability and within-class variability levels when different numbers of classes are used. Their relationships are shown by lines connecting the three

axes, forming polygons. Each polygon is a combination of class number, averaged separability and within-class variability levels. Class numbers and separability levels are inversed related as discussed above. Class numbers and within-class variability are usually correlated. The polygons (lines) in the star plot are clickable so that users can select a classification scheme based on these three criteria, and the map display will be updated by the chosen scheme. In other words, users can interactively select classification scheme through the star plot and visualize the resultant classification from the map window.

References

Sun, Min, and David W. S. Wong. 2010. "Incorporating Data Quality Information in Mapping American Community Survey Data." *Cartography and Geographic Information Science* 37 (4): 285–99. https://doi.org/10.1559/152304010793454363.

Sun, Min, David W. Wong, and Barry J. Kronenfeld. 2015. "A Classification Method for Choropleth Maps Incorporating Data Reliability Information." *The Professional Geographer* 67 (1): 72–83. https://doi.org/10.1080/00330124.2014.888627.

Sun, Min, David Wong, and Barry Kronenfeld. 2017. "A Heuristic Multi-Criteria Classification Approach Incorporating Data Quality Information for Choropleth Mapping." *Cartography and Geographic Information Science* 44 (3): 246–58. https://doi.org/10.1080/15230406.2016.1145072.