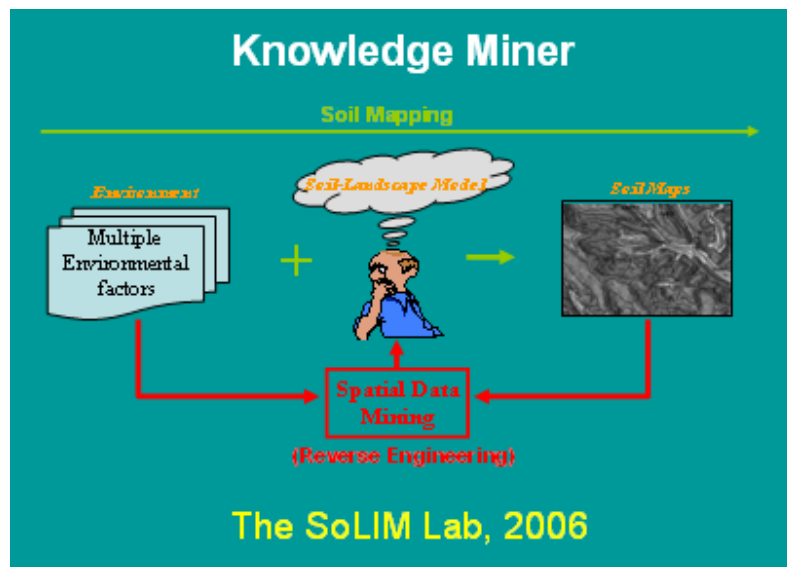


Knowledge Miner 1.0

User Manual



The SoLIM Lab
Department of Geography
University of Wisconsin – Madison
February, 2007

Table of Contents

1 Overview	4
1 Overview	4
2 Basic idea and methodology	6
2.1 Basic Idea.....	6
2.2 Methodology	7
3 Data Files	11
3.1 Environmental data layers.....	11
3.2 Soil survey file	11
3.3 Environmental data layer list file.....	11
3.4 Map unit list file.....	12
4 Knowledge Extraction	13
5 Knowledge Analysis	15
5.1 Initial knowledge curve.....	17
5.2 General pattern.....	18
5.2.1 Similarity distribution	18
5.2.2 Mode distribution.....	19
5.3 Outliers.....	20
5.3.2 Outliers in terms of low similarities.....	21
5.3.2 Outliers in terms of large mode distances.....	24
6 Knowledge Refinement	28
6.1 Automated refinement	28
6.1.1 Removing soil polygons with low similarity measures	28
6.1.2 Removing soil polygons with large mode distance measures	30
6.2 Manual refinement	31
6.2.1 Fitting the extracted knowledge curve.....	32
6.2.2 Adjusting the extracted knowledge curve.....	32
6.2.3 Saving knowledge.....	33
6.3 Knowledge Comparator	34
7 Use of the Knowledge.....	35
Appendix A: Tutorial.....	36

List of Figures

Figure 1 Basic idea of data mining	7
Figure 2 Histogram of an environment for a map unit	8
Figure 3 Normalized frequency distribution curve of an environment for a map unit	8
Figure 4 Construction of frequency distribution.....	10
Figure 5 An example of environmental data layer list file	12
Figure 6 An example of map unit list file	12
Figure 7 Knowledge extraction interface.....	13
Figure 8 Map unit and environment layer specification	15
Figure 9 Main interface.....	16
Figure 10 Knowledge curve legend.....	17
Figure 11 Initial knowledge curve	17
Figure 12 An example of similarity distribution.....	19
Figure 13 An example of mode distribution	20
Figure 14 Summary statistics.....	21
Figure 15 Potential outliers with low similarity measures.....	22
Figure 16 Looking into an individual curve	23
Figure 17 Similarity threshold for saving outlier polygon list.....	24
Figure 18 Potential outliers with large mode distance measures	25
Figure 19 Looking into an individual curve	26
Figure 20 Mode distance threshold for saving outlier polygon list	27
Figure 21 Similarity threshold and sample percentage for recomputing knowledge curve	28
Figure 22 Reconstructed knowledge based on similarity	29
Figure 23 Updated similarity distribution and mode distribution.....	29
Figure 24 Mode distance threshold and sample percentage for recomputing knowledge curve.....	30
Figure 25 Reconstructed knowledge curve based on mode distance.....	31
Figure 26 Updated similarity distribution and mode distribution.....	31
Figure 27 Number of control points.....	32
Figure 28 Manual adjustment of the knowledge curve.....	33
Figure 29 Knowledge comparator	34

1 Overview

In the United States, most cooperative soil survey activities are directed toward revision of existing surveys. Such surveys, which are typically decades old, can be thought of as spatial expressions of the surveyor's conception of soil-landscape relations. That is, based on field investigation, laboratory analysis, review of other surveys, etc., the scientist develops soil concepts (soil classes) and associates those concepts with geology, landscape position, slope, curvature, vegetation, and other environmental indicators that can be exploited in mapping. Obviously, those indicators which are directly tied to pedogenesis play a vital role in the soil-landscape model. Traditional survey practice was largely manual, with heavy reliance on stereo aerial photography for line placement (application of the model). There are two primary motivations for survey updates. First, limitations inherent in the manual process prevent totally consistent application of the model, regardless of how complete and well-conceived that model might be. Second, over time soils knowledge improves, and it is obviously desirable to incorporate that knowledge in a revised survey.

The existing surveys, though not perfect, contain a wealth of information that can potentially serve as a starting point for a revised survey. Although the scientist's soil-landscape model might not be explicitly documented anywhere, it is implicit in the survey. Our project is motivated by belief that if the model can be recovered from the existing survey, it can be used to jumpstart model development for the new survey. To that end, we have developed a set of data mining tools for extracting soil-landscape relations from a published survey. Map polygons are overlaid on raster data of various sorts, including a number of variables computed from a digital elevation matrix (DEM). The DEM-derived data include slope, aspect, planform and profile curvatures, wetness index, and other terrain indices. Together with geology and other non-DEM data, they define a suite of environmental variables identified by a soil scientist as likely to be important in the study area.

Our tools extract knowledge in the form of frequency distributions of pixels within map polygons. That is, for any map unit there is a distribution of elevation, slope, etc.

Knowledge Miner 1.0 User Manual

indicating the range of environmental conditions over which the unit has been mapped. By comparing distributions of one map unit with another, we obtain information about how the original surveyor chose to map those soils; that is, we find similarities and differences in the environments occupied by the soil map units. By comparing frequency distributions for soil polygons of the same map unit, we obtain information about the consistency of mapping, and can identify polygons that occupy anomalous environmental settings.

This user manual describes the basic idea, methodology, required data files, and usage of the data mining tools we have developed. We show how the mined data can be used to uncover soil-landscape models, and its utility in both revising soil concepts and suggesting the need for new concepts. We also show that after editing, the mined data can be used in SoLIM – a predictive modeling method – to produce an improved soil survey.

2 Basic idea and methodology

2.1 Basic Idea

The basic idea behind data mining is “reversing engineering” (Figure 1). During soil survey, empirical soil-landscape models are developed by inductive reasoning from field observations and these models are then applied to make predicative statements about the spatial distribution of soil types during soil mapping process. This can be illustrated as:

$$f + E \Rightarrow S \quad (1)$$

where E stands for soil “formative” environmental data; f stands for soil-landscape model (soil scientist’s knowledge about the relationship between soil and its “formative” environmental factors); and S stands for soil. Please note that we use the term “formative” to include variables that drive soil formation as well as factors that merely covary in a systematic way with soils and are exploited in soil mapping. In other words, we don’t necessarily imply formative in a causal sense. Hereafter we will use “formative” without the quotes. Data mining seeks to reverse the mapping process, i.e., given the soil formative environmental data layers and existing soil maps, we are trying to extract the soil-landscape model that was used in the mapping process. This can be illustrated as:

$$f \Leftarrow E + S \quad (2)$$

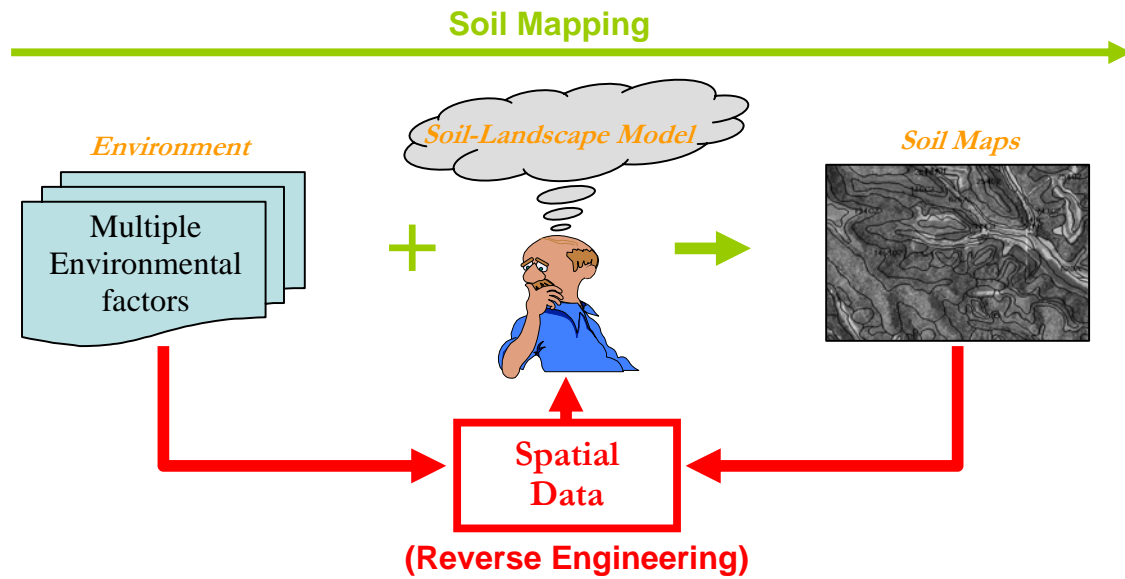


Figure 1 Basic idea of data mining

2.2 Methodology

Consider some soil, say soil “A”, and some environmental variable such as slope. The fundamental assumption of this project is that if soil A is seldom mapped for a particular range of slope values, the surveyor considered those slope values to be sub-optimal for that soil. Under that assumption, we can use the frequency distribution of occurrences to obtain information about the optimality of different values of an environmental variable. That is, for each value of an environmental variable, we will find the relative frequency of occurrence of soil A and use that distribution as an indicator of soil-environment model implicit in the map. Environmental factors can be categorical or continuous (please see section 3.3 for details). For a categorical environmental factor, the frequency distribution has the appearance of a histogram. This is illustrated in Figure 2, which shows the relative frequency for a number of discrete values of some hypothetical variable. For a continuous environmental factor, we use mapped occurrences to estimate

a continuous frequency distribution curve (Figure 3).

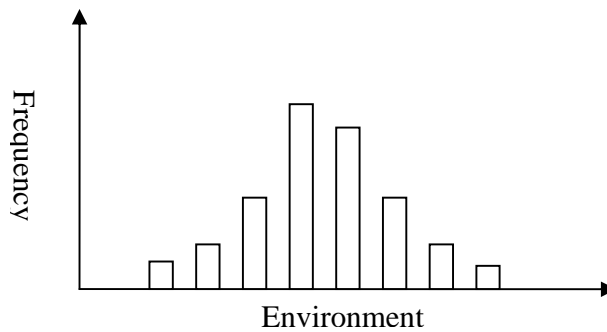


Figure 2 Histogram of an environment for a map unit

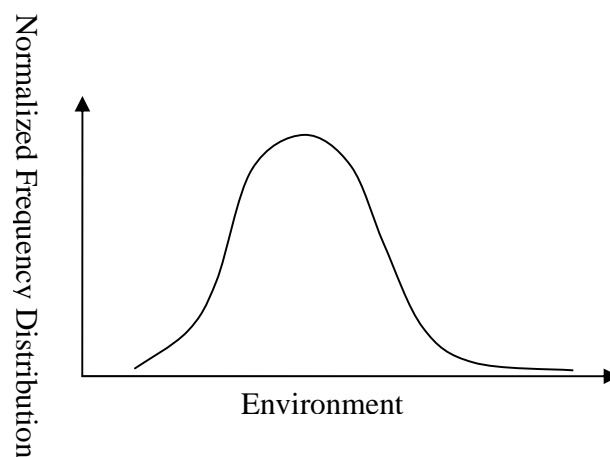


Figure 3 Normalized frequency distribution curve of an environment for a map unit

For the purposes of explanation we will use continuous variables as examples throughout this manual. Thus we will speak exclusively of a frequency distribution “curve”. The reader will understand that the ideas apply equally to categorical variables, for which the “curve” is actually a histogram.

Knowledge Miner 1.0 User Manual

In this manual, we use two different terms more or less interchangeably: frequency distribution and knowledge curve. Both refer to the frequency distribution curve we construct for continuous environmental data.

Soil maps are overlaid on various environmental raster data layers, including a number of variables computed from a digital elevation matrix (DEM). Then, knowledge in the form of frequency distributions of pixels within map polygons is extracted (Figure 4). That is, for any map unit and any soil polygon we calculate the distribution of elevation, slope, etc. indicating the range of environmental conditions over which that unit or soil polygon has been mapped. By comparing distributions of one map unit with another, we obtain information about how the original surveyor chose to map those soils; that is, we find similarities and differences in the environments occupied by the map units. By comparing frequency distributions for individual polygons of the same map unit, we obtain information about the consistency of mapping, and can identify polygons that occupy anomalous environmental settings. This can help both in revising soil concepts and suggesting the need for new concepts.

A frequency distribution curve is constructed using the kernel density estimation method. This is loosely analogous to the familiar histogram approach, in which each observation is treated as a point. If the point lies in a probability interval, the entire mass of the point, $1/n$, is added to the probability. Kernel methods, on the other hand, spread the mass of each observation around the observed value. The amount of spread or smearing is governed by a function called the kernel. The width of the kernel is adjustable – narrow kernels concentrate mass tightly around central (observed) values, whereas a wide kernel gives more smearing. The operation can be expressed mathematically as:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) \quad (3)$$

where $K(\cdot)$ represents the kernel function and h represents the degree of smearing, or bandwidth. Looking at the argument to K , we see that the distance between a point x and

each data value is expressed as a multiple of h . Because h has the same units as x , the ratio is dimensionless. This scaled distance is used in the kernel function to find the contribution of each observation. With K a decreasing function of distance, the contribution of a point far from x will be small, whereas nearby points will contribute more. The sum of all contributions divided by nh becomes the density estimate. Division by h is required to make $\hat{f}(x)$ a density, with dimensions of probability per unit X .

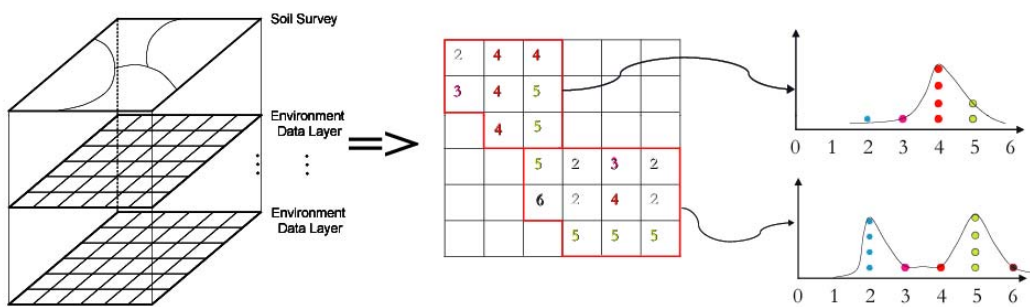


Figure 4 Construction of frequency distribution

Kernel methods are preferred because they are flexible, non-parametric, and they deliver a smooth density function (assuming K is suitably chosen).

3 Data Files

3.1 Environmental data layers

Environmental data layers such as elevation, slope gradient, slope aspect, planform curvature, profile curvature, topographic wetness index, etc., are required in 3dMapper 3dr format. The variables used are selected by the soil scientist according to his/her belief regarding what variables are likely to be helpful in separating soil classes from one another. Please check www.terrainanalytics.com for information on 3dMapper [Note from JB: the format is not described on the website---we have never published the format. I don't think it would help users of the tool if we were to do that].

3.2 Soil survey file

The existing soil survey needs to be in ArcView polygon shape file format and the attributes should contain at least two fields: Polygon ID and Map Unit Key. Polygon ID is a unique identifier of each soil polygon, and Map Unit Key is a unique identification of each map unit name. Polygon ID must be integer and Map Unit Key can be integer or string (no spaces). Please note that every polygon must be assigned a map unit. The map units need not be identical to those in the original survey. For example, the soil scientist might want to merge several map units into a single new unit. In that case all polygons of the original units would be given the same Map Unit Key. Though not essential, we recommend that users name these two fields as “Poly_ID” and “MU_KEY”.

3.3 Environmental data layer list file

This file lists the environmental data layers to be used to perform the analysis. It is a plain text file and can be created through the “File --> New Environment Layer List ...” menu item. The file format is:

Variable:	FileName	DataType
Variable:	FileName	DataType
... ..		

“Variable” is the label or tag for a data layer, “FileName” contains the full path of this data layer, and “DataType” indicates the data type of this data layer. “DataType” needs to be “1” (for continuous data type) or “2” (for categorical data type). “Variable”, “FileName”, and “DataType” are delimited using space or “Tab”. Figure 5 illustrates an example of an environmental data layer list file.

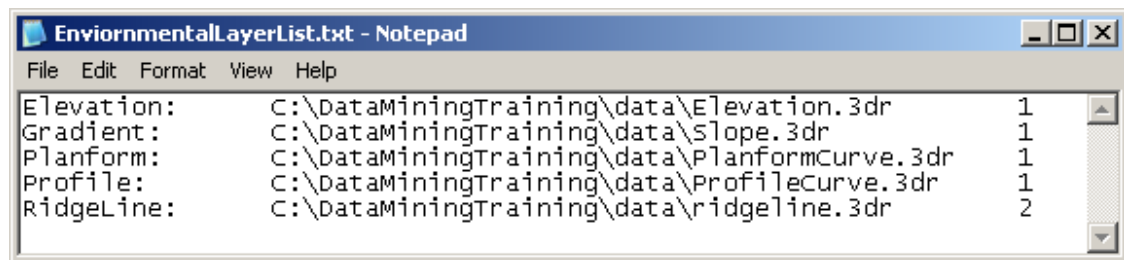


Figure 5 An example of environmental data layer list file

3.4 Map unit list file

This file lists the map units to be used to perform the analysis. It's a plain text file and can be created from the “File --> New Map Unit List ...” menu item. The file format is:

MUKey	SoilLabel
MUKey	SoilLabel
... ..	

“MUKey” is the key for a map unit and “SoilLabel” is the corresponding label for the map unit. “MUKey” and “SoilLabel” are delimited by space or “Tab”. Figure 6 illustrates an example of a map unit list file.

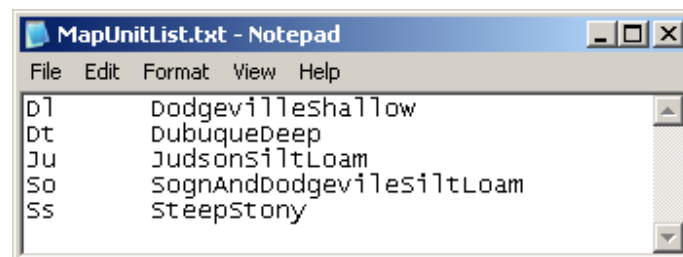



Figure 6 An example of map unit list file

4 Knowledge Extraction

This is a batch procedure used to compute and save all the necessary information into a base data file for later analysis. The extraction process is lengthy, thus the procedure is run off-line rather than in an interactive session.

The knowledge extraction interface (Figure 7) can be accessed from the main menu “Knowledge --> Extract ...” or from the tool bar “”.

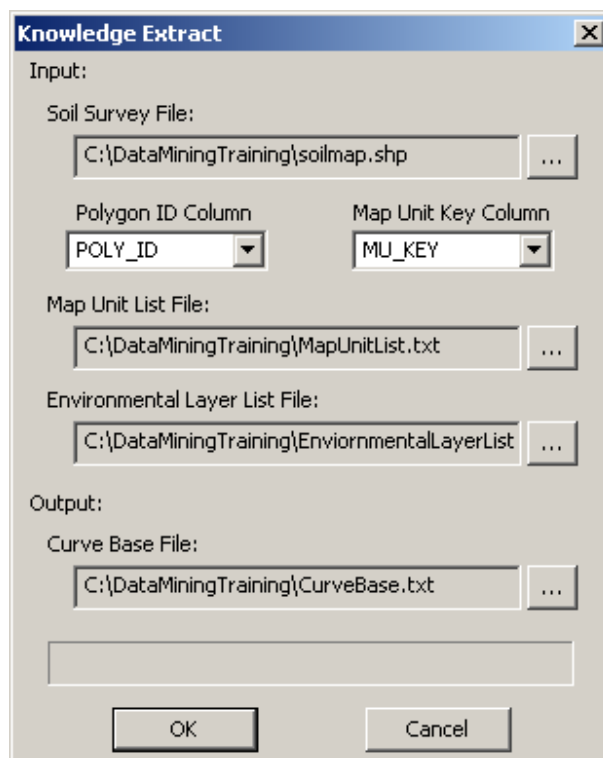



Figure 7 Knowledge extraction interface

“Soil Survey File” is the file of the original soil survey in ArcView shape file format, as explained earlier. “Polygon ID Column” is the field used to store polygon ID and “Map Unit Key Column” is the field used to store map unit key. One needs to specify “Polygon ID Column” and “Map Unit Key Column” from the dropdown control. “Map Unit List File” is the file that lists the map units used to perform analysis. “Environmental Layer List File” is the file that lists the environmental data layers used to perform analysis.

Knowledge Miner 1.0 User Manual

“Curve Base File” is the output file storing the results of the extraction process. In particular, the Curve Base File will have contain pixel frequencies for all soil types measure on all environmental variables indicated. These frequencies are stored on a polygon by polygon basis. Each polygon is tagged with its Map Unit Key.

5 Knowledge Analysis

Knowledge analysis is normally performed for every combination of map unit and environmental data layer. To specify a map unit and an environmental data layer, one can use the main menu “Knowledge --> Analyze ...” or the tool bar “” to call the interface (Figure 8).

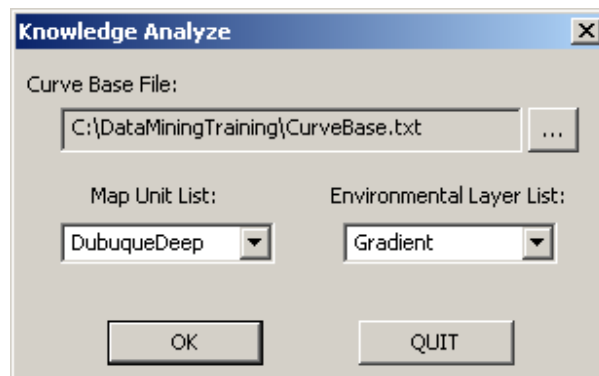


Figure 8 Map unit and environment layer specification

“Curve Base File” is the knowledge curve base file created in “Extracting Knowledge” step and it contains all the necessary information for knowledge analysis. “Map Unit List” lists all the map units that can be used in the analysis and “Environmental Layer List” lists all the environmental data layers can be used in the analysis.

After specifying map unit and environmental data layer, click “OK” button. The four windows in the main interface will be initialized for the specified map unit and environmental data layer (Figure 9).

Knowledge Miner 1.0 User Manual

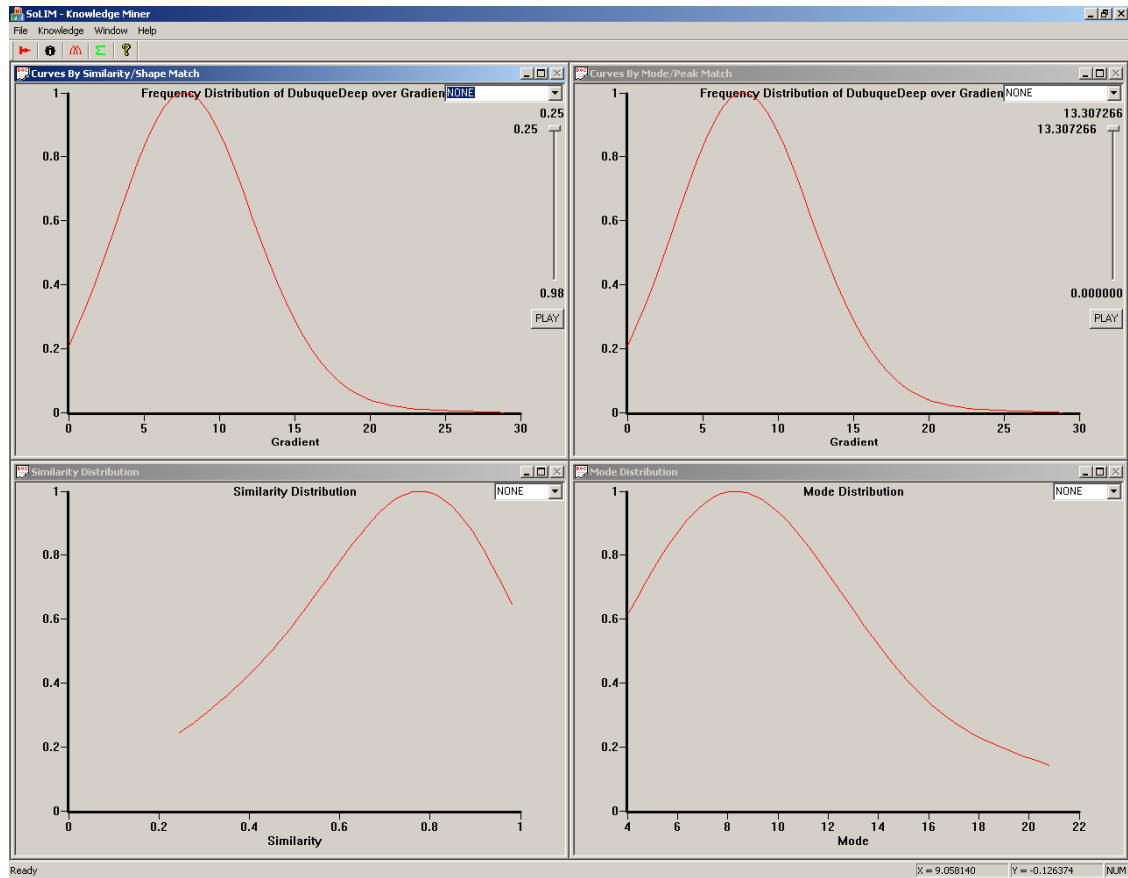


Figure 9 Main interface

These four windows are named “Curves By Similarity/Shape Match”, “Curves By Mode/Peak Match”, “Similarity Distribution”, and “Mode Distribution”.

Whenever you want to know the mouse location within a graph, you can check the status bar. The “**X = 14.069767** **Y = 0.296703**” portion in the status bar tells the X, Y coordinate of the mouse position.

Also, the legend is always available by right clicking in the windows and selecting the “Legend ...” submenu. A dialog will pop up and display the meaning the curves with different colors (Figure 10).

Knowledge Miner 1.0 User Manual

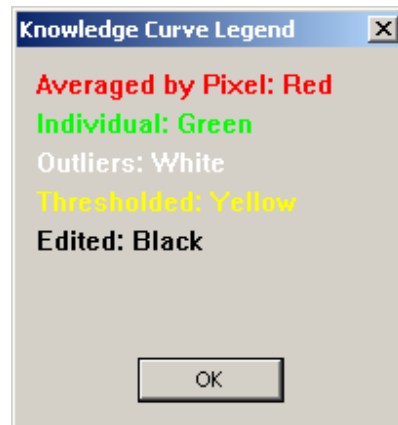


Figure 10 Knowledge curve legend

5.1 Initial knowledge curve

The red curve in the “Curves By Similarity/Shape Match” and “Curves By Mode/Peak Match” windows is the initial knowledge curve extracted from the original soil survey and reflects how the surveyor mapped the map unit on the corresponding environmental factor (Figure 11). It is constructed as the relative frequency of occurrence of various values of the specified environmental variable. It represents a composite of the entire map unit across all polygons. We name it as “Averaged by Pixel”.

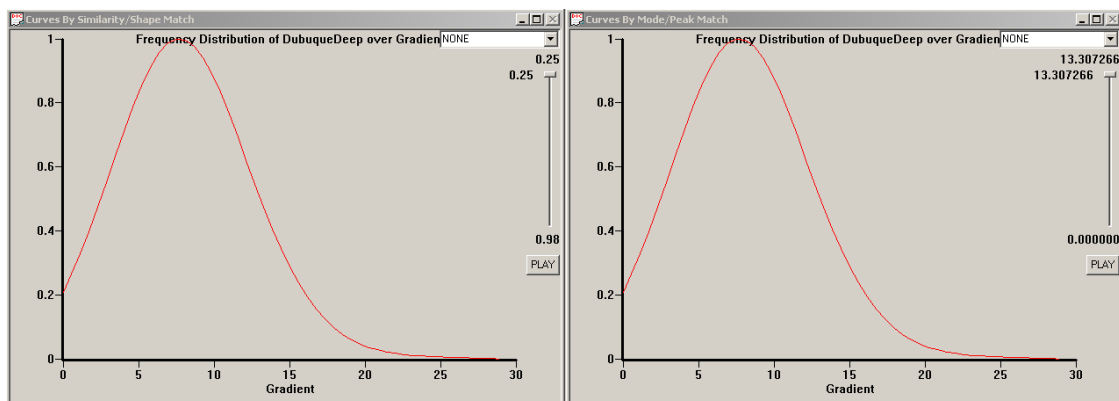


Figure 11 Initial knowledge curve

5.2 Viewing Knowledge: General pattern

We anticipate that the knowledge in terms of the frequency distribution of a specified environmental factor for a specified soil map unit might vary from polygon to polygon. For the same map unit, we construct the frequency distributions on a polygon-by-polygon basis. Polygons occupying anomalous environmental settings can be identified. This can help in both revising soil concepts and removing inconsistency in mapping. Knowledge consistency can be examined from two different perspectives: a general pattern view and an outliers view.

The general pattern view gives the overall idea how consistent the existing soil survey was mapped. We provide two ways to check the overall consistency.

5.2.1 General view similarity distribution

As indicated above, each polygon potentially has a unique frequency distribution. We can measure the similarity between each polygon's frequency distribution and the overall (average) frequency distribution. If the similarities are high, it implies the polygons are not very different from the average or from each other. In that case the soil was consistently mapped. As a help in assessing consistency, we compare each polygon's frequency curve with the average curve. Each comparison yields a single similarity value, thus there will be a range of similarities, with potentially as many values as there are polygons. For easier interpretation, we assemble the values into a frequency distribution of similarity (Figure 12). The horizontal axis is the similarity (range from 0 to 1) and the vertical axis is the relative frequency. In the example, we see that for most polygons, the similarity to the average curve is near 0.8. A few polygons have very low similarity (below 0.4), and a few are virtually identical to the composite curve.

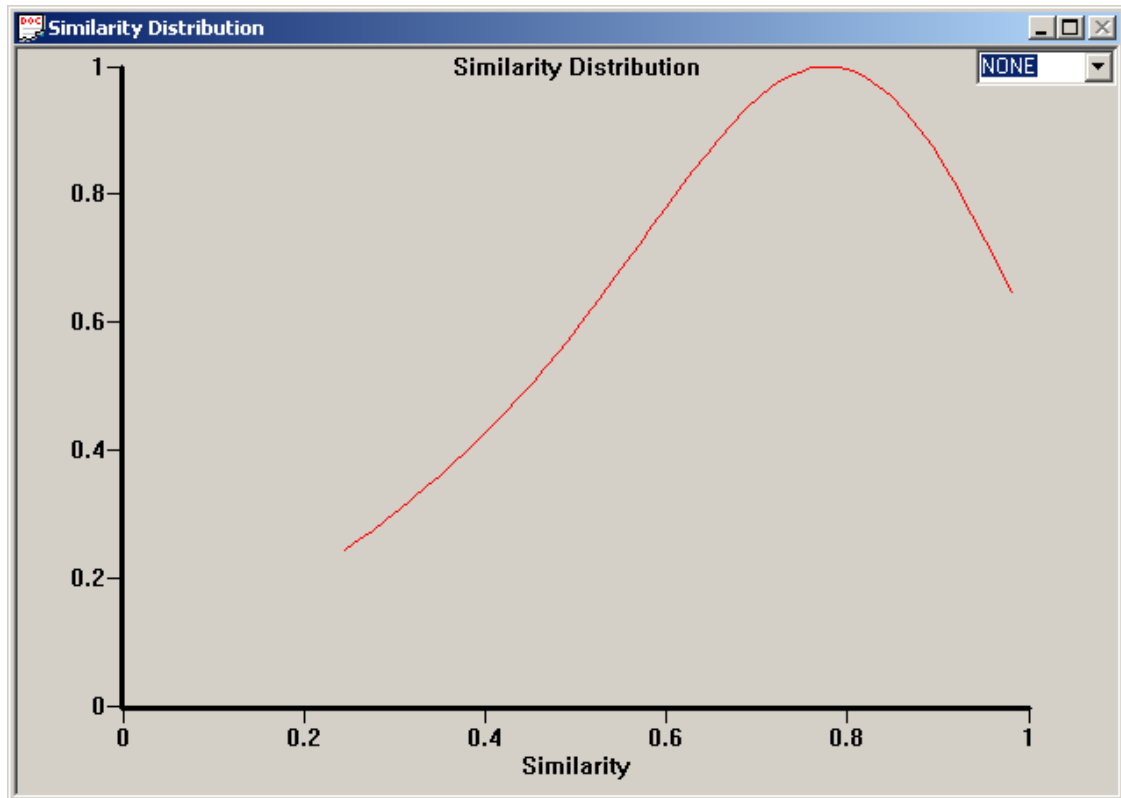


Figure 12 An example of similarity distribution

5.2.2 General view mode distribution

The mode of an individual polygon indicates the most frequently occurring value of an environmental variable. The frequency distribution of these modes provides another way to examine how consistently the soil type is mapped in the existing survey. In particular, it reveals the dispersion in the central value of polygons. For easier interpretation, we assemble the values into a frequency distribution modal values (Figure 13). The horizontal axis is the mode and the vertical axis is the relative frequency of the modal values. In this example we see that the modal slope for most polygons is around 8%, but there are at least a few polygons whose most common slope is above 20%.

Knowledge Miner 1.0 User Manual

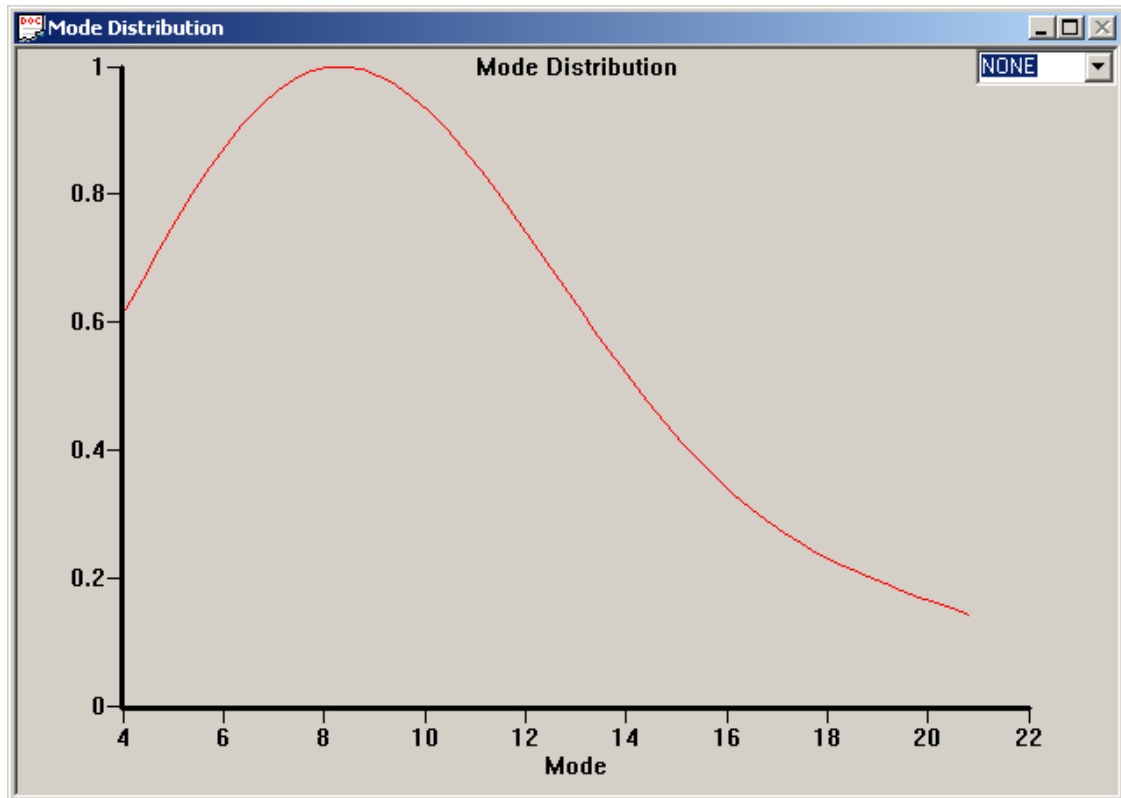



Figure 13 An example of mode distribution

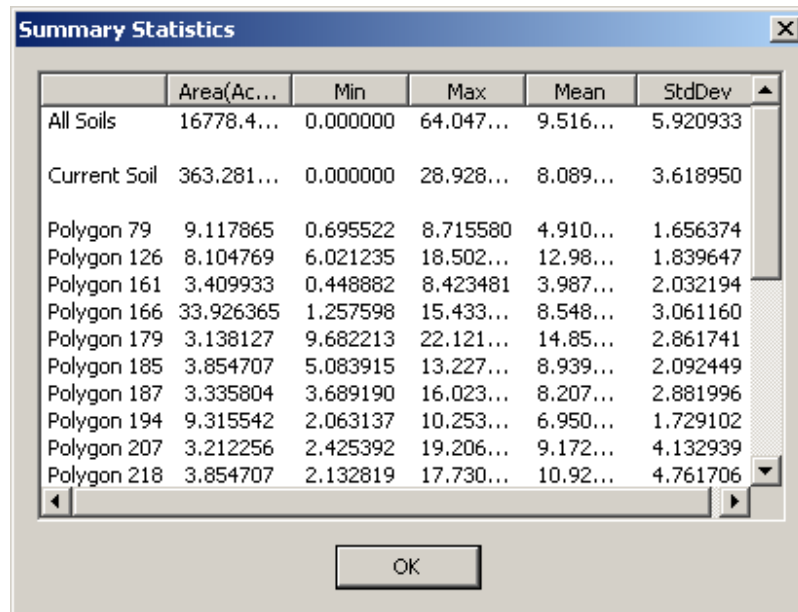
Ideally, all the similarities will be 1 and all the modes will be the same. These two diagrams tell us the overall degree of consistency of the original soil map.

5.3 Viewing knowledge: outliers

Individual polygons can be indexed either by the similarity measure (indicating shape match) or by the distance between its mode and the mode of the overall frequency distribution (indicating peak match). Polygons with low similarities and/or with large mode distances occupy anomalous environmental settings and are considered to be outlier polygons. These polygons can be flagged for further investigation in the field. Basic summary statistics information of those polygons is provided (Figure 14).

The summary statistics information box can be called from main menu “Tool --> Summary Statistics ...”, or from tool bar “”, or by right clicking in the “Curves By

Similarity/Shape Match” or “Curves By Mode/Peak Match” window and selecting “Summary Statistics ...” submenu.



	Area(Ac...	Min	Max	Mean	StdDev
All Soils	16778.4...	0.000000	64.047...	9.516...	5.920933
Current Soil	363.281...	0.000000	28.928...	8.089...	3.618950
Polygon 79	9.117865	0.695522	8.715580	4.910...	1.656374
Polygon 126	8.104769	6.021235	18.502...	12.98...	1.839647
Polygon 161	3.409933	0.448882	8.423481	3.987...	2.032194
Polygon 166	33.926365	1.257598	15.433...	8.548...	3.061160
Polygon 179	3.138127	9.682213	22.121...	14.85...	2.861741
Polygon 185	3.854707	5.083915	13.227...	8.939...	2.092449
Polygon 187	3.335804	3.689190	16.023...	8.207...	2.881996
Polygon 194	9.315542	2.063137	10.253...	6.950...	1.729102
Polygon 207	3.212256	2.425392	19.206...	9.172...	4.132939
Polygon 218	3.854707	2.132819	17.730...	10.92...	4.761706

Figure 14 Summary statistics

5.3.2 Outliers in terms of low similarities

5.3.2.1 Potential outliers

For the specified map unit, all the soil polygons are sorted based on their similarity measures (shape match). They are displayed in the “Curves By Similarity/Shape Match” window. Dragging the slider causes individual frequency distribution curves for soil polygons (white curve in Figure 15) to appear in the order of shape match, i.e., those appearing first are less similar to the overall distribution (Figure 15). Soil polygons with low similarity measures can be thought as potential outliers.

One can also click the “PLAY” button  below the slider to display these curves automatically at constant speed.

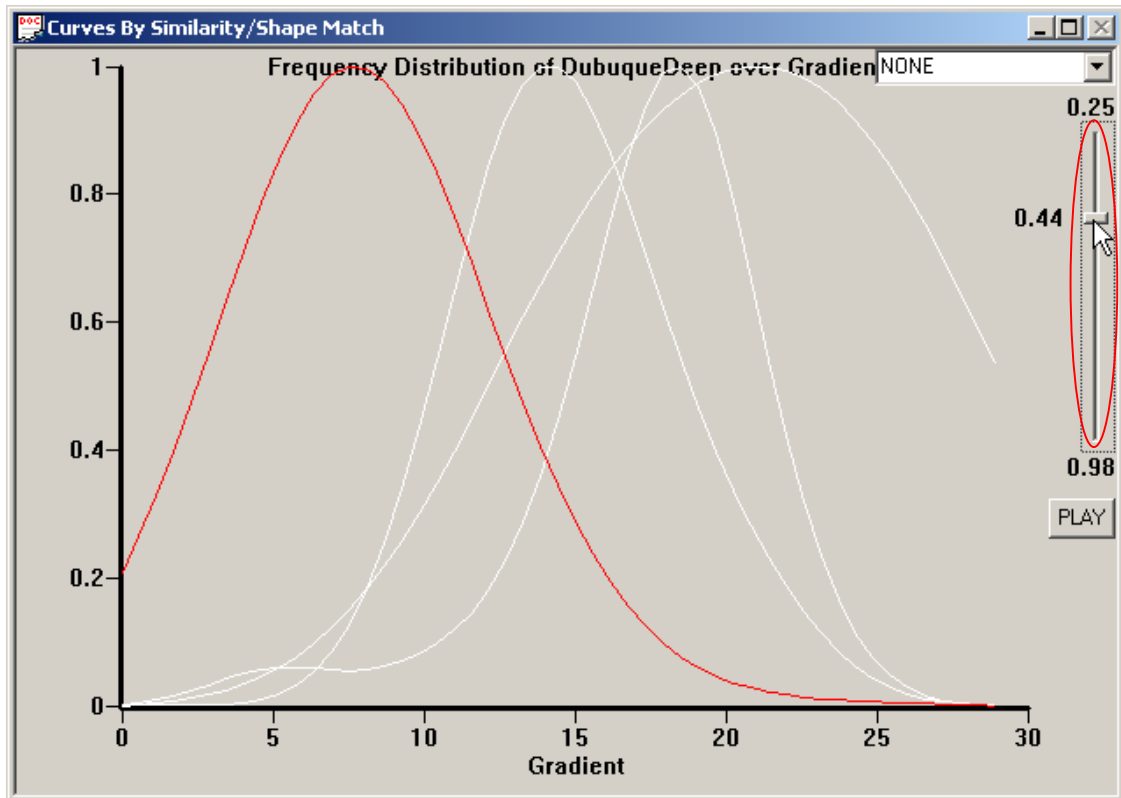
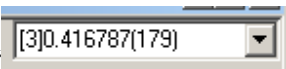


Figure 15 Potential outliers with low similarity measures

5.3.2.2 Looking into an individual curve

If a particular outlier knowledge curve (in white) is of interest and one wants to examine the corresponding soil polygon, one can use the combo list tool right above the sliding

bar “”.

There are three fields in the combo list control, with interpretation as follows:

- [3] – this is the order of the selected polygon (3rd least similar in this example)
- 0.416787 – this is the similarity of the selected polygon to overall distribution
- (179) – this is the polygon ID

The knowledge curve for the polygon selected in the combo list control will display in green (Figure 16). This provides an easy way to cycle through the individual polygons while viewing polygon distributions. When one finds that the green curve matches the

outlier curve of interest (in white), the corresponding polygon ID will be seen. Using this ID, one can find the corresponding soil polygon using ArcView or 3dMapper. One can also find the summary statistics information for the selected (green) polygon.

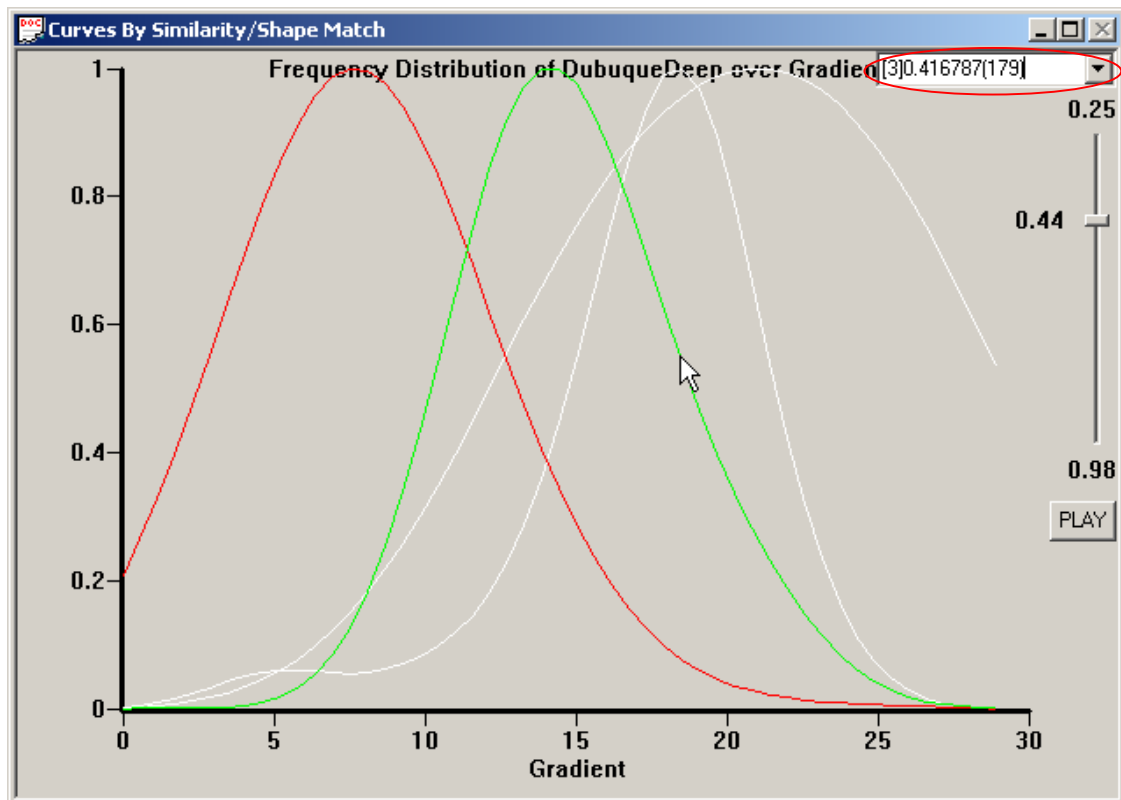


Figure 16 Looking into an individual curve

5.3.2.3 Saving potential outlier polygons

Soil polygons with low similarity measures can be saved into a dbf file. Right clicking in the “Curves By Similarity/Shape Match” window and clicking “Save Outlier Polygon List ...” submenu, the “Save Outlier Polygon List” dialog will pop up (Figure 17). The default threshold is the current value of the slider bar. All soil polygons with similarity measure less than the threshold will be saved into a dbf file.

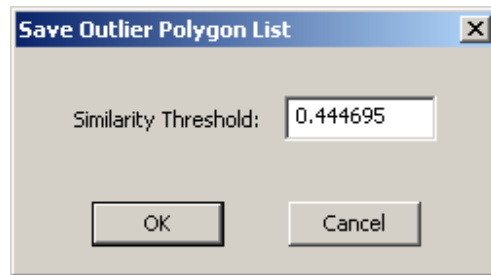



Figure 17 Similarity threshold for saving outlier polygon list

5.3.2 Outliers in terms of large mode distances

5.3.2.1 Potential outliers

Outlier polygons can also be examined based on their mode distance values. The controls for this work identically to the similarity controls, except that polygons are sorted according to increasing modal distance, and the values displayed are mode distance rather than similarity. Thus as seen in Figure 18, frequency distribution curves for soil polygons (white curve in Figure 18) will appear in the order of peak match. That is, polygons at the top of the list are more distant from the overall distribution (Figure 18). Soil polygons with large modal distance measures can be thought as are potential outliers.

Again, one can also click the “PLAY” button “” below the slider bar to display these curves automatically at constant speed.

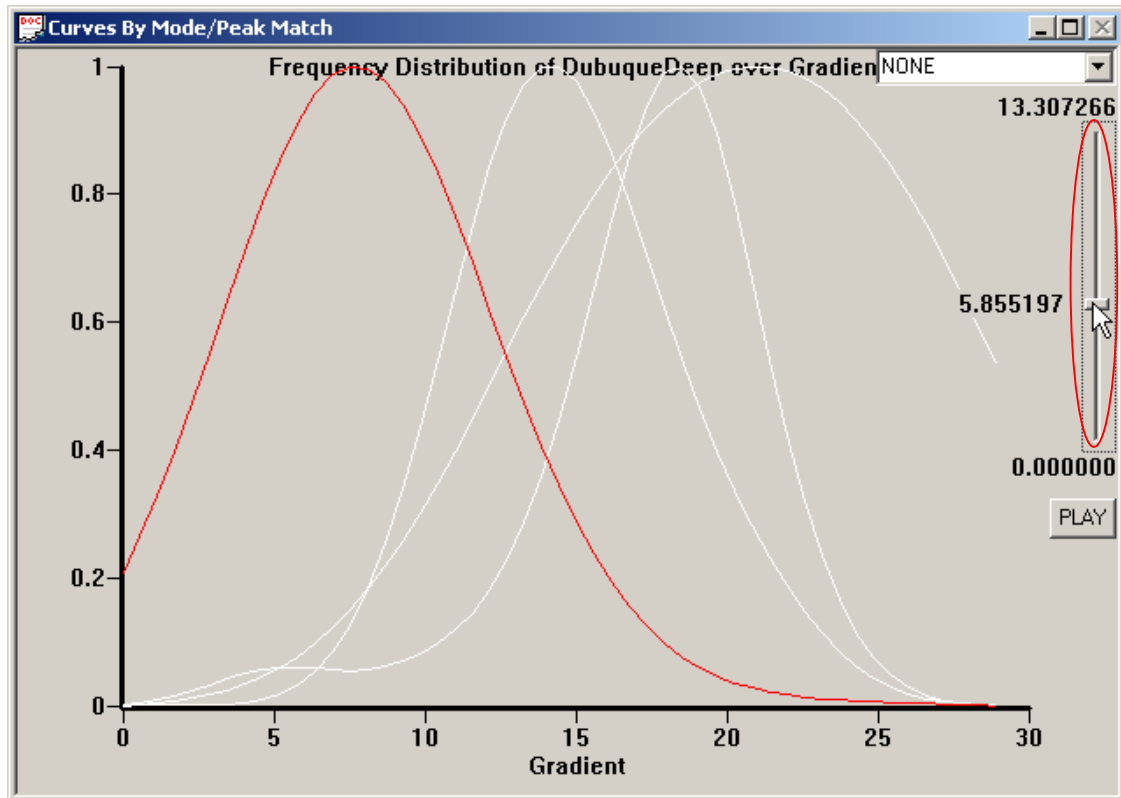
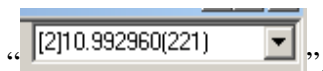


Figure 18 Potential outliers with large mode distance measures

5.3.2.2 Looking into an individual curve

Once again there is a combo list control to use in examining individual polygons:



The information in the control is interpreted much like before:

- [2] – this is the order of the selected polygon
- 10.992960 – this is the distance between the mode of the selected polygon and the mode of the overall distribution
- (221) – this is the polygon ID

The knowledge curve for the polygon selected in the combo list control will display in green (Figure 19). When the selected curve matches the outlier curve of interest, one can get the corresponding polygon ID from the combo control. Using this ID, one can find

the corresponding soil polygon using ArcView or 3dMapper. One can also find the summary statistics information for this polygon.

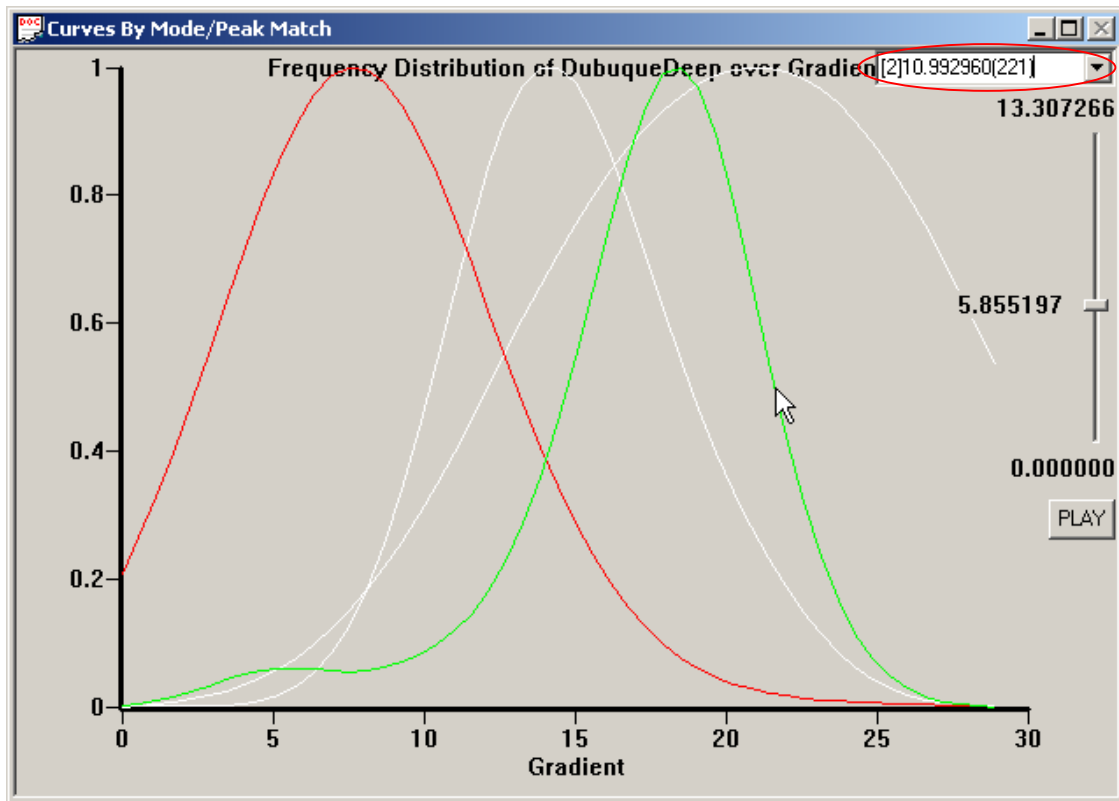


Figure 19 Looking into an individual curve

5.3.2.3 Saving potential outlier polygon

Soil polygons with large modal distances can be saved into a dbf file. Right clicking in the “Curves By Mode/Peak Match” window and clicking “Save Outlier Polygon List ...” submenu, the “Save Outlier Polygon List” dialog will pop up (Figure 20). The default threshold is the current value of the bar. Soil polygons with mode distance measure greater than the threshold will be saved into a dbf file.

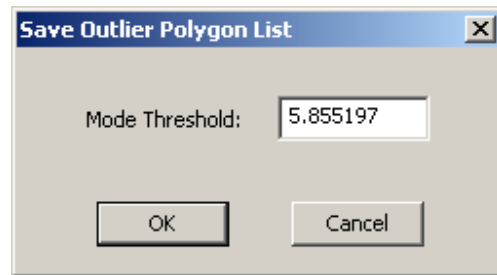


Figure 20 Mode distance threshold for saving outlier polygon list

6 Knowledge Refinement

6.1 Automated refinement

Polygons with low similarities (to the overall curve) and/or with large mode distances (from the overall curve) occupy anomalous environmental settings and are considered to be outlier polygons. A user can remove these soil polygons and reconstruct the knowledge curve (frequency distribution).

6.1.1 Removing soil polygons with low similarity measures

Right click in the “Curves By Similarity/Shape Match” window and click “Recompute Knowledge Curve ...” submenu. “Recompute ...” dialog will pop up (Figure 21). The default threshold is the current value of the slider bar. One can change it freely. Soil polygons with similarity measure less than the threshold will be removed and improved knowledge curve will be reconstructed (Figure 22, the yellow curve is the reconstructed knowledge curve). “Similarity Distribution” and “Mode Distribution” will be updated as well (Figure 23, the yellow curves are the updated distributions). “Sample Percentage” indicates the percentage of pixels used in re-computation. The default percentage is 100, i.e., all the pixels will be used in re-computation. Again, one can change it freely. Smaller percentage will save computing time at the expense of conformity to actual knowledge curve.

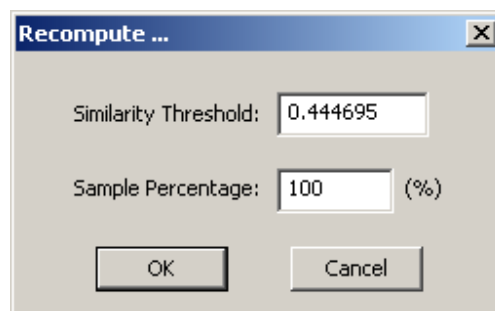


Figure 21 Similarity threshold and sample percentage for recomputing knowledge curve

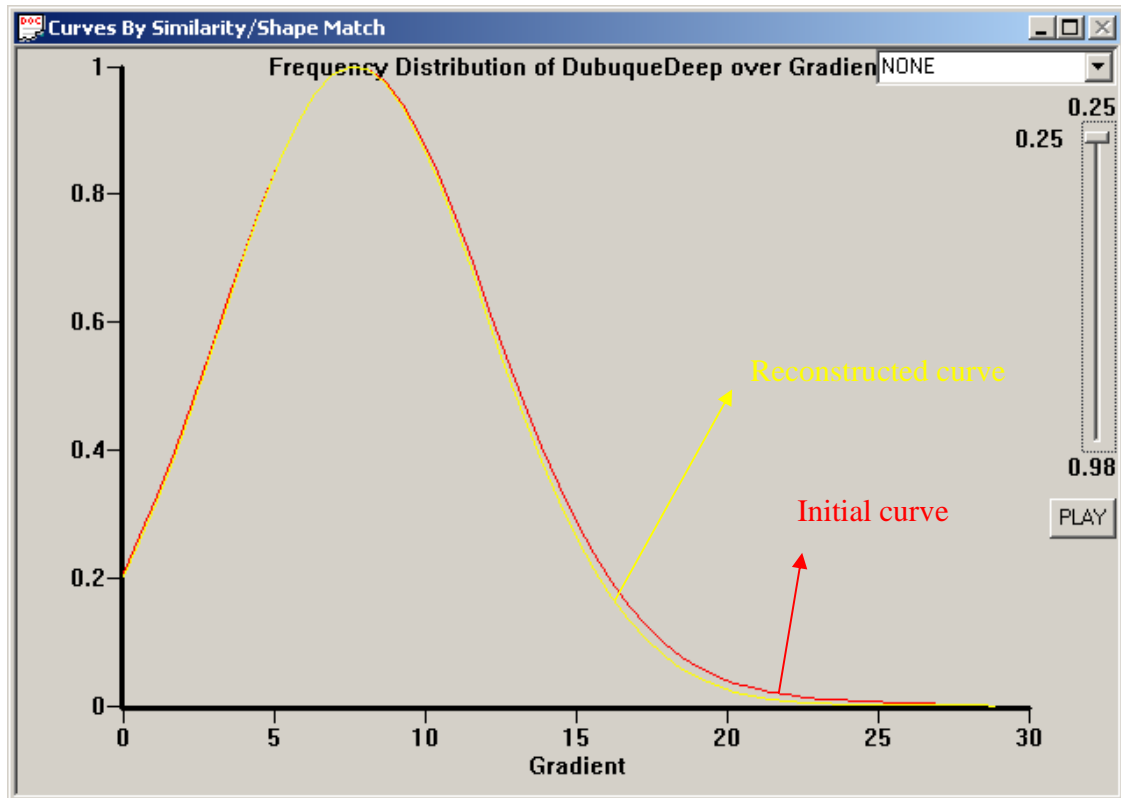


Figure 22 Reconstructed knowledge based on similarity

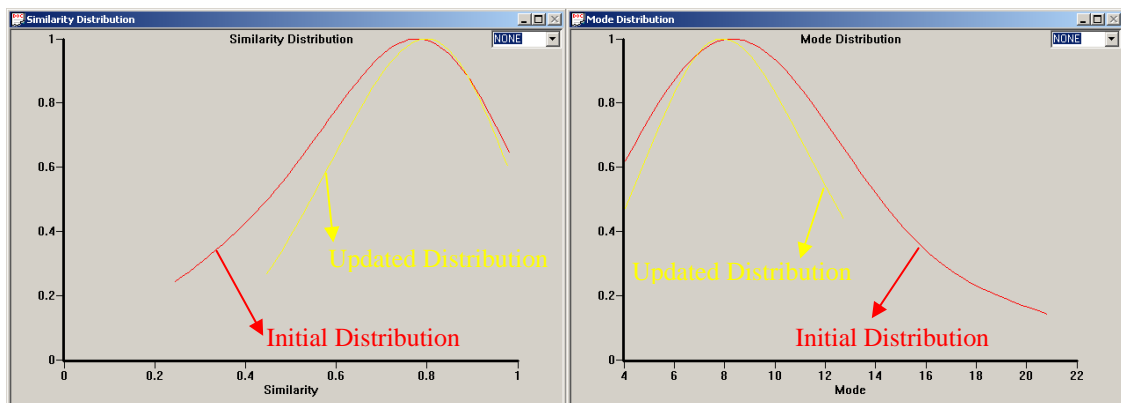


Figure 23 Updated similarity distribution and mode distribution

6.1.2 Removing soil polygons with large mode distance measures

Right click in the “Curves By Mode/Peak Match” window and click “Recompute Knowledge Curve ...” submenu. “Recompute ...” dialog will pop up (Figure 24). The default threshold is the current value of the slider bar Soil polygons with mode distance measure greater than the threshold will be removed and improved knowledge curve will be reconstructed (Figure 25, the yellow one is the reconstructed knowledge curve). “Similarity Distribution” and “Mode Distribution” will be updated as well (Figure 26, the yellow curves are the updated distributions). “Sample Percentage” indicates the percentage of pixels used in re-computation. The default percentage is 100, i.e., all the pixels will be used in re-computation. Again, one can change it freely. Smaller percentages will save computing time at the expense of conformity to actual knowledge curve.

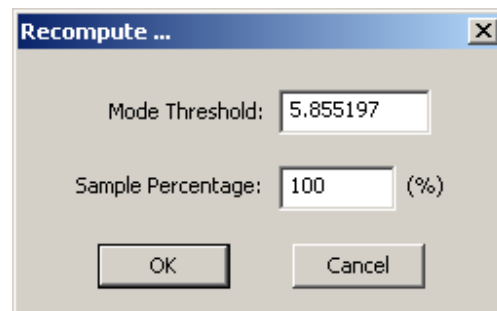


Figure 24 Mode distance threshold and sample percentage for recomputing knowledge curve

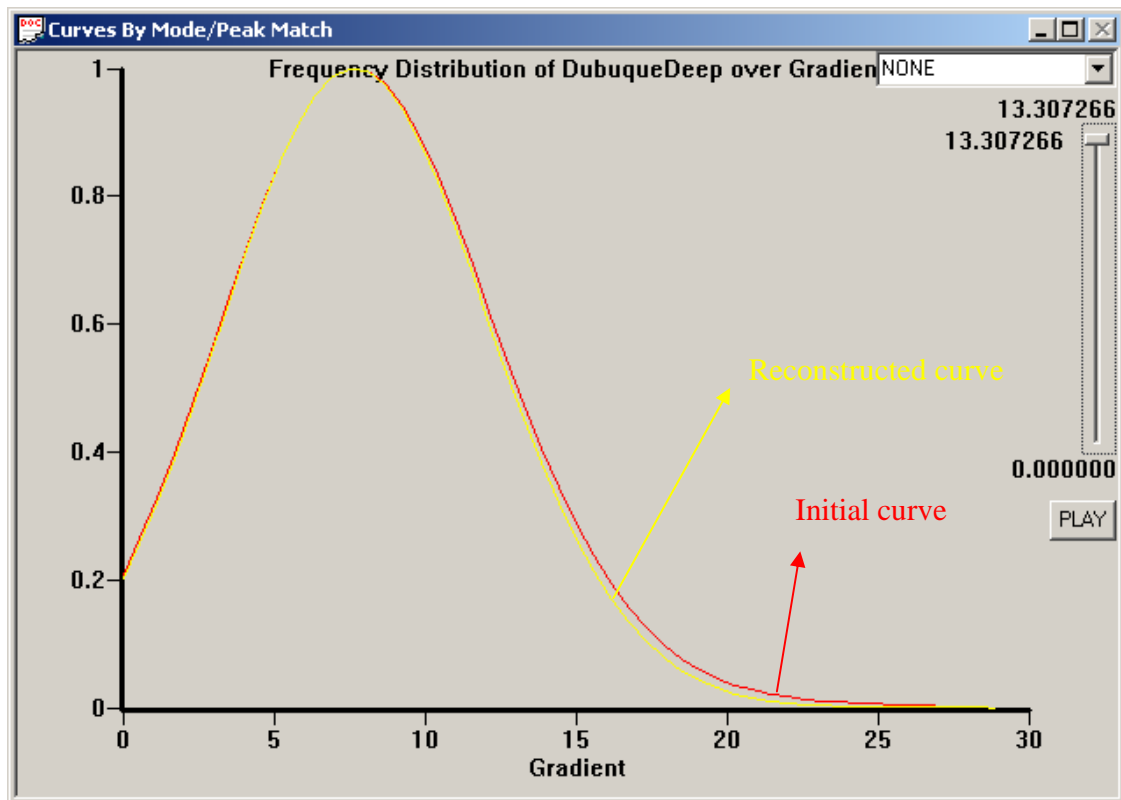


Figure 25 Reconstructed knowledge curve based on mode distance

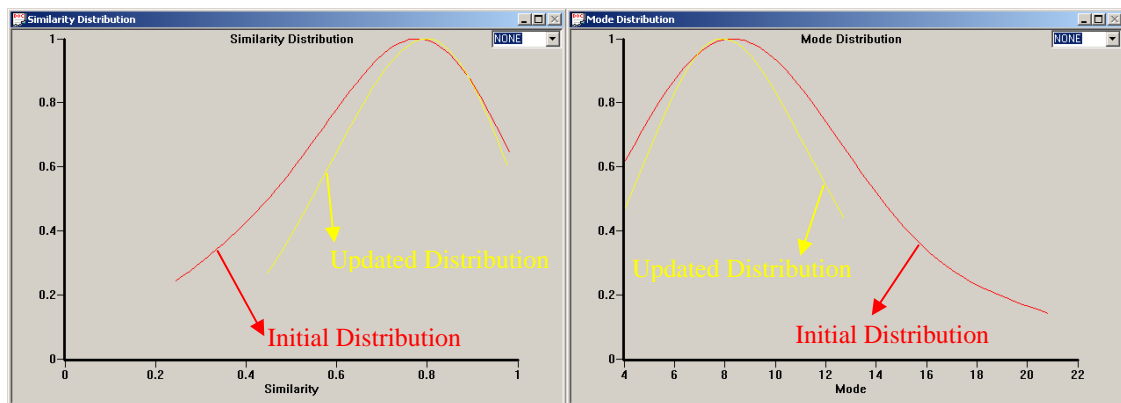


Figure 26 Updated similarity distribution and mode distribution

6.2 Manual refinement

Alternatively, soil scientists may want to revise the knowledge by manually adjusting the curve based on their experience. The shape of the knowledge curve is governed by a

small number of control points. The scientist drags the control points to change the shape of the curve.

6.2.1 Fitting the extracted knowledge curve

Right click in the “Curves By Similarity/Shape Match” or “Curves By Mode/Peak Match” window and click “Edit Knowledge Curve ...” submenu. A dialog will pop up asking the number of control points used to fit the knowledge curve (Figure 27). The default number is “15”. A large number of points provides very detailed control over the curve, but requires more effort when editing than just a few points.



Figure 27 Number of control points

6.2.2 Adjusting the extracted knowledge curve

After fitting, one can adjust the extracted knowledge curve by moving, adding, and removing control points (Figure 28, the black curve is the fitted curve which can be edited via the control points).

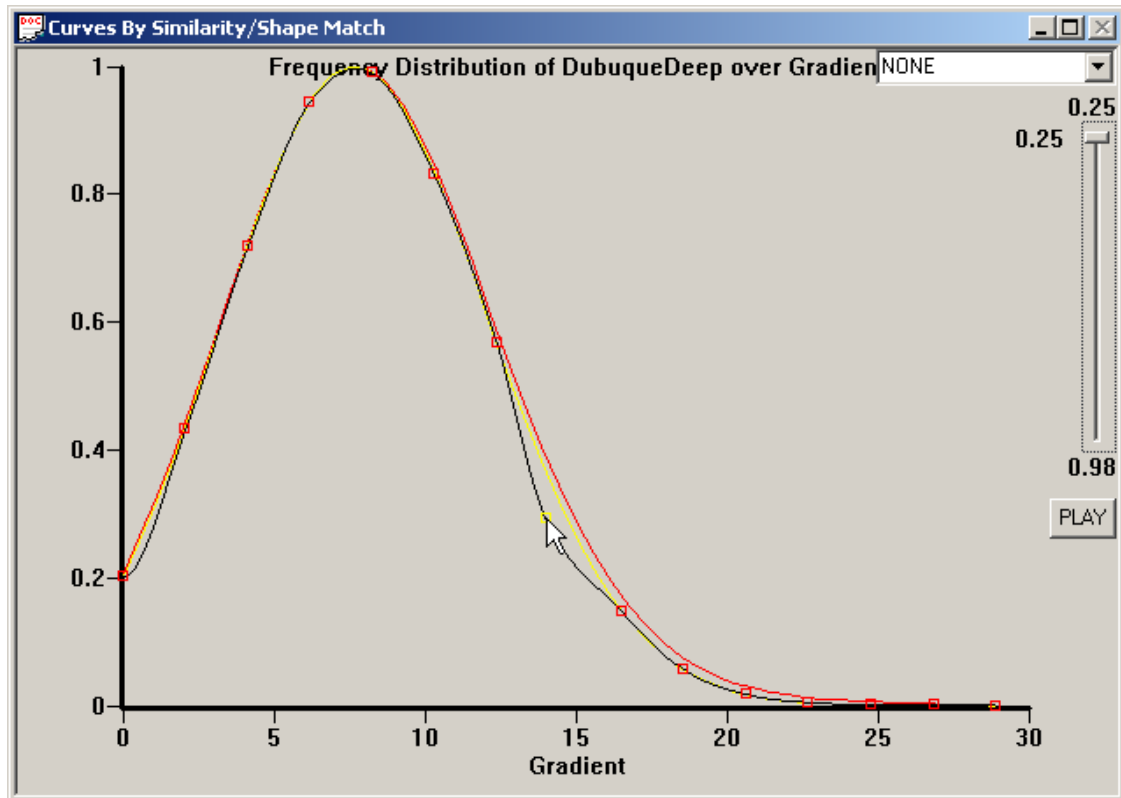


Figure 28 Manual adjustment of the knowledge curve

6.2.2.1 Moving control points

Select a control point and drag it to the desired location.

6.2.2.2 Adding control points

Double click in the desired location and a new control point will be added to the curve.

6.2.2.3 Deleting control points

Hold “shift” key and double click the control point which needs to be removed.

6.2.3 Saving knowledge

One can save the knowledge curves for examination later, or for use as knowledge in SoLIM soil inference.

To save the knowledge curve, right click in the “Curves By Similarity/Shape Match” or “Curves By Mode/Peak Match” window and click “Save Knowledge Curve...” submenu. A popup dialog will guide you to store the knowledge curve into files.

6.3 Knowledge Comparator

There is another tool called knowledge comparator that can be used to reload the saved knowledge curves (Figure 29). This is useful for comparing the knowledge of one map unit to the knowledge of another unit. Curves for different map units can be plotted on the same screen and edited side-by-side.

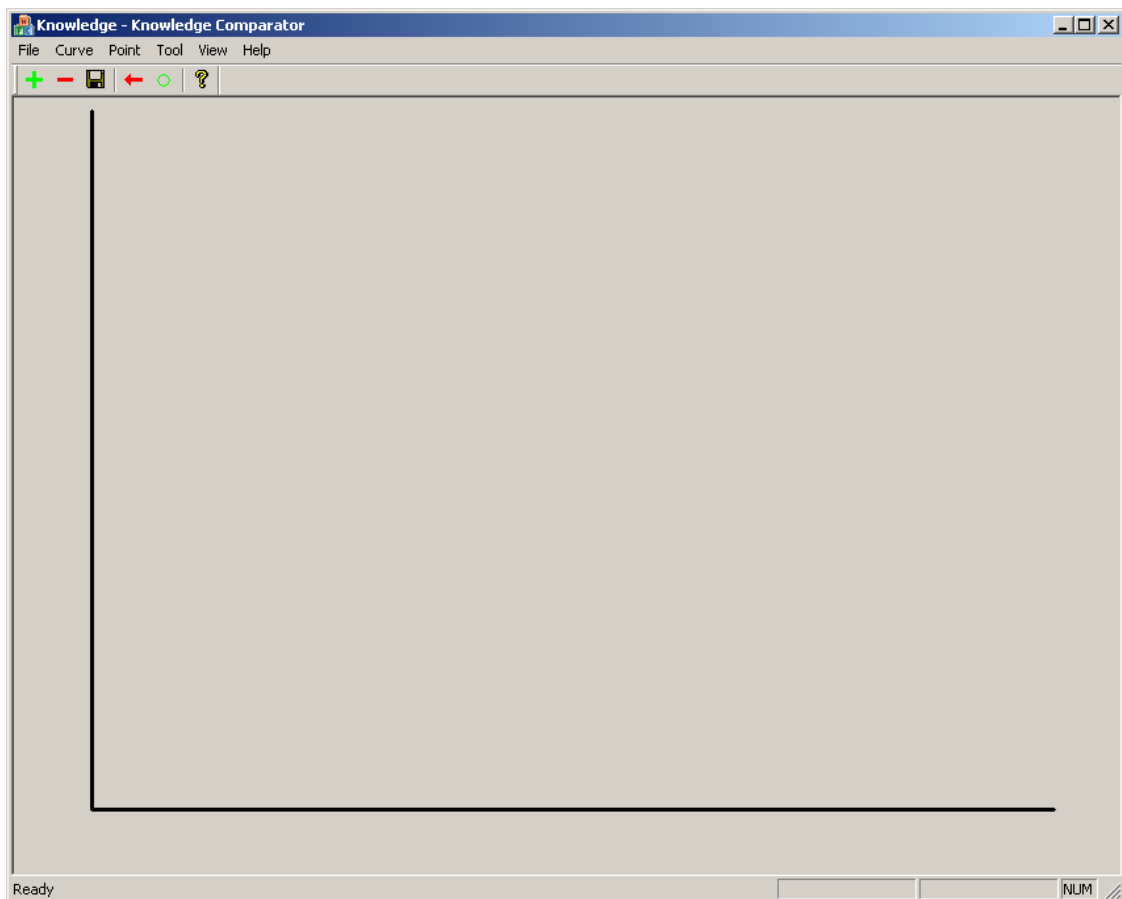


Figure 29 Knowledge comparator

7 Use of the Knowledge

If desired, the refined knowledge curves can be used for documentation for future use and/or can be used in knowledge-based inference system such as SoLIM (<http://solim.geography.wisc.edu>) to produce a revised survey. A special version of SoLIM has been coded to read the output knowledge from the data mining tools. The knowledge curves become rules that are used to compute membership values. Of course, if the soil scientist has created new soil concepts in the course of data mining, the new soils will appear in the revised survey. Both pre-existing and new concepts will be mapped by SoLIM consistently across the landscape based on the formative environmental data.

Appendix A: Tutorial

This tutorial session will walk you through the procedures of data preparation, knowledge extraction, knowledge analysis, and knowledge refinement.

Step 1: Data preparation

1. Copy tutorial data:

- 1) Create a “DataMiningTraining” folder at C drive root directory, i.e., “C:\DataMiningTraing”;
- 2) Copy all the contents in the “TutorialData” folder to “C:\DataMingTraining”.

2. Create an environmental data layer list file:

- 1) Create an “EnvironmentalLayerList.txt” plain text file in “C:\DataMingTraining” using notepad or using “File --> New Environmental Layer List File ...” menu in “Knowledge Miner”. The file format is explained in section 3.3 and the file content is shown below. In this example all variables except for RidgeLine are continuous.






Elevation:	C:\DataMiningTraining\data\Elevation.3dr	1
Gradient:	C:\DataMiningTraining\data\Slope.3dr	1
Planform:	C:\DataMiningTraining\data\PlanformCurve.3dr	1
Profile:	C:\DataMiningTraining\data\ProfileCurve.3dr	1
RidgeLine:	C:\DataMiningTraining\data\ridgeline.3dr	2

3. Create a map unit list file:

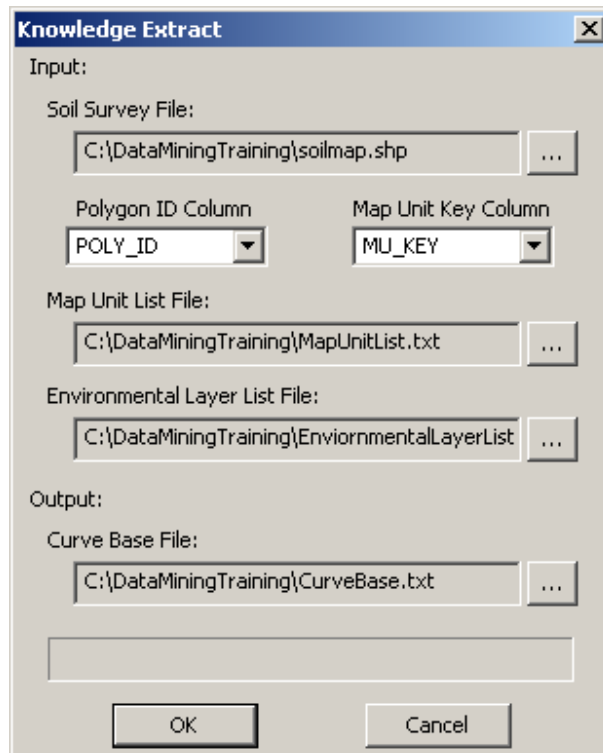
- 1) Create a “MapUnitList.txt” plain text file in “C:\DataMingTraining” using notepad or using “File --> New Map Unit List File ...” menu in “Knowledge Miner”. The file format is explained in section 3.4 and the file content is:

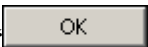
Dl	DodgevilleShallow
Dt	DubuqueDeep
Ju	JudsonSiltLoam
So	SognAndDodgevileSiltLoam
Ss	SteepStony

Step 2: Knowledge extraction

- 1) In “Knowledge Miner”, click “” toolbar or click “Knowledge --> Extract ...” menu to launch the “Knowledge Extract” dialog;
- 2) Click “” button for “Soil Survey File:” and specify “soilmap.shp” as the input soil survey file;
- 3) In the dropdown list, specify “POLY_ID” for “Polygon ID Column” and “MU_KEY” for “Map Unit Key Column”;
- 4) Click “” button for “Map Unit List File:” and specify “MapUnitList.txt” as the input map unit list file;
- 5) Click “” button for “Environmental Layer List File:” and specify “EnvironmentalLayerList.txt” as the input environmental layer list file;
- 6) Click “” button for “Curve Base File:” and name “CurveBase.txt” as the output curve base file.

The “Knowledge Extract” dialog should now look as below:





- 7) Click “” to perform the knowledge extraction. This can take hours to complete, depending on the dataset size. The result will be saved in “C:\DataMiningTraining\CurveBase.txt”.

Step 3: Knowledge analysis

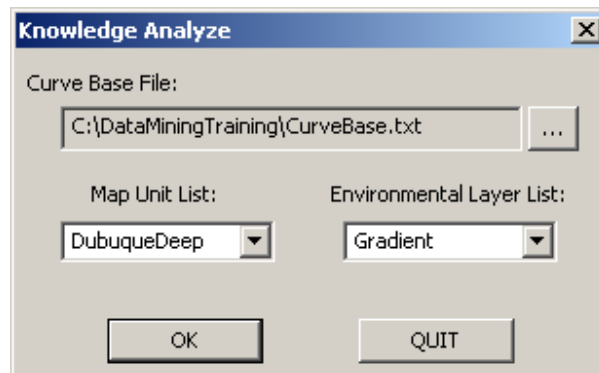
Knowledge analysis is performed for every map unit on every environmental data layer. We are using map unit “DubuqueDeep” and environmental data layer “Gradient” as an example.

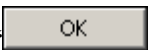
1. Load in knowledge base:

- 1) In “Knowledge Miner”, click “” toolbar or click “Knowledge --> Analyze ...” menu to launch the “Knowledge Analyze” dialog;
- 2) Click “” button for “Curve Base File:” and specify “CurveBase.txt” as the input curve base file, if it is not specified yet;

- 3) In the dropdown list, specify “DubuqueDeep” for “Map Unit List:” and “Gradient” for “Environmental Layer List:”;

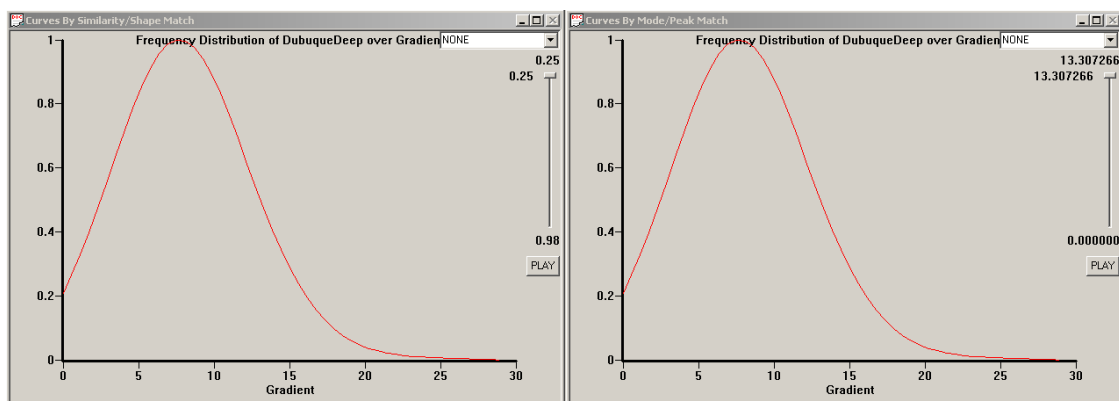
The “Knowledge Analyze” dialog should now look as below:



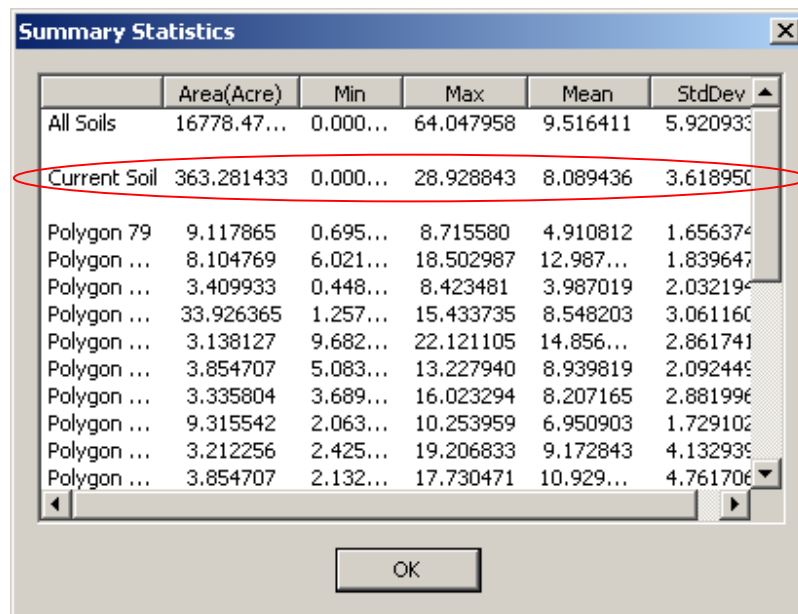
- 4) Click “” to load in the knowledge for “DubuqueDeep” over “Gradient”.

2. Investigate the initial knowledge curve

- 1) “Curves By Similarity/Shape Match” window and “Curves By Mode/Peak Match” window show the initial knowledge curve extracted from the old soil survey and reflects how the original surveyor mapped DubuqueDeep over Gradient.



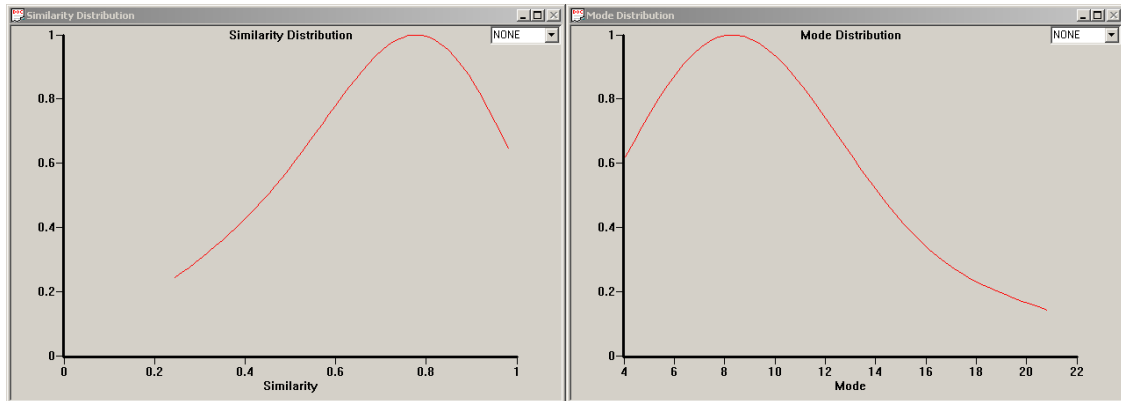
- 2) Move the mouse around in the window and the coordinate information will be displayed in the status bar area “ $X = 14.069767$ $Y = 0.296703$ ”.
- 3) Right click in “Curves By Similarity/Shape Match” window or “Curves By Mode/Peak Match” window and select “Summary Statistics ...” menu or click “ Σ ” toolbar, a summary statistics dialog will popup. The “Current Soil” row shows the summary statistics information for Gradient of DubuqueDeep.




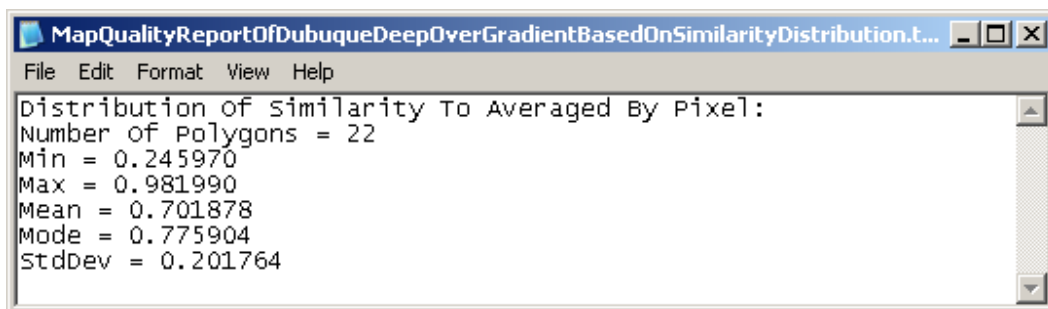
	Area(Acre)	Min	Max	Mean	StdDev
All Soils	16778.47...	0.000...	64.047958	9.516411	5.920933
Current Soil	363.281433	0.000...	28.928843	8.089436	3.618950
Polygon 79	9.117865	0.695...	8.715580	4.910812	1.656374
Polygon ...	8.104769	6.021...	18.502987	12.987...	1.839647
Polygon ...	3.409933	0.448...	8.423481	3.987019	2.032194
Polygon ...	33.926365	1.257...	15.433735	8.548203	3.061160
Polygon ...	3.138127	9.682...	22.121105	14.856...	2.861741
Polygon ...	3.854707	5.083...	13.227940	8.939819	2.092449
Polygon ...	3.335804	3.689...	16.023294	8.207165	2.881996
Polygon ...	9.315542	2.063...	10.253959	6.950903	1.729102
Polygon ...	3.212256	2.425...	19.206833	9.172843	4.132939
Polygon ...	3.854707	2.132...	17.730471	10.929...	4.761706


3. Investigate the general pattern

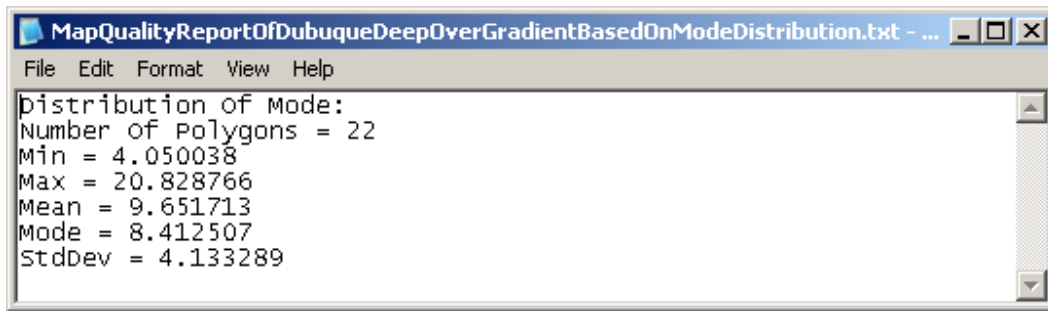
- 1) “Similarity Distribution” window and “Mode Distribution” window show the overall mapping consistency of DubuqueDeep over Gradient. For the similarity distribution curve, the horizontal axis is the similarity (range from 0 to 1) and the vertical axis is the relative frequency. We see that for most polygons, the similarity to the average curve is near 0.8. A few polygons have very low similarity (below 0.4), and a few are virtually identical to the composite curve (close to 1). For the mode distribution curve, the horizontal axis is the mode and the vertical axis is the relative frequency of the modal values. We see that most of the polygon modes are around 8.2 while some polygons have very different modes (greater than 21).





- 2) Right click in “Similarity Distribution” window and select “Map Quality Report ...” menu, accept the default file name and click “” in the “Save As” dialog. Some summary statistics for the similarity distribution curve will be saved and this information can be used as map quality report. Take a look at the saved file and it should look like:

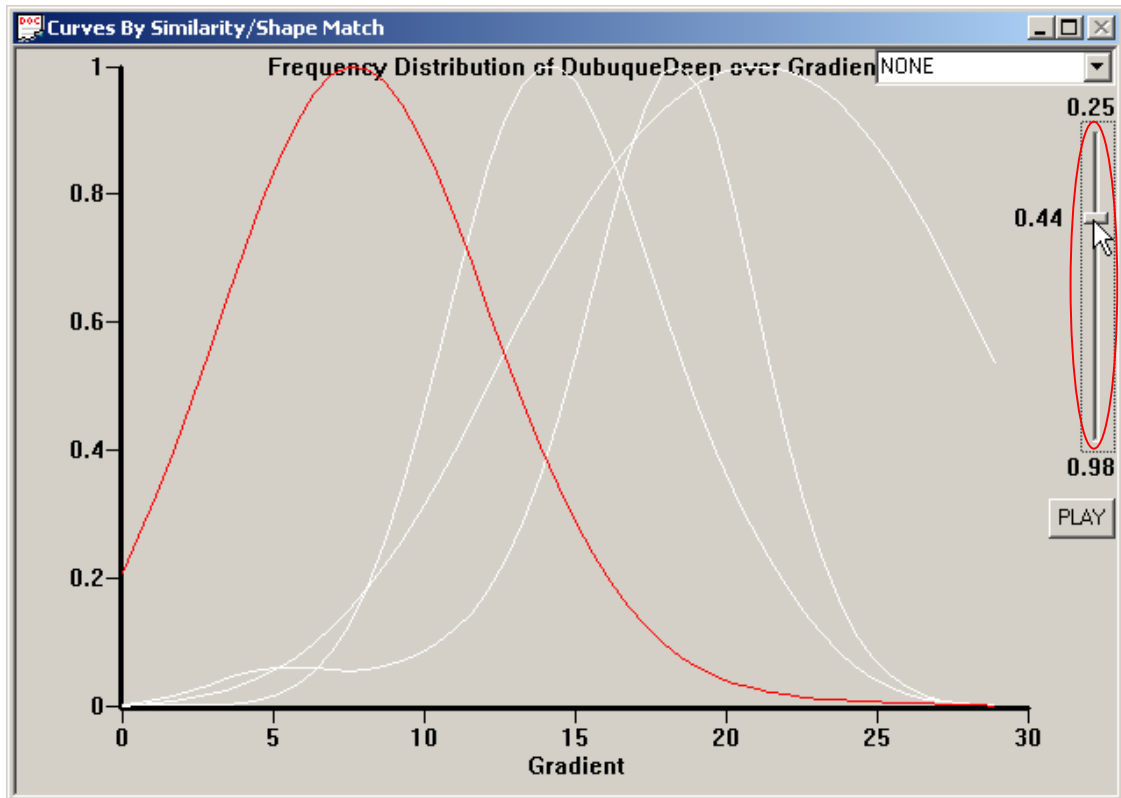


- 3) Right click in “Mode Distribution” window and select “Map Quality Report ...” menu, accept the default file name and click “” in the “Save As” dialog. Some summary statistics for the mode distribution curve will be saved and this information can be used as map quality report. Take a look at the saved file and it should look like:

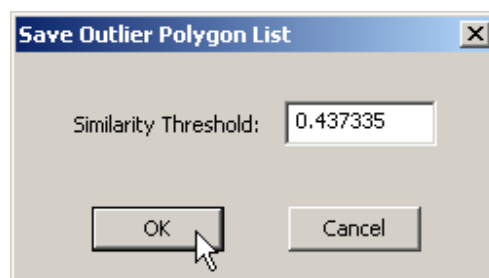


4. Investigate the outliers

- 1) In the “Curves By Similarity/Shape Match” window, drag down the slider slowly. Frequency distribution curves for DubuqueDeep soil polygons will show up one by one in the order of shape match, i.e., those appearing first are less similar to the overall distribution. Soil polygons with low similarity measures can be thought as potential outliers. Experiment with this tool and compare the knowledge curves for each DubuqueDeep soil polygon with the knowledge curve for the DubuqueDeep map unit. This process can also be automated by clicking the   button below the slider. Drag down the slider to 0.44.



- 2) Right click in the “Curves By Similarity/Shape Match” window and select “Save Outlier Polygon List ...” submenu. The “Save Outlier Polygon List” dialog will pop up. Leave the default threshold as is and click “OK”. Leave the default file name as is in the popup “Save As ...” dialog.



- 3) Open the saved outlier polygon list file
 “OutlierListOfDubuqueDeepOverGradientBasedOnSimilarity.dbf”. Polygon 221, 273, and 179 are identified as outlier polygons. These polygons can be located based on their IDs using GIS software like ArcGIS, ArcView, or 3dMapper.

POLY_ID	Grad_Sim	Grad_Min	Grad_Max	Grad_Mean	Grad_SD
221	0.246	4.439	23.192	17.357	3.548
273	0.283	10.558	28.929	20.639	4.861
179	0.417	9.682	22.121	14.856	2.862

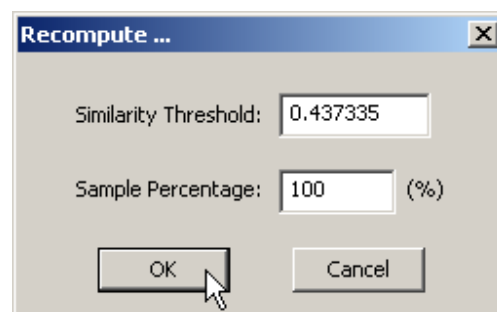
Exercise:

Identify some outlier polygons in terms of large mode distances (hint: operate in the “Curves By Mode/Peak Match” window).

Step 4: Knowledge refinement

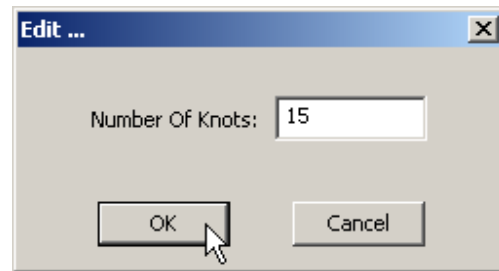
1. Remove soil polygons with low similarity measures

- 1) Right click in the “Curves By Similarity/Shape Match” window and select “Recompute Knowledge Curve ...” submenu. “Recompute ...” dialog will pop up. Leave the default threshold as is and click “OK”. Soil polygons with similarity measure less than the threshold will be removed and improved knowledge curve will be reconstructed (displayed in yellow color).



2. Refine the improved knowledge curve manually

- 1) Right click in the “Curves By Similarity/Shape Match” window and select “Edit Knowledge Curve ...” submenu. A dialog will pop up to ask the number of control points used to fit the knowledge curve. Leave the default number “15” and click “OK”.





- 2) The black curve is a fitted curve which approximates the yellow curve. It serves as the starting point of manual adjustment. Based on your knowledge, you can edit the black knowledge curve by moving/adding/deleting the control points. In this example here, just leave it as is.
- 3) Right click in the “Curves By Similarity/Shape Match” and select “Save Knowledge Curve...” submenu. In the popup “Save As ...” dialog, leave the default file name “DubuqueDeepOverGradient.txt” as is.

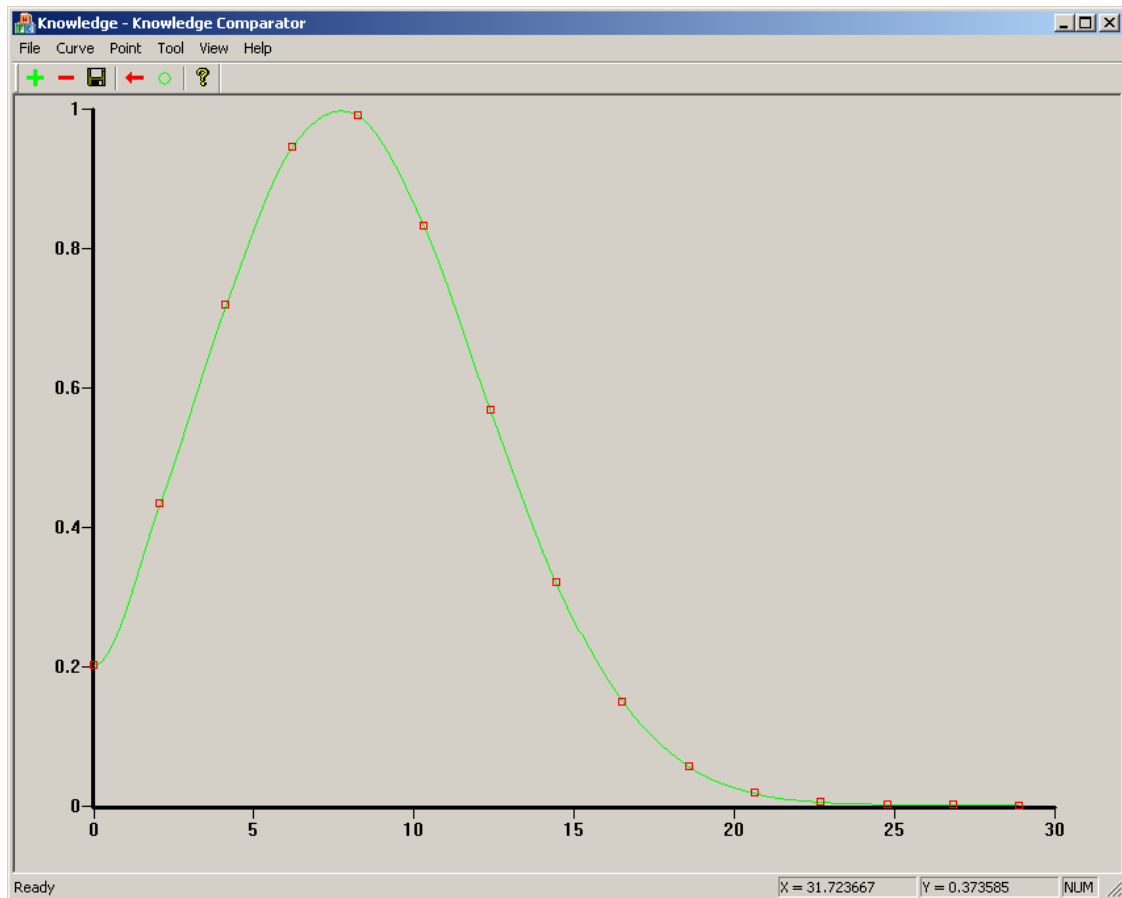
Exercise:

Refine the extracted knowledge by removing outlier polygons in terms of large mode distances and save it to a file (hint: operate in the “Curves By Mode/Peak Match” window).

3. Refine the knowledge using Knowledge Comparator

- 1) In “Knowledge Miner”, click “” toolbar or click “Knowledge --> Compare ...” menu to launch the “Knowledge Comparator” tool.
- 2) In “Knowledge Comparator”, click “” toolbar or right click and select “Add Curve ...”. In the open file dialog, select the saved knowledge. In this example, select “DubuqueDeepOverGradient.txt”. The saved knowledge curve will be reloaded.

Knowledge Miner 1.0 User Manual



- 3) This tool is ideal for knowledge comparison. Curves for different map units (for the same environmental variable) can be plotted on the same screen and edited side-by-side.