NAME: Find Duplicate Shapes or Records v. 1.2

Aka: Find_dupes.avx

Last modified: October 10, 2005

TOPICS: Duplicate Records Identical Overlap Same Compare Attributes Count

AUTHOR: Jeff Jenness
　　　　Wildlife Biologist, GIS Analyst   Email: jeffj@jennessent.com
　　　　Jenness Enterprises   Web Site: http://www.jennessent.com)
　　　　3020 N. Schevene Blvd.   Phone: 1-928-607-4638
　　　　Flagstaff, AZ 86004   USA

DESCRIPTION: This is a fairly basic extension that allows you to identify and optionally delete duplicate records in your dataset. You can define "Duplicate" to mean either identical shapes, identical attributes, or both.

OUTPUT:

- The "Identify Duplicates" function will produce a report summarizing all the duplicate records and optionally add a field to your table identifying these records.

- The "Delete Duplicates" function will produce a new dataset without the duplicated values, without modifying your original theme or table.

- The "Identify Unique Values" function will produce a table listing all unique values, plus counts, of all the specified fields in a table or theme.

REQUIRES: This extension requires that the file "avdlog.dll" be present in the ArcView/BIN32 directory (or $AVBIN/avdlog.dll) and that the Dialog Designer extension be located in your ArcView/ext32 directory, which they usually are if you're running AV 3.1 or better. The Dialog Designer doesn't have to be loaded; it just has to be available. If you are running AV 3.0a, you can download the appropriate files for free from ESRI at:

　http://support.esri.com/index.cfm?fa=downloads.patchesServicePacks.viewPatch&PID=25&MetaID=483

REVISIONS: Version 1.1 (March 24, 2005) adds the option to identify unique values in multiple fields, similar to the Summarize function.

Version 1.21 (October 10, 2005) fixes:

1) A bug that occurs when the current working directory is not writable, producing an error message stating that "A(n) Nil object does not recognize request WriteElt".

2) A bug that occurs when a Date field is selected as the ID Field, producing a message stating that "A(n) Date object does not recognize request SetFormatPrecision.

Recommended Citation Format: For those who wish to cite this extension, the author recommends something similar to:

Jenness, J. 2005. Find Duplicate Shapes or Records (find_dupes.avx) extension for ArcView 3.x, v. 1.21. Jenness Enterprises. Available at: http://www.jennessent.com/arcview/find_dupes.htm.

Please let me know if you cite this extension in a publication (jeffj@jennessent.com). I will update the citation list to include any publications that I am told about.
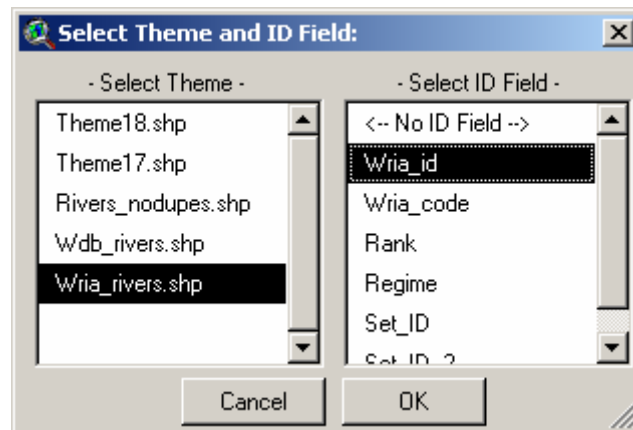

General Instructions:

1) Begin by placing the "**find_dupes.avx**" file into the ArcView extensions directory (../../Av_gis30/Arcview/ext32/).
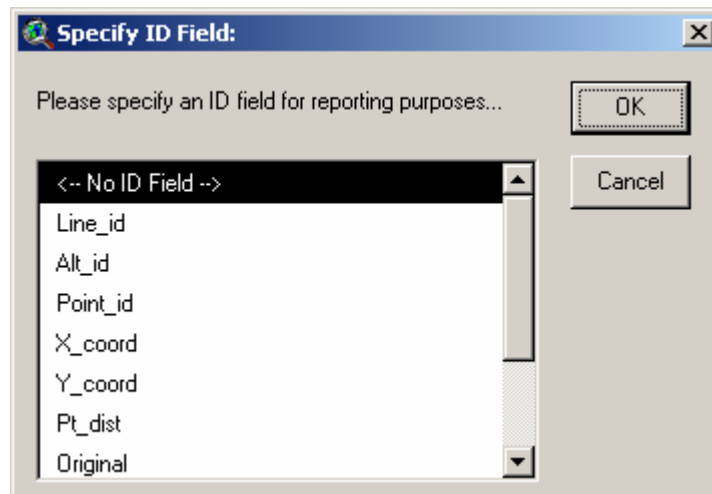
2) After starting ArcView, load the extension by clicking on **<u>F</u>ile --> Extensions… ,** scrolling down through the list of available extensions, and then clicking on the checkbox next to the extension called "**Find Duplicate Shapes or Records.**"

3) This extension will add 3 buttons to both your View and Table button bars. These buttons will not be enabled in your view unless you have at least 1 feature theme in your view.

a. Identifies duplicate records in your table or theme, producing a report and optionally adding a new field to the attribute table identifying any duplicates.

b. Deletes duplicate records from the table or theme.

c. Identifies unique values in specified fields from a table or theme.

Identifying Duplicate Records:

1) Click the [D?] button from your View or Table button bar. If you are in a view, you will first be prompted to identify your theme and ID field:
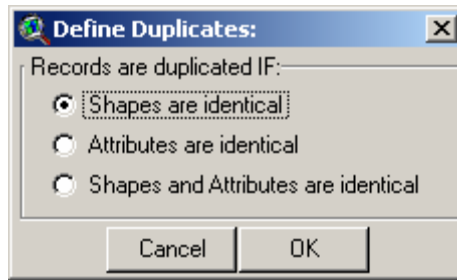

If you are in a Table, then you will only be asked to identify your ID field:

IMPORTANT - A NOTE REGARDING ID FIELDS: The ID values are used primarily for the reports so generally the choice of ID field is not critical. However, if you are looking for duplicates based only on attributes, then the ID values are used to make a first pass on the dataset to speed up the analysis. If you have a choice, the analysis will go faster if you pick an ID field with a relatively large number of unique values. If your dataset has thousands of records, then the analysis will go orders of magnitude faster if your ID field has thousands of unique values than if it has only 1 unique value. If you choose the option "No ID Field", then the extension will briefly examine your data and make a first pass on the dataset using a field that appears to have the most unique values.

2) Next, IF you have shape field in your dataset (i.e. if you are analyzing a theme or a theme attribute table), you will be asked to define what the extension should consider to be "Duplicate".

3

Shapes are identical: This identifies all the points, lines or polygons that are exactly duplicated and overlay each other. They may have different attribute values.

Attributes are identical: This identifies all records that have exactly the same set of attribute values, regardless of the location or shape of the point, line or polygon object.

Shapes and Attributes are identical: Both the shapes and the associated attribute values must be exactly identical in order to be considered a duplicate value.

If you are working with a standard table with no shape component, then your only option is to search for duplicate attribute values and you will not see the dialog above.

3) The extension will then go to work searching for duplicate shapes and/or records. If it finds any, it will ask you if you would like to add a field to the table identifying those duplicate records:
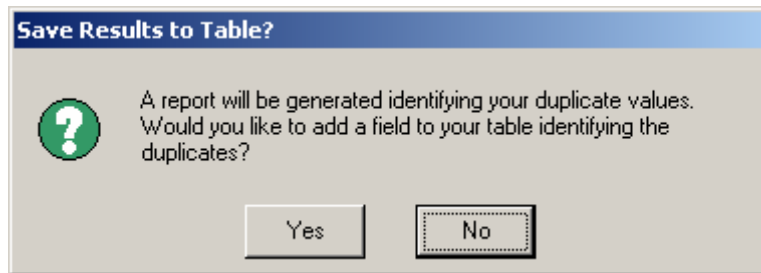
If you click "Yes" to this option, a new field will be added to your original dataset. Duplicate records will be sorted into sets with each set assigned a unique number, and this set number will then be added to the new field. Records that have no duplicates will have no value in that field.

If the extension fails to find any duplicate records, it will go directly to the report below.

4) When all else is done, the extension will produce a report describing what it did and what it found:
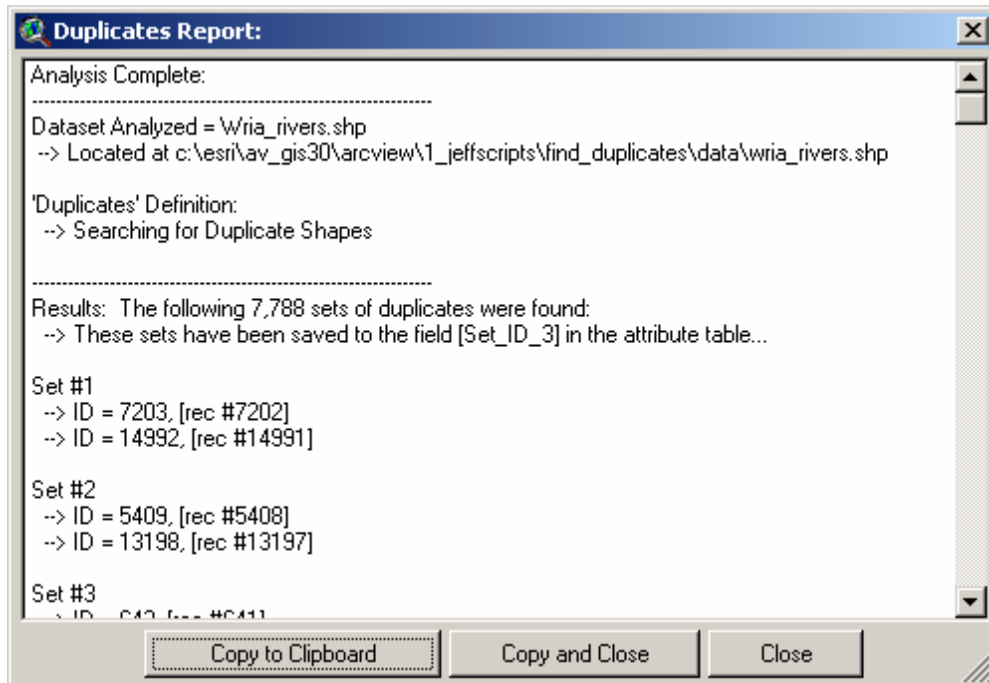
Deleting Duplicate Records:

1) This function behaves in a similar manner to the "Identify Duplicate Records" function described above. Begin by clicking the ⬛ button from your View or Table button bar. If you are in a view, you will first be prompted to identify your theme and ID field:

If you are in a Table, then you will only be asked to identify your ID field:

IMPORTANT - A NOTE REGARDING ID FIELDS: The ID values are used primarily for the reports so generally the choice of ID field is not critical. However, if you are looking for duplicates based only on attributes, then the ID values are used to make a first pass on the dataset to speed up the analysis. If you have a choice, the analysis will go faster if you pick an ID field with a relatively large number of unique values. If your dataset has thousands of records, then the analysis will go orders of magnitude faster if your ID field has thousands of unique values than if it has only 1 unique value. If you choose the option "No ID Field", then the extension will briefly examine your data and make a first pass on the dataset using a field that appears to have the most unique values.

2) Next, IF you have shape field in your dataset (i.e. if you are analyzing a theme or a theme attribute table), you will be asked to define what the extension should consider to be "Duplicate".

Shapes are identical: This identifies all the points, lines or polygons that are exactly duplicated and overlay each other. They may have different attribute values.

Attributes are identical: This identifies all records that have exactly the same set of attribute values, regardless of the location or shape of the point, line or polygon object.

Shapes and Attributes are identical: Both the shapes and the associated attribute values must be exactly identical in order to be considered a duplicate value.

If you are working with a standard table with no shape component, then your only option is to search for duplicate attribute values and you will not see the dialog above.
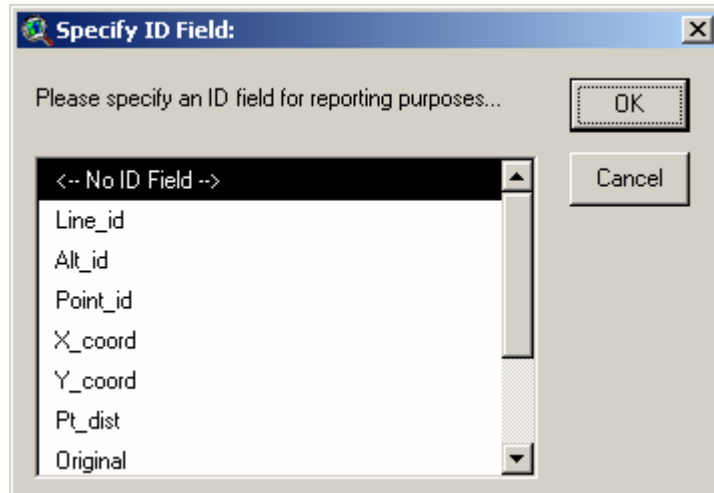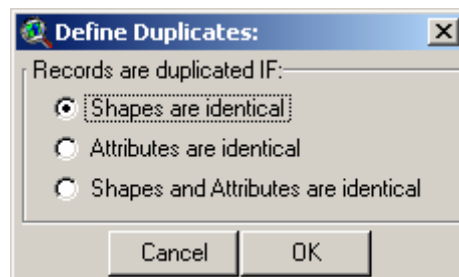
3) The extension will ask you to identify where to save the new file. This function will not modify your original theme or table, but rather create a new one which does not include the duplicate values. It saves the first instance of the duplicate value and deletes all the rest.

6

Please specify a name for your new shapefile:

File Name:
theme24.shp

ponds.dbf
ponds_original.dbf
river_nodups.dbf
table.dbf
table2.dbf
theme.dbf
theme10.dbf
theme11.dbf

Directories:
c:\esri\av_gis30\arcview\1_jeffscript

c:\
esri
av_gis30
arcview
1_jeffscripts
find_duplicates
data

Drives:
c:

OK
Cancel
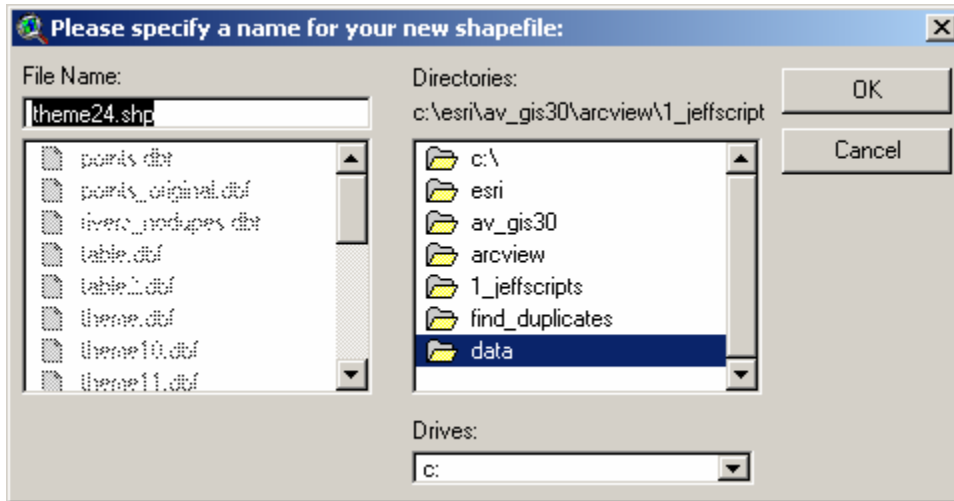
4) The extension will then go to work searching for duplicates. If it does not find any, it will not create a new theme or table. Regardless of whether it finds duplicates, it will produce a report describing exactly what it did:

Duplicates Report:

Analysis Complete:

Dataset Analyzed = Wria_rivers.shp
  --> Located at c:\esri\av_gis30\arcview\1_jeffscripts\find_duplicates\data\wria_rivers.shp
  --> Wria_rivers.shp has 40,424 records...

'Duplicates' Definition:
  --> Searching for Duplicate Shapes

Duplicates saved to:
  --> c:\esri\av_gis30\arcview\1_jeffscripts\find_duplicates\data\theme24.shp
  --> theme24.shp has 32,636 records...

Results: The following 7,788 sets of duplicates were found:

Set #1
  --> ID = 7203, [rec #7202] -- Saved
  --> ID = 14992, [rec #14991] -- Deleted

Set #2
  --> ID = 5409, [rec #5408] -- Saved
  --> ID = 13198, [rec #13197] -- Deleted

Set #3
  --> ID = 642, [rec #641] -- Saved
  --> ID = 8431, [rec #8430] -- Deleted

Copy to Clipboard | Copy and Close | Close

5) Finally, the extension will add either your new theme to your view, or your new table to your project.

Identifying Unique Values:

This function will identify all unique values in a set of fields and report those values, plus counts, in both a table and a text report. This is similar to the Summarize function except that it operates on multiple fields and only provides a count of each unique value in each field, and does not combine field values before summarizing them.

Begin by clicking the [U?] button in your View or Table. If you are in a View, you will first be asked to identify the theme to analyze:
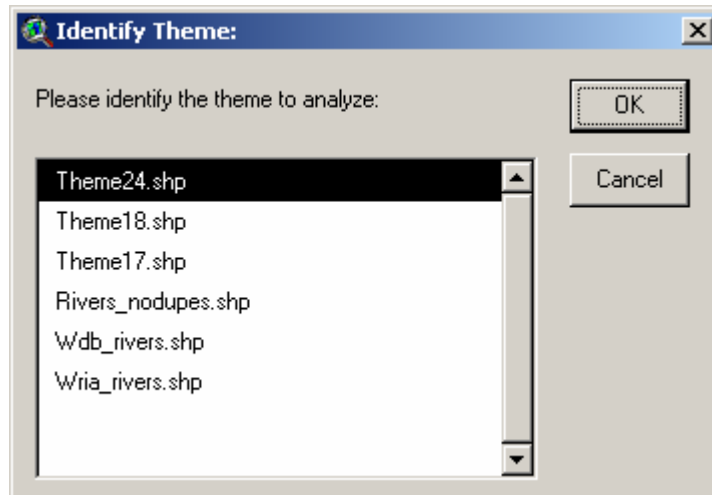
You will then be asked to identify the fields to summarize:
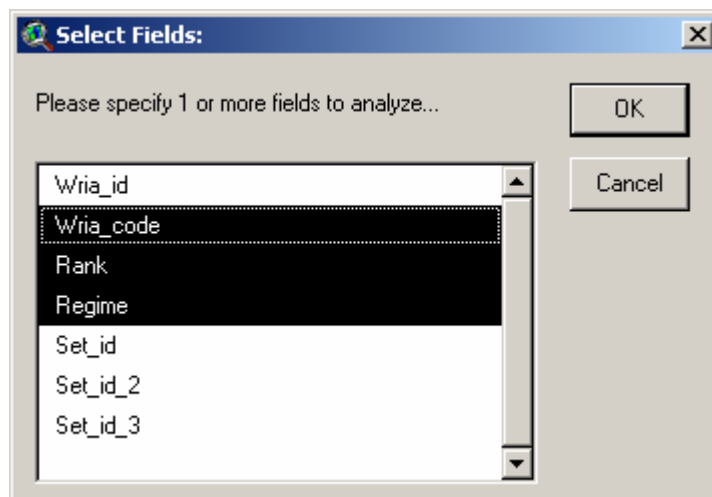
Finally you will be asked where to save your summary table:

Please specify a name for your table:

File Name: value_counts27.dbf

Directories: c:\esri\av_gis30\arcview\1_jeffscript

[OK] [Cancel]

Drives: c:
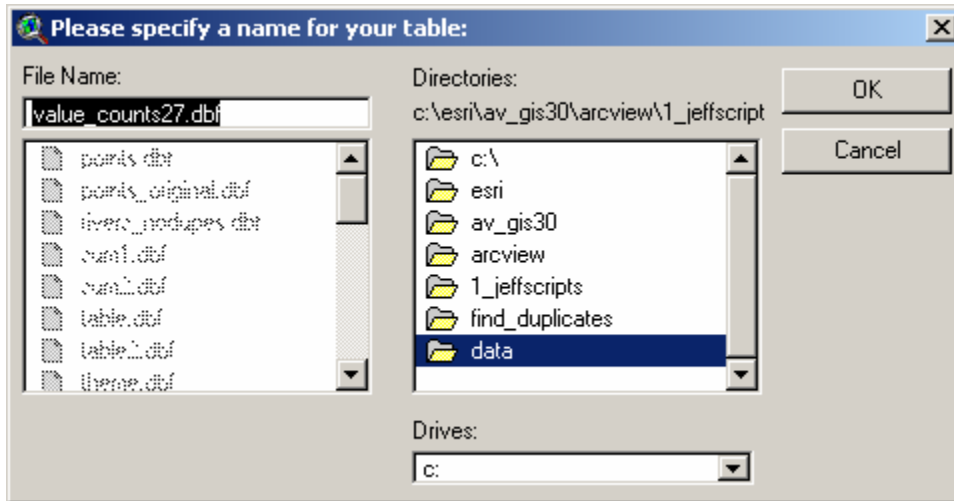
When the extension finishes the analysis, it will produce a table of unique values per field, plus counts of those unique values, and a text report summarizing the results:

Counts of Field Values in 'Theme24.shp' [value_counts27.dbf]

Unique_ID	Whia_code	Whia_code Count	Rank	Rank Count	Regime	Regime Count
0	1	8	0	9	-1	99
1	3	1	1	17099	1	21624
2	11	10043	2	8201	2	10911
3	12	7056	3	3731	3	2
4	21	5656	4	1973		
5	22	2545	5	762		
6	31	2775	6	633		
7	32	956	7	129		
8	41	1671	10	99		
9	42	301				
10	43	1				
11	51	710				
12	52	52				
13	61	632				
14	62	1				
15	71	129				
16	99	99				

9

```
Duplicates Report:                                                    ×

Analysis Complete:
--------------------------------------------------------
Dataset Analyzed = Theme 'Theme24.shp' in View 'View1'
 --> Located at c:\esri\av_gis30\arcview\1_jeffscripts\find_duplicates\data\theme24.dbf


The following fields were analyzed:
 --> Wria_code [Field name = Wria_code]
 --> Rank [Field name = Rank]
 --> Regime [Field name = Regime]

All 32,636 records analyzed...

--------------------------------------------------------
UNIQUE VALUES FOUND:
 --> Wria_code: 1 (n=8); 3 (n=1); 11 (n=10,043); 12 (n=7,056); 21 (n=5,656); 22 (n=2,545); 31
(n=2,775); 32 (n=956); 41 (n=1,671); 42 (n=301); 43 (n=1); 51 (n=710); 52 (n=52); 61 (n=632); 62
(n=1); 71 (n=129); 99 (n=99)
 --> Rank: 0 (n=9); 1 (n=17,099); 2 (n=8,201); 3 (n=3,731); 4 (n=1,973); 5 (n=762); 6 (n=633); 7
(n=129); 10 (n=99)
 --> Regime: -1 (n=99); 1 (n=21,624); 2 (n=10,911); 3 (n=2)
--------------------------------------------------------
Analysis Began: March 24, 10:44:08 AM
Analysis Complete: March 24, 10:44:11 AM
Time Elapsed: 3 seconds...

       Copy to Clipboard          Copy and Close          Close
```

Enjoy! Please contact the author if you have problems or find bugs.

Jeff Jenness	jeffj@jennessent.com
Jenness Enterprises	http://www.jennessent.com
3020 N. Schevene Blvd.	(928) 607-4638
Flagstaff, AZ 86004	
USA	

Updates to this extension and an on-line version of this manual are available at

http://www.jennessent.com/arcview/find_dupes.htm

Please visit *Jenness Enterprises* ArcView Extensions site for more ArcView Extensions and other software by the author. We also offer customized ArcView-based GIS consultation services to help you meet your specific data analysis and application development needs.