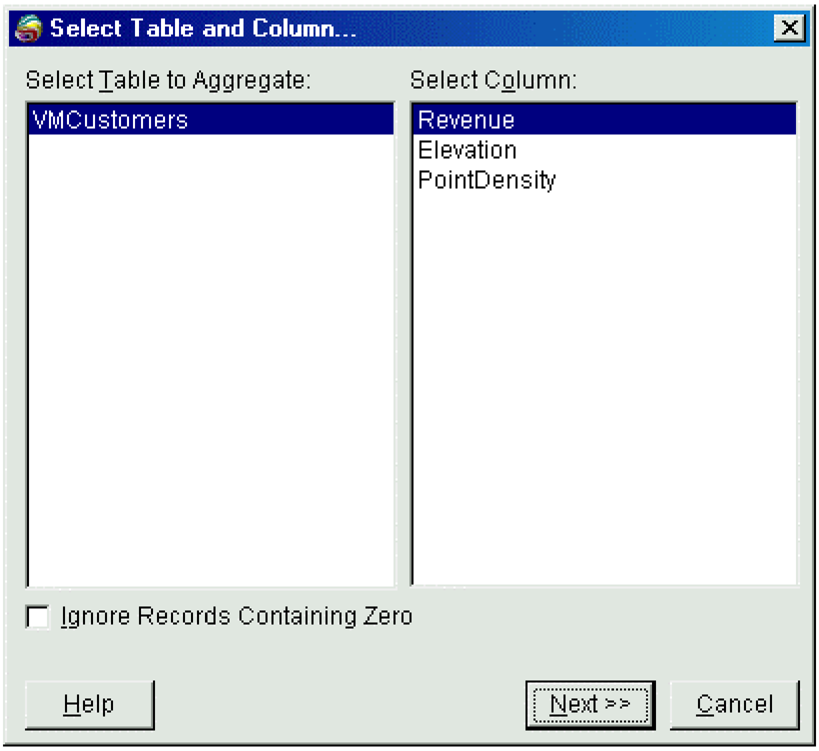
Point Aggregation with Statistics

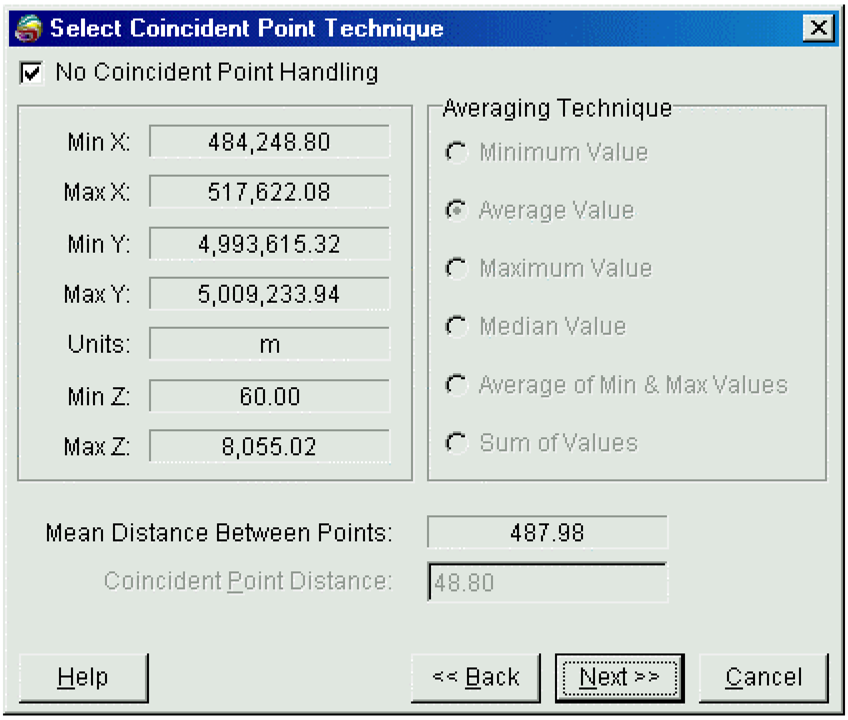
The *Simple Point Aggregation* technique is fast and effective but does not reveal much about the mathematical and spatial characteristics of the aggregation process. To quantitatively measure the aggregation process, *Vertical Mapper* employs three other techniques, namely *Forward Stepping*, *Cluster Density*, and *Square Cell*, two of which take into consideration the nature of distribution of the data points. These methods are grouped under the category *Point Aggregation With Statistics*.

Toaccess the ***Point Aggregation With Statistics*** command, from the *Vertical Mapper* pull-down menu, choose the command *Data Aggregation > Point Aggregation With Statistics.*


Select Table and Column...
Select Table to Aggregate:
VMCustomers
Select Column:
Revenue
Elevation
PointDensity
□ Ignore Records Containing Zero
Help
Next >>
Cancel

- From the ***Select Tables and Column*** dialogue box, select the point table to aggregate from the list of open MapInfo tables and select the column that contains the data to be transferred to the new aggregated file. Check ***Ignore Records Containing Zero*** to include only non-zero records. Once this dialogue has been completed, choose the ***Next >>*** button.

1 The ***Select Coincident Point Technique*** dialogue that appears next is the same as the one used to set parameters for simple aggregation. It is important to note that statistical aggregation proceeds in two steps. This first step is designed to allow the user to deal with virtually coincident points separately before proceeding with further aggregation. For instance, in the soil sample collection example mentioned earlier, the user may want to average all the samples collected at each site before proceeding to average and sum all the sample sites which occur in a given unit area.



Select Coincident Point Technique

☑ No Coincident Point Handling

Min X: 484,248.80

Max X: 517,622.08

Min Y: 4,993,615.32

Max Y: 5,009,233.94

Units: m

Min Z: 60.00

Max Z: 8,055.02

Averaging Technique

○ Minimum Value

● Average Value

○ Maximum Value

○ Median Value

○ Average of Min & Max Values

○ Sum of Values

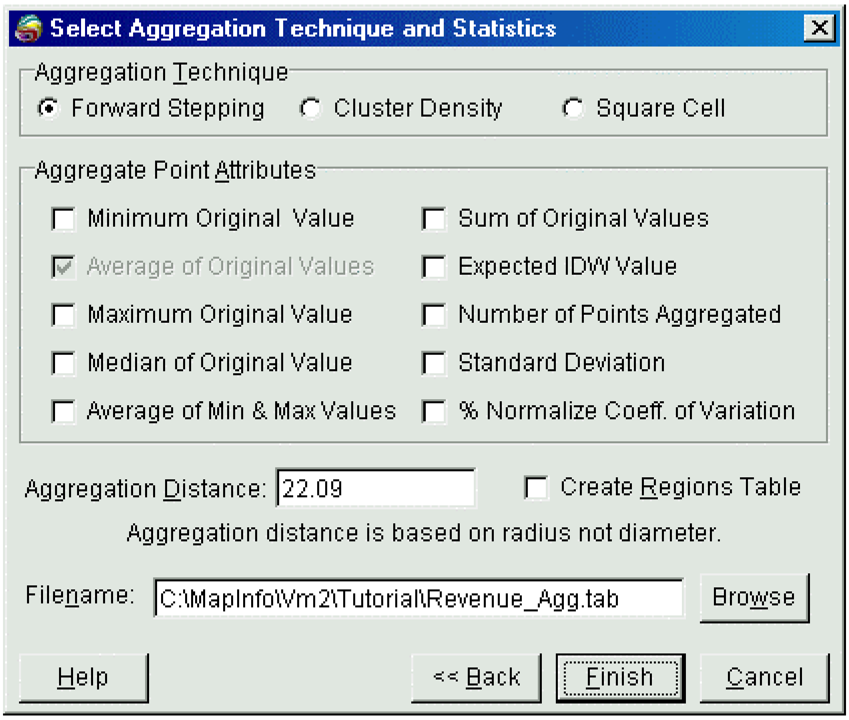
Mean Distance Between Points: 487.98

Coincident Point Distance: 48.80

1. This dialogue box presents the same parameters as the *Simple Point Aggregation* dialogue and has been discussed in the previous section. There is the addition, however, of the ***Mean Distance Between Points*** parameter.

1 By default, the ***No Coincident Point Handling*** check box is selected and the entire dialogue is greyed. It is assumed that, in most cases, aggregation of virtually coincident points is not required before proceeding with statistical aggregation and the settings in this dialogue will be ignored. This statement does not hold true for *Cluster Density* aggregation. With *Cluster Density*, coincident point handling is always performed regardless of the dialogue being greyed out. If the user does not choose any of the settings on this dialogue, the default values will be used. A reminder to this fact will appear when this technique is chosen.

- Once this dialogue has been completed, choose the ***Next*** >> button. The ***Select Aggregation Technique and Statistics*** dialogue will appear.



Select Aggregation Technique and Statistics

Aggregation Technique

● Forward Stepping ○ Cluster Density ○ Square Cell

Aggregate Point Attributes

□ Minimum Original Value	□ Sum of Original Values
☑ Average of Original Values	□ Expected IDW Value
□ Maximum Original Value	□ Number of Points Aggregated
□ Median of Original Value	□ Standard Deviation
□ Average of Min & Max Values	□ % Normalize Coeff. of Variation

Aggregation Distance: 22.09 □ Create Regions Table

Aggregation distance is based on radius not diameter.

Filename: C:\MapInfo\Vm2\Tutorial\Revenue_Agg.tab

1 The ***Aggregation Techniques*** section of the dialogue allows the user to select one of the three aggregation techniques supported by *Vertical Mapper*.

2 The ***Aggregate Point Attributes*** section allows the user to choose the statistical parameters, computed during the aggregation process, that will be passed to the aggregated point file and region file (if selected). The options are discussed below.

a) ***Minimum Original Value:*** The minimum value of all values of every point selected to be aggregated.

b) ***Average of Original Value:*** The average value of all the points selected to be aggregated. Note that this value must <u>always</u> be passed to the new aggregated tables.

c) ***Maximum Original Value:*** The maximum value of all values of every point selected to be aggregated.

d) ***Median Original Value:*** The median value is the middle occurring value of all points selected to be aggregated. If there is an even number of points, the median value is the average of the two middle values.

e) ***Average of Min & Max:*** This value is the average between the minimum value and the maximum value.

f) ***Sum of Original Values:*** The value obtained by summing all values of all points selected to be aggregated.

g) ***Expected IDW (Inverse Distance Weighted) Value:*** The calculated value is obtained by summing the weighted value for each point selected to be aggregated, and dividing this value-weighted sum by the sum of the weights. The weight associated with each value is inversely proportional to the normalized distance between the value and the reference location, raised to the exponent 2. The normalized distance is the actual distance from the reference point divided by 1.01 times the "same point" distance. Therefore, values close to the reference point will have more influence than points farther from the reference point. The inverse distance weighted value is considered to be a reasonable estimate of the value at the aggregated coordinate.

h) ***Number of Points Aggregated:*** This value is the total number of points that were selected to be aggregated. It does not include the points processed in the *Coincident Point Handling* step.

i) ***Standard Deviation:*** This value measures the degree of dispersion about the mean of the values for those points selected for aggregation.

j) ***% Normalized Coeff. Of Variation:*** The coefficient of variation is the standard deviation divided by the average expressed as a percentage (i.e., multiplied by 100). The value is dimensionless and is representative of data dispersion, especially when the average is not close to zero. A "normalized" coefficient of variation is calculated using the standard deviation divided by the difference between the average and the minimum value i.e., the minimum value of the complete data set not just the data in the local aggregated region. The coefficient can still be large when the average value is close to the minimum value but the normalized coefficient is representative of the original range of values and not their absolute value.

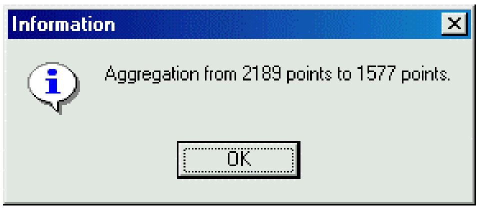
1 The ***Aggregation Distance*** setting is the distance used to group points for aggregation. This distance has a different meaning for each of the three methods. For the *Cluster Density* and *Forward Stepping* techniques, *Aggregation Distance* is defined by the radius of a user-specified circular search area centred about each aggregation cell. For the *Square Cell* technique, *Aggregation Distance* is defined by width of the square aggregation cell.

2 Selecting the ***Create Regions Table*** check box will build a MapInfo table of the regions used to group the point data for aggregation. This option is useful in allowing the user to visually inspect the results of the aggregation process. The order in which these regions are created is the same order in which the point file was aggregated. Therefore by opening a Browser window of the region file and selecting each record in the list one at a time, the user can visually see the process order in the Map window.

Tip: Use the created aggregation regions to produce a MapInfo coloured thematic map, where each region is thematically shaded according to one of the computed statistical values.

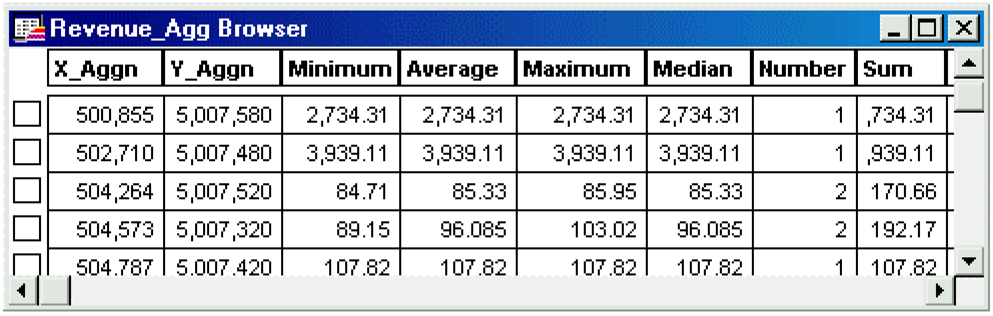
- Enter a unique file name in the ***Filename*** edit box and select either the ***Finish*** button to complete the aggregation process or, if modifications to the previous dialogue are required, the ***Back*** >> button to return to one or more dialogues back.

- Once the process is complete, a dialogue will appear stating the extent of the aggregation as shown below. Also, the aggregated point file will appear in a new Map window with a default symbol style applied to each point. If *the Create Regions Table* option was selected, then the aggregation regions table will also appear in a separate Map window and is automatically assigned the suffix "*AggRegion*" to the file name.



Aggregation from 2189 points to 1577 points.

- The statistical information calculated is retained in the Browser window of both the aggregated point file and the aggregation regions file as shown below.



	X_Aggn	Y_Aggn	Minimum	Average	Maximum	Median	Number	Sum
□	500,855	5,007,580	2,734.31	2,734.31	2,734.31	2,734.31	1	,734.31
□	502,710	5,007,480	3,939.11	3,939.11	3,939.11	3,939.11	1	,939.11
□	504,264	5,007,520	84.71	85.33	85.95	85.33	2	170.66
□	504,573	5,007,320	89.15	96.085	103.02	96.085	2	192.17
□	504,787	5,007,420	107.82	107.82	107.82	107.82	1	107.82

Forward Stepping Aggregation

The ***Forward Stepping Aggregation*** technique is an appropriate method for any general aggregation application due to its speed and effectiveness. This technique could be employed when your data has a truly random distribution or when the other two techniques are not appropriate.

This aggregation process can be simply stated as one which aggregates by moving through the data set from left to right and then top to bottom. The process begins by sorting the data points into rows that are generally three times the aggregation distance. This sorting is performed to aid the program in determining where the aggregation will begin each left-to-right swath. There are no settings the user can set to alter this process. Beginning with the data point in the upper left (northwest) corner of the data set, a circular search radius is created as specified by the *Aggregation Distance* setting. Then all the data points which fall inside this search radius are selected and flagged. This is to prevent these data points from being aggregated to another location. The geocentre of the selected points is then determined. This becomes the location of the new aggregated point. The last step is to perform the aggregation calculations on the selected points as specified in the Point Aggregation dialogue, and the results are then attributed to the new aggregated point.

Once the processing for the first point is complete, the procedure sweeps from left to right and top to bottom across the study area selecting and aggregating unflagged points. It is important to note that not every data point will be aggregated on the first pass through the data set. Because of this a second is normally required to aggregate those data points missed on the first pass. The results are shown in

Figure 12.3 below.

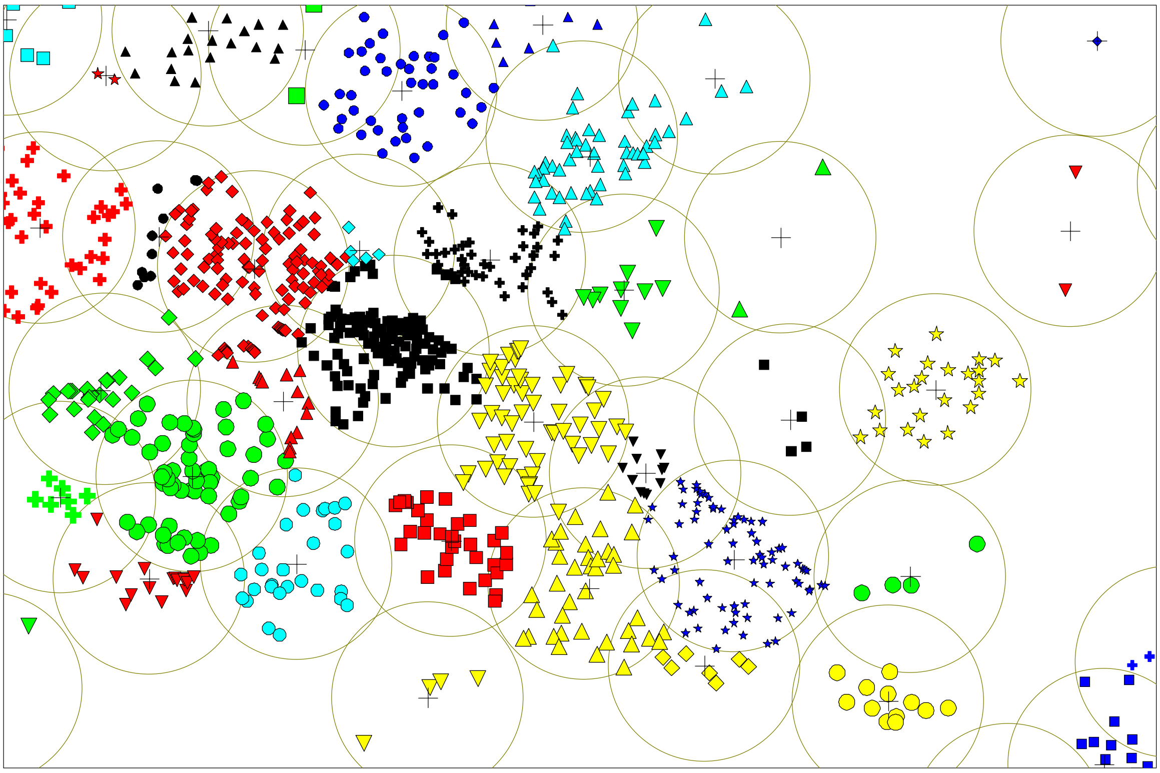


Figure 12.3. Example of circle aggregation using the Forward Stepping aggregation technique. The shaded points are the original data points. The crosses represent the newly aggregated points, and the circles represent the aggregation region. The original points have been coded to help show which aggregation region they belong to.

In Figure 12.3 the user should notice the inappropriate aggregation decisions that are made in certain locations of the point file as well as the degree of overlap of the aggregation regions. In the upper left corner of the diagram there are two examples of an inappropriate aggregation, marked by the letters **A** and **B**. In both cases one would likely aggregate these points differently if it were to be done manually. The reason these points get aggregated this way has to do with the two aggregation passes this technique performs; the second pass aggregates the remaining unflagged points, resulting in a large degree of overlap of the aggregation regions. Some of the aggregation regions in the above diagram have been numbered to show the aggregation process order. The letters show which points were aggregated on the second pass.

Cluster Density Aggregation

The ***Cluster Density*** technique is typically used when there is a visual clustering effect occurring in the dispersion of the data points. For example, demographic data representing rural areas may often exhibit a ‘shot-gun’ pattern (see Figure 12.4). Data from each small community is considerably more densely distributed than in the surrounding countryside. But this is not to say that the *Cluster Density* technique can only be performed on clustered data, any type of distribution can be processed. *Cluster Density* does, however, on large data sets process more slowly than the other aggregation techniques but is not significantly slower on small to medium size data sets.

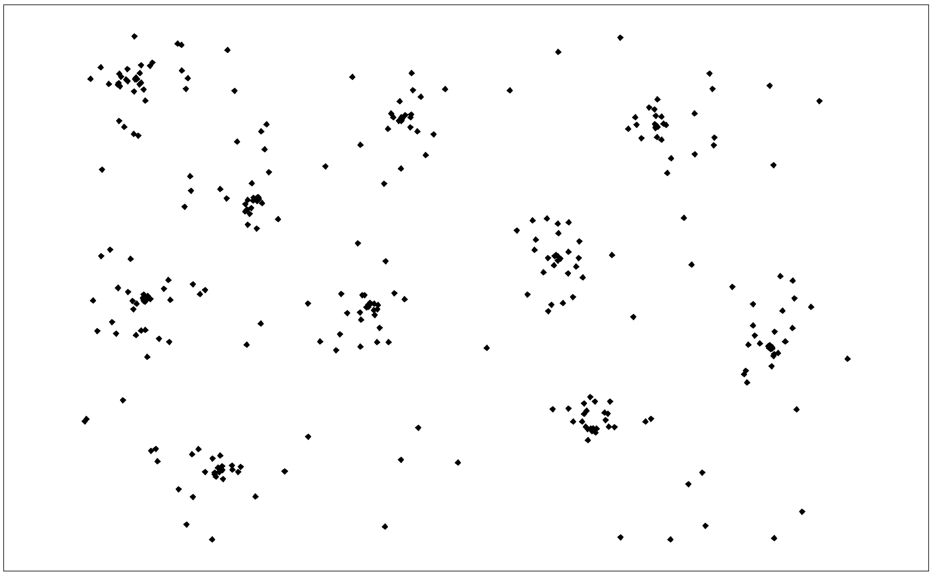


Figure 12.4. Original distribution of data points.

The *Cluster Density* technique looks at the entire data set prior to aggregation and determines the single most densely populated area that would fall inside the user-specified search radius (Aggregation Distance). Points that fall inside the search radius are selected and flagged. The geocentre of these points is calculated and that position becomes the location of the new aggregated point. Calculations are performed on the values of the selected points (as specified in the Point Aggregation dialogue) and the results are attributed to the new geocentred point. After this, the area with the second highest density of points is chosen and the process is repeated. At each stage, the entire remaining data set must be examined for its density patterns to avoid using previously aggregated points and to factor in the removed points in the density analysis.

As mentioned earlier *Cluster Density* aggregation always includes coincident point handling regardless of whether or not the user requested this type of handling. The reason for this is that the point density calculation used in the aggregation technique does not handle points that occupy the same space (coincident). If the user does not choose any of the settings on the *Coincident Point Handling* dialogue, the default values will be used. A warning message will appear reminding the user of this fact when this technique is chosen.

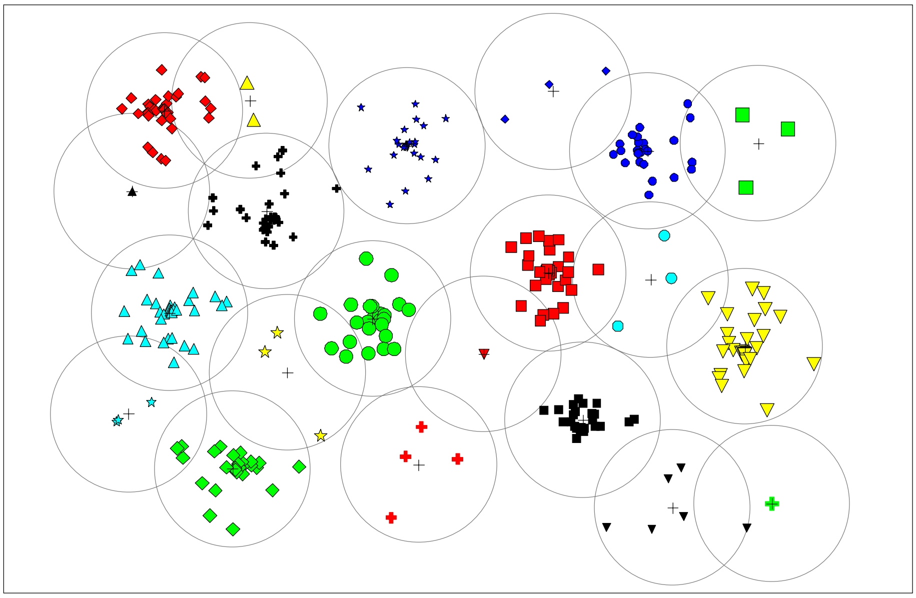


Figure 12.5. Circle aggregation based on the unbiased cluster density technique

There are a total of 330 points in the sample data set shown on the previous figure. It is obvious, after using the *Cluster Density* aggregator, that the data is clustered in about ten separate areas. The more randomly distributed points lying outside these clustered zones (circles) are also aggregated, but include significantly fewer points. Although many circles overlap, the degree of overlap is significantly less than the *Forward Stepping* method. One could then assume that there are far less occurrences of inappropriate aggregations.

Note: Generally, the *Cluster Density* aggregation technique is the best technique for most small to medium size data sets, because it makes better decisions during the aggregation process, i.e., it aggregates the more densely clustered points first and has a reasonable processing time. However, it is not appropriate if the *Number of Points Aggregated* statistic is required and there are coincident points in the data set, because coincident points are always aggregated first and are not included in the statistics appended to the new point file. An example of when this will occur is with the aggregation of crime data. Typically, crimes are committed at the same location and therefore this is reflected in the point file. The fact that there was more than one crime committed at any given location is important. The only alternative is to use the *Forward Stepping* aggregation technique.

Square Cell Aggregation

The ***Square Cell*** technique is generally used when values are required that are representative of specific area, e.g., create a density map of the number of new housing units per square kilometre, or when trying to avoid any areas of overlap of the aggregation regions. This method divides the area covered by the point file into adjacent squares determined by *Aggregation Distance*. The points that fall inside any of these squares are aggregated to a new point created at the geocentre of the aggregated points (not the centre of the square). As with the other two techniques, the specified statistical information is then attached as attributes to the new aggregated point.

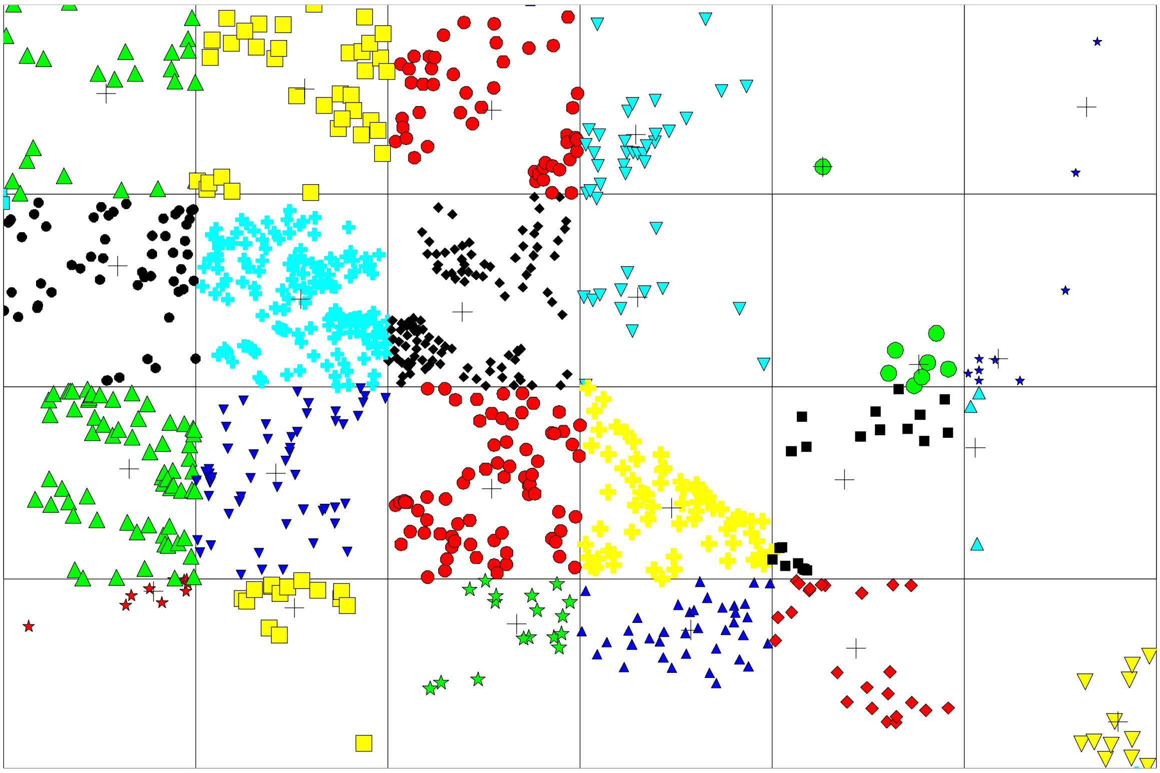


Figure 12.6. Example of square aggregation

Although there is no overlap of the aggregation regions in Figure 12.6, there are several areas where points have been aggregated inappropriately. Therefore the best results require a certain degree of overlap.