Near Repeat Calculator

Program manual for

Version 2.0

TABLE OF CONTENTS

DISCLAIMER

The software program, Near Repeat Calculator (hereafter referred to as the "program"), was supported by Grant 2006-IJ-CX-K006 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. Points of view in this document are those of the author and do not necessarily represent the official position or policies of the US Department of Justice.

ABOUT THE PROGRAM

The development of this computer program was made possible through a grant from the Office of Research and Evaluation, National Institute of Justice (NIJ), Washington, DC, grant number 2006-IJ-CX-K006. The software was developed by Jerry H. Ratcliffe, of the Department of Criminal Justice at Temple University, Philadelphia PA, however the BANUS group were instrumental in pioneering the approach (Shane Johnson, Kate Bowers, Michael Townsley, Henk Elffers, Wim Bernasco, & George Rengert).

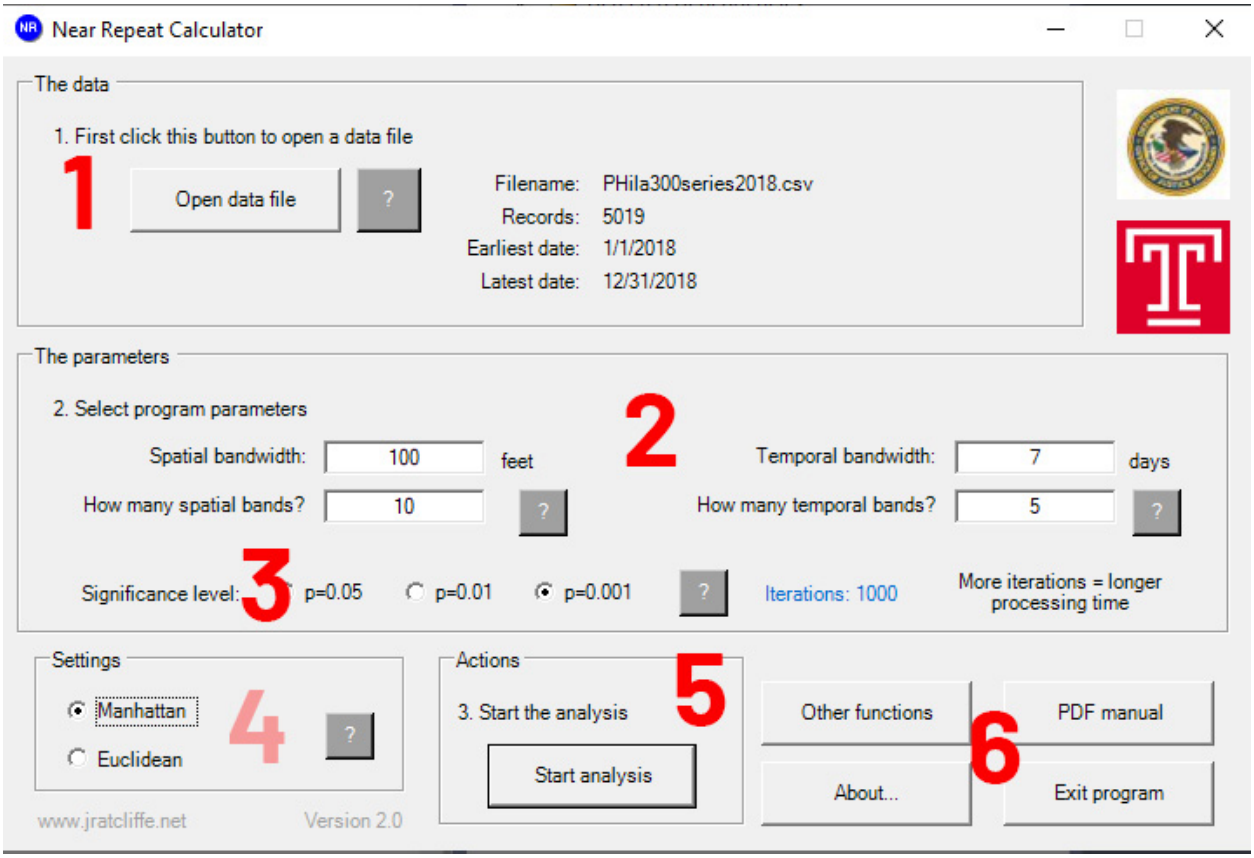
For details of disclaimer and other details, see www.jratcliffe.net or this file.

Please check www.jratcliffe.net for updates.

The recommended citation is: Ratcliffe, JH, Near Repeat Calculator (version 2.0). Temple University, Philadelphia, PA and the National Institute of Justice, Washington, DC. May 2008 (updated March 2020).

PROGRAM PARAMETERS AND THE MAIN DIALOG BOX

The main dialog box for the program is shown here. Functionality for areas shown in red is explained below.



When the program opens, only the data section (1) is available. Open a data file here. For brief help, click the '?' button. See the Data file section of this manual for detail of the acceptable format.

1. Once a data set is loaded, you can adjust the program parameters in the 'parameters' section (2). The three main settings that are adjustable are the Spatial bandwidth and spatial bands, the Temporal bandwidth and temporal bands, and the Significance level (3). When in the program, you can click the relevant '?' button for help.
2. The distance settings are adjustable (4). See the Distance settings (Manhattan/Euclidean) section for more details.
3. The 'Start analysis' button commences the analysis (5).
4. Some buttons at (6) open this pdf help manual from within the program, explain program origins, conduct other helpful functions, and close the program.

DATA FILE

There is one data file format for this program. **The program accepts text files in comma separated values (*.csv) format**. This is a common type of output from most GIS programs and from Microsoft Excel (please see the help sections of those programs for further details). The data requirements for this program are simple. Each data point is expected to have an x-coordinate, a y-coordinate, and a date value. Each row of the data file should contain a single record. Do not include any header rows of information as these will be ignored by the program. The data format is simply x, y, date; as follows:

24326, 123978, 8/12/06

11698, 335122, 8/11/06

22100, 290888, 8/11/06

(and so on)

To open a data file, click 'Open data file', select the data file, and click 'Open'.

The x and y coordinates (in that order) should indicate the projected location of the crime event. The date should represent the date of the crime. If your data are in a text file that does not have a *.csv suffix, then when the open file dialog is open, click 'Files of type' and select from the dropdown list either text files (*.txt) or all files (*.*).

Including a header row will not cause the program to stop and will not cause an error, unless the header row happens to be similar in format (number, number, date) to the expected data format. However if the header row is text, the program will simply ignore the row and later report that there was one row of data that were not included in the analysis.

The format of the date is fairly flexible, as long as the date format complies with standard Microsoft date formatting. Acceptable examples include mm/dd/yyyy or mm-dd-yy. If you are using a computer where the date format is set to European format where the day precedes the month (for example, dd/mm/yy) then the program should detect this and expect data in this format accordingly.

Data are accepted in any projected format and in meters or feet. At present, the program is unable to accept data in latitude and longitude because distance calculations are not currently programmed to function. As there are other ways to achieve this approach, or data can be converted in a GIS, this functionality is not currently planned for future versions.

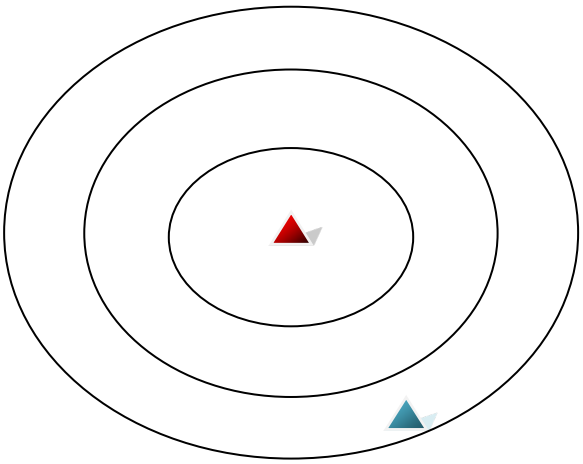
SPATIAL BANDWIDTH AND SPATIAL BANDS

The program looks for unusual patterns in the spatio-temporal relationships between all points in the data set. To make any interesting relationships easier to interpret, spatial patterns should be disaggregated into distance bands. For example, if a city has a block pattern where the length of each block is about 400 feet, then a spatial bandwidth of 400 feet might be appropriate. Any interesting results can then be estimated as

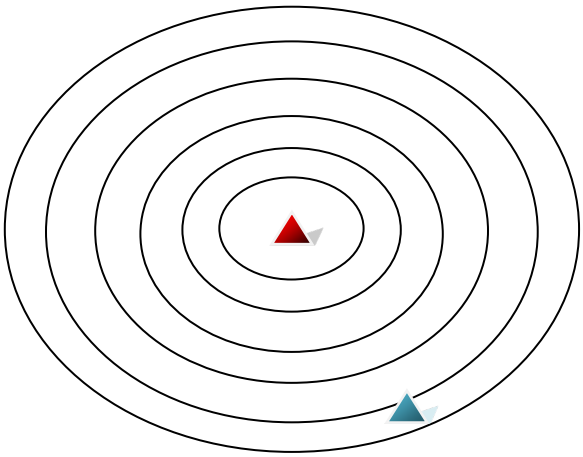
affecting a certain number of blocks from each data point. The choice of a spatial bandwidth could therefore reflect a feature of the urban landscape.

The number of spatial bands is dependent on how far you expect a pattern of near repeats to extend. For example, most environmental criminology research suggests that near repeat patterns occur for only a few blocks or a few hundred meters at most. Any effects appear to peter out beyond this distance. Adding additional bands beyond any identified effects rarely adds much value; however the program is often most effective when you experiment with various settings. For fairly large data sets, ten spatial bands is often a good starting point, though experimentation is encouraged.

These concepts are demonstrated in this picture. In the first scenario, there are three spatial bands (the first starts at the red location). If each band is 800 feet wide, then the eventual display would have five distance categories (same location, >0 to 800 feet, >800 to 1600 feet, >1600 to 2400 feet, and > 2400 feet). The second scenario shows six spatial bands or 400 feet. The second scenario provides greater accuracy in determining the spatial extent of the near repeat phenomenon, as long as there are sufficient data points to populate the increased number of categories.



Three spatial bands with moderate bandwidths. The blue site is located in the third spatial band.



Six spatial bands with narrow bandwidths. The blue site is located in the sixth spatial band.

Be advised that selecting too many bands or too narrow a bandwidth can reduce the number of observations in each category and potentially limit the findings by creating too many categories with low (or zero) values in the observed or expected matrices. If you have zero event-pairs in an analysis category it will show as zero events in the observed frequencies table, and analytical output tables will show **NaN**, indicating that the software could not calculate a number for that cell (NaN means Not a Number).

TEMPORAL BANDWIDTH AND TEMPORAL BANDS

The program looks for unusual patterns in the spatio-temporal relationships between all points in the data set. To make any interesting relationships easier to interpret, temporal patterns (the number of days between

events) should be disaggregated into temporal bands. For example, if a burglary occurred at a house and then another burglary occurred two days later at a neighboring home, then that might be interesting. If there were an unusually high number of incidents that were close in time and space, then this would definitely be worth knowing. Investigators may therefore feel that a temporal bandwidth of a week – 7 days – is a suitable setting. Alternatively, a month - 28 or 30 days – is a common choice for a temporal bandwidth.

The number of temporal bands is dependent on how long you expect a pattern of near repeats to extend. For example, the research on repeat victimization suggests that a risk of repeat burglary increases rapidly after an initial burglary, but that this risk dissipates in the months after the incident and the risk returns to the background (normal level for the area) after a few months. Settings of a 30 day temporal bandwidth with 12 temporal bands, or 14 days with 13 temporal bands (to cover a six month period) are common, but experimentation is encouraged.

MONTE CARLO ITERATIONS

The program compares the actual pattern of spatio-temporal relationships between all points (called the observed pattern) with the pattern one would expect if there were no near repeat process taking place (called the expected pattern). The expected pattern is derived from a redistribution of date values randomly reallocated to the spatial points. For this process to be statistically valid, this random reallocation has to be performed many times. Within social sciences, the standard minimum threshold for statistical significance is $p = 0.05$. This can be achieved with 20 reallocations (called iterations). The best statistical level the program can achieve is $p = 0.001$, reached with 1000 iterations. The greater the number of iterations, the more reliable the result; however, more iterations takes a lot longer to process.

The best way is to start with a quicker analysis, say $p = 0.05$, and then see if you can accept the extra processing time for more iterations. Statistically, $p = 0.001$ is the best (statisticians might say 'robust') though $p = 0.01$ would produce results that are statistically valid and universally acceptable. The program will highlight relationships as important if the difference between the observed pattern and the expected pattern is quite large, and if the statistical significance is at least to 0.05 or lower. A statistical significance value of 0.05 means that if the null hypothesis is true (there is no near repeat pattern) and if you performed the analysis a large number of times and in the same way, you would still get the same or greater difference between the expected results and the observed results five percent of the time. With 999 iterations it is possible to get this number down to 0.001 – the chance of an error by chance then being one in a thousand.

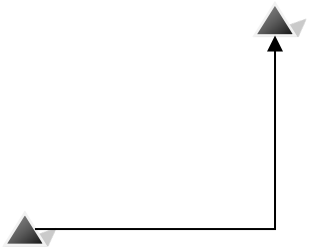
DISTANCE UNITS

Once you load a data file, the program will ask for the distance units. The main reason for this is to improve the quality of the output by labeling the tables appropriately. Select from the four choices. An incorrect selection will not adversely hamper the program but may confuse people reading the output when feet are mislabeled as meters, or similar. Please note that the program is not currently set up to work with data in latitude and longitude format. Please reformat data into a projected format before using the program.

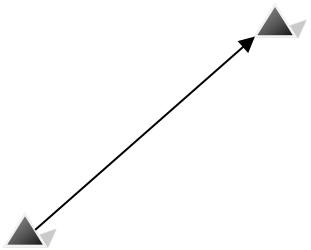
DISTANCE SETTINGS (MANHATTAN/EUCLIDEAN)

Manhattan and Euclidean are two different ways of measuring the distance between two points. Manhattan distance simply adds the different between the x coordinates of two points to the difference between the y coordinates of two points. It is the same as travelling from point to point first horizontally and then vertically. Euclidean distance uses the Pythagorean equation to measure the direct ('crows flight') distance between the points; that is, the square root of the summed squared horizontal and vertical distances. In the program, the default option is Manhattan.

Manhattan is the default because, for urban environments where it is not possible to calculate actual route between two points, research suggests that the Manhattan distance is a closer approximation of the actual route between points than the Euclidean measure which has a tendency to underestimate distances to a greater degree. Details of the research can be found in Chainey and Ratcliffe (2005) *GIS and Crime Mapping* (Wiley: London) and Rossmo (2000) *Geographic Profiling* (CRC Press: Boco Raton).



Manhattan distance



Euclidean distance

PROGRAM OUTPUT

When the program finishes, the program creates two output files.

Both the summary htm file and a comma-separated-values output file can be found in the same folder as the source data set. The verbose file (csv) contains a wealth of information for researchers wishing to understand what occurred during each iteration of the Monte Carlo process. See the section of this manual called

Verbose file details for an explanation of the contents of this file.

The main output is an htm file that can be read with a standard web browser program. If your computer is set up with a default program to handle internet files, then once the program finishes analyzing the data, your browser will automatically open and show the summary file. Details of the contents of the summary file can be found in the following section on Summary file details.

SUMMARY FILE DETAILS

The first section of the program provides a text summary of the findings of the analysis.

This section is created by a program function that examines your data for you and estimates a brief synopsis of the findings. The threshold for what counts as significant to the program are any results that indicate an over-representation of events that are close to an original event in both time and space. Events are considered an over-representation if the pattern has a p value < 0.05, and having an odds ratio of at least 1.20 – in other words suggesting that the any increased occurrence of events is at least 20 percent greater than the pattern we would expect by chance.

The summary section examines both near repeats and repeat victimization (events occurring at the same location).

Observed over mean expected frequencies table

Two tables then follow. The first shows the observed over mean expected frequencies table. This ratio is the difference between the average expected number of point-to-point space time links expected in each cell, and the actual number that were found with your data. The higher the number, the greater the importance and the difference between your data and the expected amount if no pattern existed. Colors are used to indicate statistical probability at both the best possible level for your chosen number of iterations, and the commonly accepted social science threshold of p < 0.05.

Observed over mean expected frequencies table

	Same day	>0 to 7 days	>7 to 14 days	>14 to 21 days
Same location	7.41	2.48	1.89	1.74
>0 to 100 feet	1.10	1.10	0.70	0.93
>100 to 200 feet	2.42	1.20	1.16	0.81
>200 to 300 feet	0.55	1.15	1.03	0.90
>300 to 400 feet	1.67	1.17	1.23	
>400 to 500 feet	1.11	1.13	1.07	
More than 500 feet	1.00	1.00	1.00	

For example, the display shown here is part of the output from an examination of burglary in a city. It shows that once a home has been targeted, the chance of the same location being targeted again on the same day in 641 percent (the 7.41 value) greater than if there were no discernible pattern in offender behavior. Because of the bright red color of the value, this result is statistically significant at the greatest level of significance selected. Repeat victimization cells – as opposed to near repeat victimization cells – are shown with a light blue background.

Importantly, there is also a near repeat pattern that exists on the same day as a burglary. There is increased risk to nearby homes between (more than) 100 feet and 200 feet away. These values are also statistically significant but at the p<0.05 level and not the highest statistical significant level selected (due to the dull red color). Other values are in grey indicating they are either not statistically significant and/or their odds ratios are less than 1.2.

If you see **NaN**, this signifies that the program was unable to make a calculation for that cell. This is usually because there are no observations in that spatio-temporal cell. You can confirm this by examining the observed frequency table in the output. By the way, NaN means Not a Number.

Statistical significance table

The second table shows the statistical significance of the finding within each cell. The limit for the statistical significance value depends on the number of Monte Carlo iterations that you selected. Greater number of iterations improves the statistical level achievable. 999 iterations may take a longer time to calculate but the maximum statistical significance that can result is a p value equal to 0.001. Running only 99 iterations means that the chance of potential error rises to one in one hundred (1:100).

Statistical significance values in the table that are at the best possible result for the number of iterations you chose are shown in a bright red color. Values that do not reach this level, but are at least statistically significant to 0.05 (a commonly accepted threshold in social sciences) are shown in a dull red color. Statistically significant values are shown in bold font.

Parameters used

This section shows the parameter choices that you made.

Additional analysis details

Source and output files

This section shows; the source file for the data; the number of data records that were read and used correctly; the number of data rows in the source file that were without useful data; and output file details.

The summary (htm format) output file takes the source data filename and adds _NRsummary_ along with the date of analysis, and the hour, minute and second that the output file was created. This prevents accidental overwriting of the output by any subsequent analysis.

The other output file is the verbose output file. This is a csv format file containing details of the Monte Carlo (MC) iterations (simulations). It has an identical filename however has a .csv file extension. As the summary file show, if you have Excel loaded on your computer (or any program that is the default program for handling csv files) you should be able to open the simulation file by clicking where indicated.

Observed frequency table

This table shows the frequency of event-pairs in each space-time cell group based on the parameters you selected. If you see cells with zero counts, then the program will not be able to estimate odds ratios and

statistical significance for these cells in other areas of the program. In this case, the Observed over mean expected frequencies table will show NaN in those cells. So will the Statistical significance table. If the program has just one value, or under some other circumstances, you might see inf. This means infinity, and is another way of saying that the program could not make a calculation for that cell. You can generally ignore NaN and inf cells.

About near repeats

This section provides an internet link to a web page with further details of near repeat research. You must have an internet connection enabled to access this link. If you cannot access the link, you can return to the summary output file by clicking 'Back' on your internet browser.

VERBOSE FILE DETAILS

The verbose output file is in csv format. The file shows the source data filename, the number of data values, and when the file was created.

There follow two tables; the observed over mean frequencies, and the table of actual observed frequency distribution. The latter may appear to have a large number of total values; however it should be remembered that the program measures the space-time distance from each point to every other, so there are a large number of space-time pairs to place in categories. For example, if there are 200 points in a dataset (n=200) the total number of pairs is n(n-1)/2. So for 200 pairs, this results in 200(199)/2 = 39,800/2 = 19,900.

These tables both show row and column numbers to help interpret the lengthy display that follows.

For each row and column in the space-time matrix, every Monte Carlo simulation result is shown. These are shown next to the first column label of Expected. If there are more than about 50, then the values will spill over onto numerous lines of the spreadsheet.

The table then shows the observed value (which should match the observed value in the second table of the output); the minimum of the expected values; the maximum of the expected values; the mean of the expected values; and the observed over mean expected value (which should match with the first table of the output).

This file is predominantly provided to allow researchers to calculate their own statistical significance values if they wish, or to better understand the overall program process. It is not expected that it is of value to the majority of users, who are directed to the output in the summary file. See Summary file details.

OTHER FUNCTIONS

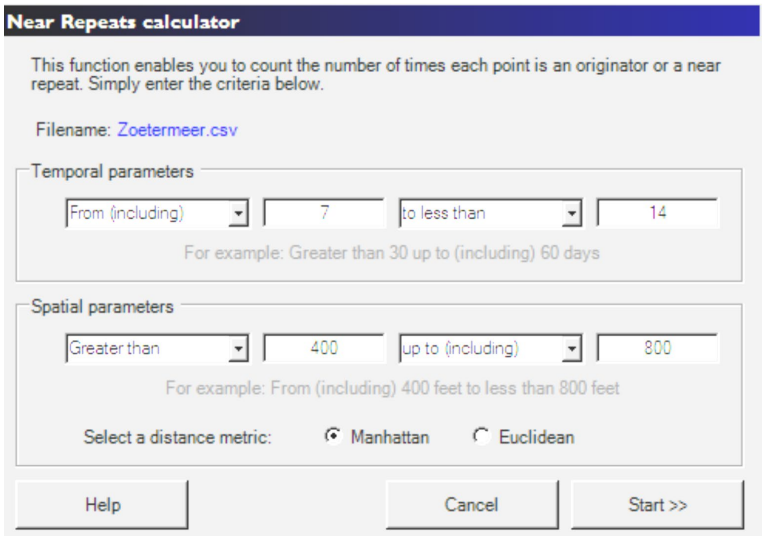
When you click the button 'Other functions' on the main dialog, additional functionality is provided. This feature is only available once you have selected a data file.

NEAR REPEAT ORIGINATOR AND REPEAT COUNTER

This function enables you to determine how many times a crime event was either the original event in a near-repeat pair, or the subsequent (near repeat) event.

The source file for the process is the data file that is already loaded into the program. You cannot access this function without selecting a data file. In the dialog box, you choose the appropriate criteria for your search. In other words, if you wanted to identify events that contributed to near repeat patterns from 7 to less than 14 days apart, and more than 400 feet up to and including 800 feet, you would enter criteria as shown in the following dialog box.

The drop-down boxes allow you to choose whether or not to include the number that follows it. For example, if you wanted to include 7 as the lower number for the days searched, you would select "From (including)", but if you wanted all near repeat pairs where the temporal distance was more than 7, you would choose "Greater than".



Some examples, with notation:

1) If x is a search parameter (it could be either days or distance), and you want to search for 7 <= x < 14 you would select and enter: **From (including) 7 to less than 14**. In interval notation, this is shown as [7, 14) because a square bracket [means include the number next to it, while a parenthesis means exclude the adjoining number.
2) For parameter 100 > x <= 800, select and enter: **Greater than 100 up to (including) 800**. It would be shown in interval notation as (100, 800].
3) For parameter 50 >= x <= 70, select and enter: **From (including) 50 up to (including) 70**. It would be shown in interval notation as [50, 70].[1]

When you click Start >> the program runs. The program run time should usually be instant with a reasonably sized data set and a modern computer. The output file is calculated from the filename of the input file. It takes the source data filename and adds _NROriginatorCount_ along with the date of analysis, and the hour, minute

[1] Wikipedia has more details on interval notation at http://en.wikipedia.org/wiki/Interval_notation

and second that the output file was created. This prevents accidental overwriting of the output by any subsequent analysis.

The output file is a comma-separated values files (*.csv) that can be opened with Microsoft Excel. When the program finishes, it shows the search criteria using interval notation, and asks if you would like to open the output file. If you do not have a program set up to automatically handle csv files it will not open, but can be opened manually with a text editor. The file will be located in the same folder as the source data file.

Example output

The output file will appear similar to the example below. For each crime event the output indicates the x-coordinate, y-coordinate, and day of the crime. Then it shows the number of times that the crime event was the originator (first) event in a near-repeat pair. It then shows how many times it was the subsequent (second) event in a near repeat pair.

X	Y	Date	Originator	NearRpt
392146	420226	4/18/2003	13	14
392144	420237	2/4/2002	5	2
392234	420570	6/24/2002	2	2
395139	422747	10/24/2003	14	24
393520	423799	1/19/2002	9	10

In this example, the first crime event at coordinates (392146, 420226) was the originator of a near repeat pair to 13 subsequent crimes, and was itself the near repeat to 14 other events that preceded it.

Important note: Please note that the definition of a near repeat is determined by the criteria entered by the user, and not from the criteria or program output from the main part of the program. This is done to give the user greater flexibility.

Why might this be useful?

While other crime analysis programs are able to tell you where clusters of crime events take place, part of the predictive power of the near repeat hypothesis is that knowing the space and time of near repeats helps to put spatial and temporal boundaries on proactive crime prevention measures. In other words, once there is a shooting in a city knowing that there is increased risk of another shooting over the next two weeks and within 500 feet helps to target prevention measures. This function adds to the value of that information by helping identify clusters (if any exist) or hotspots for originator events. It may be that in some parts of a jurisdiction the risk of near repeats is much greater than in other areas, irrespective of the actual distribution of the general crime pattern. This function can help to identify the high-count originator events. Mapping the originator event locations may also be helpful.