Identification and cleansing of scatter in GPS-surveys in urban environments

Jensen, Anders Sorgenfri; Bro, Peter; Harder, Henrik

Publication date:
2009

Document Version
Publisher's PDF, also known as Version of record

Identification and cleansing of scatter in GPS-surveys in urban environments

Anders Sorgenfri Jensen, Peter Bro, Henrik Harder

Identification and cleansing of scatter in GPS-surveys in urban environments
Jensen, A. S., Bro, P. & Harder, H.

Aalborg University
Department of Architecture and Design
ISSN: 1603-6204
Series: 27

This paper is a documentation of the workflow of the labour performed prior to a GIS-based analysis. The paper explains which methods are used to prepare the raw data for further analysis, and describes why and how the processes are being executed.

Datasets, data types and terminology

The surveys performed by Diverse Urban Spaces operate with several data sets and data types. Furthermore two different types of error are associated to the GPS loggings. This section seeks to give an overview of the complexity of datasets and errors through explanation of the terminology used in this paper.

Raw data

Raw data is the entire collection of all points recorded during the time when the GPS receivers has been in the custody of a respondent and has been turned on. Raw data consists of several parameters including geo references, timestamp, attributes associated with the precision of the recorded point and technical values related to the either the GPS-receiver or the network used for data transmission.

Survey data

Is the data collected through the questionnaire which the respondents had to fill out each day during the survey. This dataset contains personal information about the respondent along with information about purpose, cost etc of the trip during which the point was recorded.

Scatter points

As stated above, the data set is affected by two types of inaccurately registered points. These two types will henceforth be denoted as Scatter type 1 and Scatter type 2.

Scatter type 1

This error happens whenever the location of the GPS-receiver is static. Since this often means that the respondent (and thus the receiver) is located inside a building, the receiver has difficulty recording its true position. This results in a doodle-like movement pattern being recorded in the dataset.

0-1 Which points belong to the actual trip? Which points belong to the activity at the location? Last ones are defined as Scatter type 1 points.

Scatter type 2

This error occurs randomly for no apparent reason. During a Scatter type 2 situation, the points recorded lose their track for a brief moment lasting a few seconds up to several minutes. During this instance, the points are recorded in a linear pattern, occasionally ascending speed levels of several hundred km/h.

2 An example of a trip affected by Scatter type 2. During the point recording process, the receiver lost track and registered points in a linear pattern in the upper right corner.

Processes

To make the data gathered through the GPS receiver and the questionnaire suitable for GIS-based analyses, the datasets have undergone several processes. The following text is an explanation of these steps.

1. Cleaning the raw dataset of scatter type 1

The first step is to split the dataset by defining whether a given point is associated with a trip or associated with a stay. The reason for this process is to make the dataset more flexible for GIS analyses based on trips. Since points affected by scatter type 1 in almost every case has been registered during the stay at the destination place, the opposite fact holds true; that a point not being affected by Scatter type 1 is a trip point. The method for identifying Scatter type 1 points is explained later in this paper. The dataset created through this process is defined as Processed dataset - Stage 1 in Figure 1.

2. Joining the raw dataset with the survey data

Second step is to join the geodata registered by the GPS receiver with the point and purpose of the trip as explained by the respondent through the questionnaire. The join is based around the timestamp registered by the GPS receiver for each point, and the time interval of each trip provided by the respondent. Every point from the raw dataset with a timestamp within the interval of a given trip from the survey dataset is joined with the questionnaire data tied to the trip. The result is a new, bigger dataset. In Figure 1, this dataset is defined as Processed dataset - Stage 2.

3. Interpolation

Third step is to expand the dataset to make it suited for time based GIS analyses. Since the GPS receiver only records a point circa every sixth second, it is possible (and necessary) to construct additional, intermediate points with the same features as the original points through interpolation. In Figure 1, the resulting dataset is defined as Processed dataset - stage 3.

4. Cleaning the dataset of Scatter type 2

The final step is to clean the dataset of Scatter type 2 points. This is being done through a combined automated and manual process. However this is not explained within this paper.

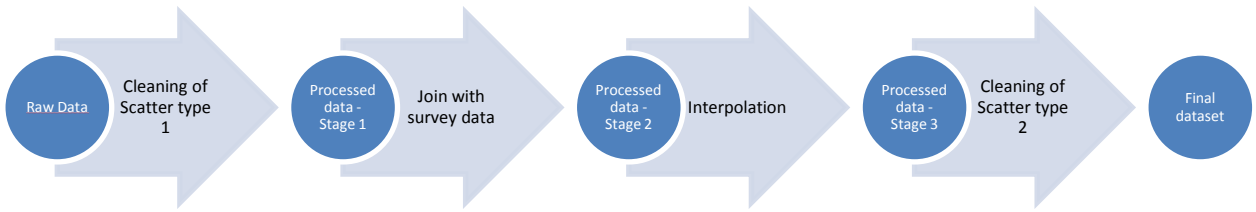


Figure 1: Workflow

Cleaning the raw dataset of Scatter type 1

The following text explains how the dataset is being cleaned of Scatter type 1. Note that the method is a test which is still being tuned and therefore contains flaws. The chapter explains how the test proceeds, along with the results and conclusions. The purpose of the test is to develop an algorithm that through the technical properties of a given point along a trip makes it possible to automate identification of whether the point belongs to the initial/destination location or belongs to the actual trip between the two positions. Such methods will reduce the amount of manual labour required to clean up a given dataset of points. The ultimate goal is to eliminate the manual labour.

Procedure

The test utilizes the values of speed, direction and distance recorded during natural movement along roads, streets etc opposed to the same values recorded during a static location of the GPS-receiver. Since roads and streets tend to be somewhat linear with minor curving and eventually turn, a given person is less likely to make sudden and repeatedly shifts in direction during his trip. Furthermore, when a respondent is on the move, the GPS-receiver will record a noticeable speed value for each point as well as a decent distance travelled. This movement pattern defined by speed, direction and distance travelled changes when a person reaches the destination place. This is due to a number of reasons:

1. The GPS-receiver has difficulties locating the true position during non-movement (especially if the receiver lies within a building, which happens commonly when a respondent reaches his destination),

and therefore records a doodle-like “trip” with random change in direction. As mentioned earlier, these points are defined as Scatter type 1.

2. When the GPS-receiver is recording Scatter type 1 points, the values of speed and distance recorded is generally lower than during the time when the receiver is recording the actual trip.

Both properties of the destination points - the sudden change in direction and the speed/distance values being near-zero - is particularly noticeable when calculating an average (or sum) of the speed values of a series of subsequent points. The procedure of the test is therefore three queries which return three new values for each point based on the sum of direction shifts, the sum of distance travelled and the sum of the recorded speed values for a set number of points post and prior to the given point.

In the end, the values of the three queries are being used to form a final query which marks a given point as either a trip or a destination point depending on its sum-values.

Calculations

Dirsum

The query is a sequence of simple calculations. This example shows how the operations proceed for determining the Dirsum value for the point with FID = 6 with the number of points post and prior to FID6 being 5.

(1) First step is to calculate the difference between the directions of two subsequent points.

$$
\begin{aligned}
Diff_{n-5n-4} &= DIR_{n-5} - DIR_{n-4} \\
Diff_{n-4n-3} &= DIR_{n-4} - DIR_{n-3} \\
&\vdots \\
Diff_{n-1n} &= DIR_{n-1} - DIR_{n} \\
Diff_{nn+1} &= DIR_{n} - DIR_{n+1} \\
&\vdots \\
Diff_{n+4n+5} &= DIR_{n+4} - DIR_{n+5}
\end{aligned}
$$

If an input table consisting of the following points:

FID	DIR
1	45
2	345
3	102
4	21
5	329
6	102
7	89
8	341
9	42
10	69
11	183

The output, sub-query table will be:

DiffID	Diff
12	-300
23	243
34	81
45	-308
56	227
67	13
78	-252
89	299
910	-27
1011	-114

(2) Next step is to convert negative values into numeric values.

$$IF\ Diff < 0$$
$$NumDiff = |Diff|$$

(3) The output table is:

DiffID	NumDiff
12	300
23	243
34	81
45	308
56	227
67	13
78	252
89	299
910	27
1011	114

(4) Now seeing as the difference between some of the points is greater than 180 degrees, the true difference between the directions of the two points is calculated

$$IF\ NumDiff > 180$$
$$TrueDiff = 360 - NumDiff$$

DiffID	TrueDiff
12	60
23	117
34	81
45	52
56	133
67	13
78	108

89	61
910	27
1011	114

(5) Finally, the sum of the differences between the implicated points is calculated

$$Dirsum = \sum_{n-5}^{n+5} TrueDiff_n$$

FID	DIR	Dirsum
1	45	
2	345	
3	102	
4	21	
5	329	
6	102	766
7	89	
8	341	
9	42	
10	69	
11	183	

The calculated value in the example is relatively high, which isn't much of a surprise if we look at the DIR values of the input table, meaning that the point with FID = 6 has potential to be a point recorded during the time spent at the destination. However, it could also be a pedestrian or bicycle driver manoeuvring through urban spaces. Therefore, we need the sum of distance and speed measured for the same points before we can decide, whether the point is a destination point.

Distsum

Like Dirsum, the query for calculating a sum of distance travelled is a series of simple calculations. Again, the sum-value for the point with FID = 6 will be calculated as an example. Note that the points in the tables are not related to the points in the example for Dirsum.

(1) First step is to calculate the distance between two subsequent points. This is done by applying the Pythagorean theorem to the projected coordinates.

$$\begin{aligned} Dist_{n-5n-4} &= \sqrt{(X_{n-5} - X_{n-4})^2 + (Y_{n-5} - Y_{n-4})^2} \\ Dist_{n-4n-3} &= \sqrt{(X_{n-4} - X_{n-3})^2 + (Y_{n-4} - Y_{n-3})^2} \\ &\vdots \\ Dist_{n-1n} &= \sqrt{(X_{n-1} - X_n)^2 + (Y_{n-1} - Y_n)^2} \\ Dist_{nn+1} &= \sqrt{(X_n - X_{n+1})^2 + (Y_n - Y_{n+1})^2} \\ &\vdots \\ Dist_{n+4n+5} &= \sqrt{(X_{n+4} - X_{n+5})^2 + (Y_{n+4} - Y_{n+5})^2} \end{aligned}$$

A table with the following points:

FID	X	Y
1	12	52
2	17	13
3	52	24
4	42	52
5	32	24
6	19	45
7	39	12
8	2	22
9	61	49
10	27	35
11	65	11

Results in the following subquery-table:

DistID	Dist
12	39.319
23	36.688
34	29.732
45	29.732
56	24.698
67	38.588
78	38.328
89	64.885
910	36.770
1011	44.944

(2) Last step is to sum the values of the sub-query table to achieve the Distsum value.

$$Distsum = \sum_{n-5}^{n+5} Dist_n$$

FID	X	Y	Distsum
1	12	52	
2	17	13	
3	52	24	
4	42	52	
5	32	24	
6	19	45	383.684
7	39	12	
8	2	22	
9	61	49	
10	27	35	

11	65	11	

The calculated value is relatively high. Since high values of distance travelled is a characteristic of motion, the point with FID = 6 is most likely to be a trip point.

Speedsum

The Speedsum query is the last and simplest of the calculations, involving only the sum of the value SPEED through a subquery of points prior to and post the current point. Again, the example will resolve around determining the sum-value of the point with FID = 6.

$$Speedsum = \sum_{n-5}^{n+5} SPEED_n$$

For a table consisting of the following points:

FID	SPEED
1	14
2	12
3	10
4	15
5	16
6	10
7	4
8	0
9	1
10	0
11	0

The Speedsum query yields:

FID	SPEED	Speedsum
1	14	
2	12	
3	10	
4	15	
5	16	
6	10	118
7	4	
8	0	
9	1	
10	0	
11	0	

The calculated value in the example is intriguing. The number is too large to be classified as Scatter type 1, but the SPEED values of points 7,8,…,11 low, indicating that FID = 6 might be one of the last actual trip points on the tour. Wether FID = 6 will be classified as scatter point or trip points depends on the other calculated sum-values.

Results of the queries

The three queries were performed on a dataset containing 447.570 records. The queries involved 1, 5, 10 and 20 points back and forth respectively for each attribute (speed, direction and distance travelled). From this dataset, approximately 10.000 points of either type (trip vs. Initial/destination) were carefully, manually selected. Graphical plots and statistical values can be found in appendix A.
The key value of interest is the approximate X-value of intersection or the approximate max value of the scatter points. This value marks the threshold between most occurrences of scatter points and most occurrences of trip points. That is, for a dataset containing the calculated Speedsum5, a selection of points where Speedsum5 > 23 will yield more trip points than scatter points and vice verse.
Note that this value of intersection/max value of occurrences has been estimated manually based on the visuals of the graph. It is therefore highly unlikely to be the optimal value for a selection that yields as many scatter points and as few trip points as possible.

The final Point type query

The final query which marks every point as either a trip point or a scatter point is constructed on the threshold values mentioned before. The 3*4 calculated values (Speedsum, Dirsum and Distsum for 1, 5, 10 and 20 points) can be combined in multiple ways. The current final Point type query uses the following sums and threshold values:

$$IF\ Distsum10 < 80\ AND\ Speedsum10 < 45\ AND\ Dirsum5 \geq 125:$$

$$Pt\_type = 1$$

$$ELSE:$$

$$Pt_{type} = 0$$

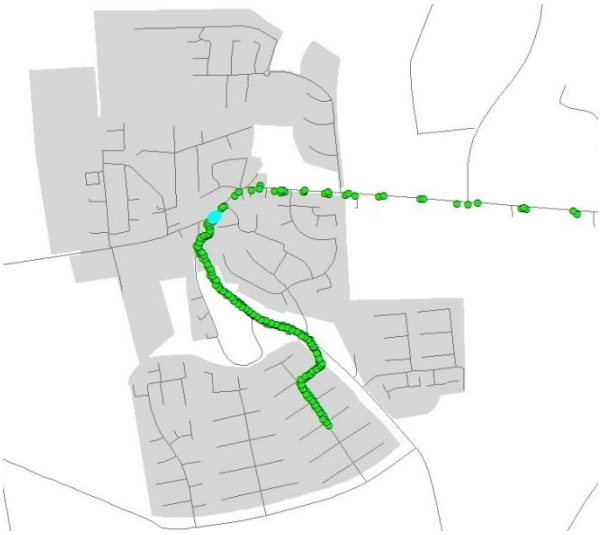
With 1 being scatter points and 0 being trip points.

Assessment of the

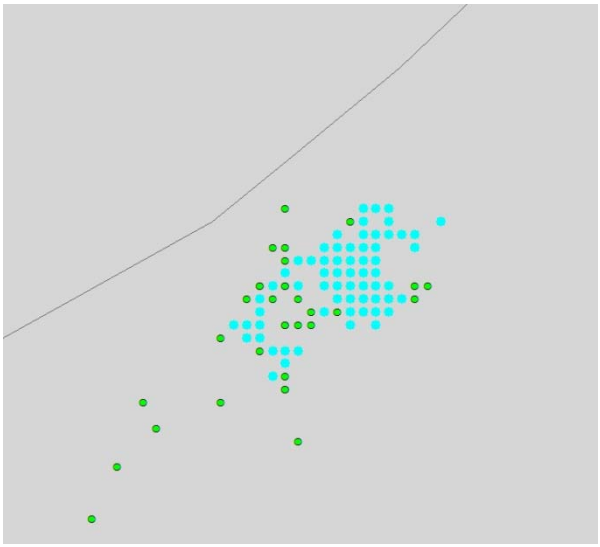
When performed on the datasets containing scatter points and trip points, which were used to identify the properties of either type, the final query yields the following.

	Trip points dataset	Scatter type 1 points dataset
Number of points	8.658	9.346
Number of points marked as scatter by the final query	197	6.679
Percentage marked	2,275 %	71,464 %

Upon further inspection of the 2,275 % of the trip points, which were wrongfully marked as scatter points, most of them turn out to actually be scatter points. This is a result of a human error made during the earlier mentioned manually selection of the trip points.



3 The dataset of trip points appears to contain scatter points by accident



4 A zoom of the suspecious points removes all doubt

If the obvious scatter points are removed from the dataset and the final query is performed again, the result is as following:

	Trip points dataset	Scatter type 1 points dataset
Number of points	8.658	9.346
Number of points marked as scatter by the final query	15	6.679
Percentage marked	0,0017 %	71,464 %

Conclusion

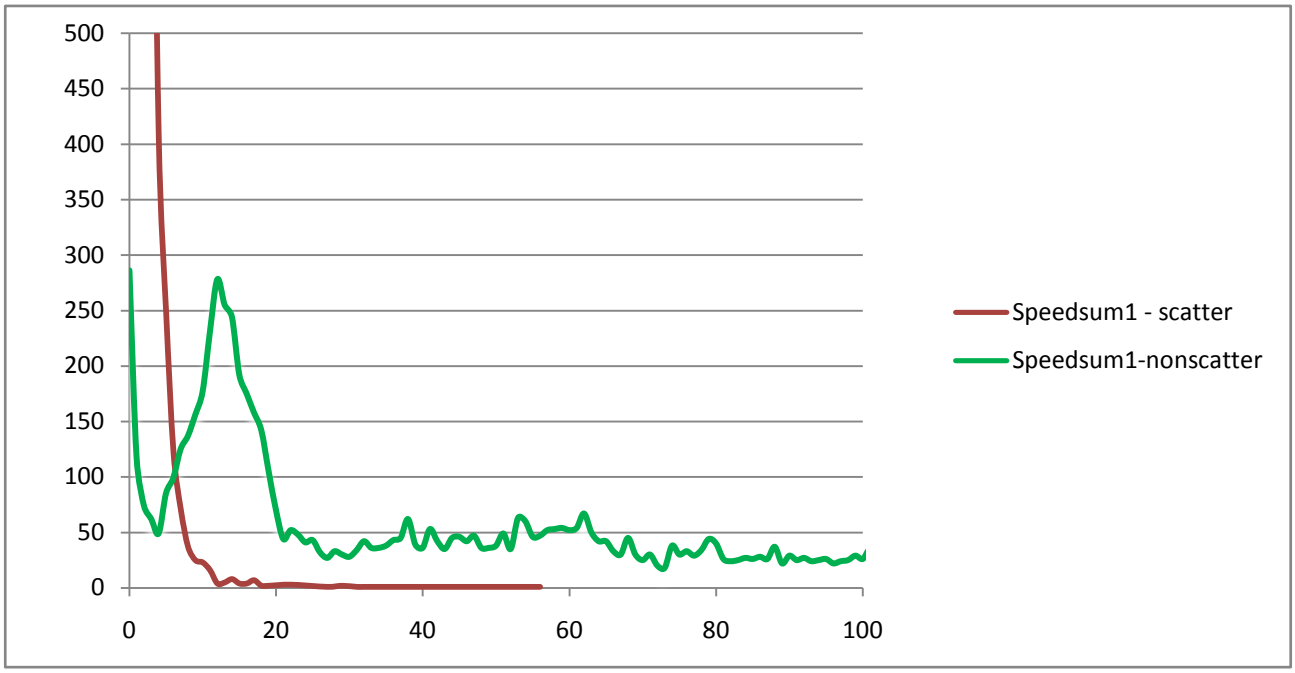
The advantage of the final query is that it manages to mark a good share of the scatter points as scatter points, whilst only a minimal amount of trip points are being incorrectly marked as scatter points. The downside of the final query is that a certain percentage of the scatter points (18.536 %) is marked as trip points, which can be quite a big amount of points depending on the size of the dataset. However, the manual labour required to phase out the remaining scatter points, in order to perform GIS-based analysis's on the trip points only should have been vastly reduced based on this query.
As mentioned earlier, the threshold values, which are the key to identify the properties of the scatter and trip points, have been estimated manually based on graph visuals. If these values can be measured more precisely, it is likely that the amount of correctly marked scatter points can be enhanced further while still keeping a minimum of erroneous marked trip points.

Appendix A – Query results

This appendix contains the results of the three sum queries (Speedsum, Distsum, Dirsum). Each query has run 4 times with the algorithm using 1, 5, 10 and 20 points back and forth respectively.

Speedsum

Speedsum1

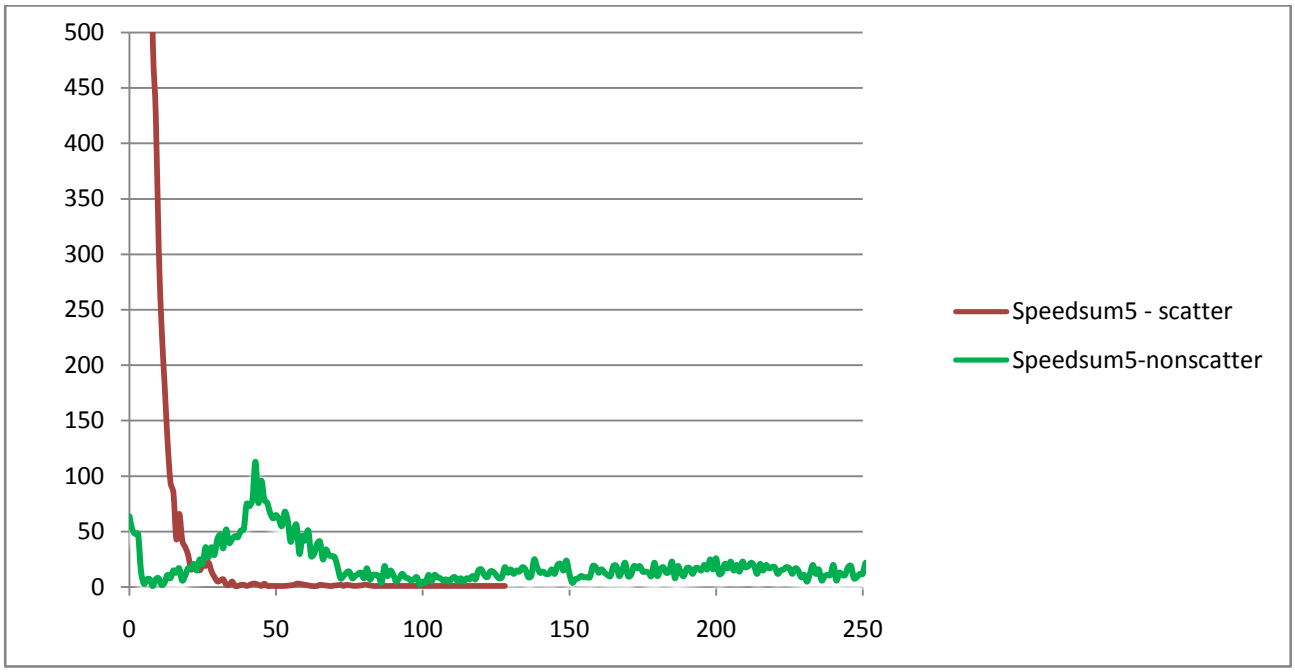


Statistical values of Scatter points	
Min. of Speedsum1	0
Max. of Speedsum1	56
Mean of Speedsum1	1,56
Stddev of Speedsum1	2,29

Statistical values of Non-scatter points	
Min. of Speedsum1	0
Max. of Speedsum1	427
Mean of Speedsum1	71,62
Stddev of Speedsum1	72,71

Correlations between scatter and non-scatter points	
Number of scatter points	9346
Number of non-scatter points	8658
Approximate X-value of intersection	6,2
Scatter points with value lower than point of intersection	9125
Scatter points with value higher than or equal to point of intersection	221
Non-scatter points with value lower than point of intersection	772
Non-scatter points with value higher than or equal to point of intersection	7886

Speedsum5

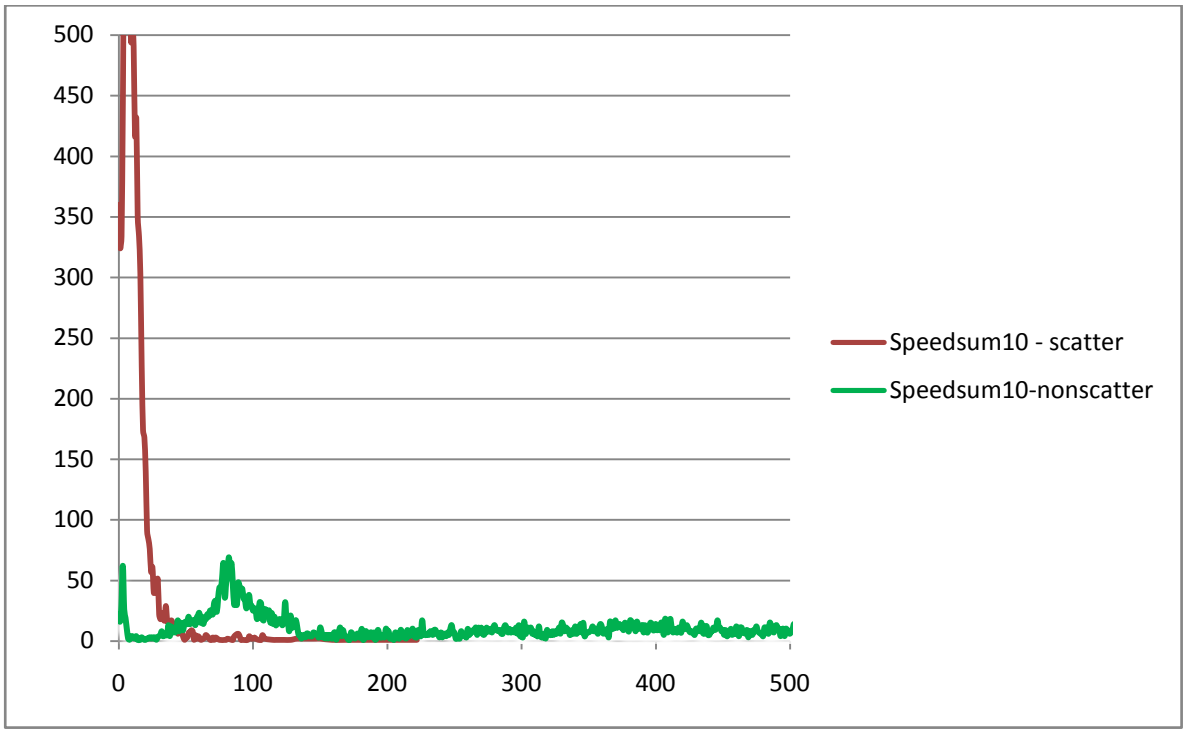


Statistical values of Scatter points	
Min. of Speedsum5	0
Max. of Speedsum5	128
Mean of Speedsum5	5,88
Stddev of Speedsum5	6,68

Statistical values of Non-scatter points	
Min. of Speedsum5	0
Max. of Speedsum5	1410
Mean of Speedsum5	256,69
Stddev of Speedsum5	239,24

Correlations between scatter and non-scatter points	
Number of scatter points	9346
Number of non-scatter points	8658
Approximate X-value of intersection	23
Scatter points with value lower than point of intersection	9145
Scatter points with value higher than or equal to point of intersection	201
Non-scatter points with value lower than point of intersection	400
Non-scatter points with value higher than or equal to point of intersection	8258

Speedsum10

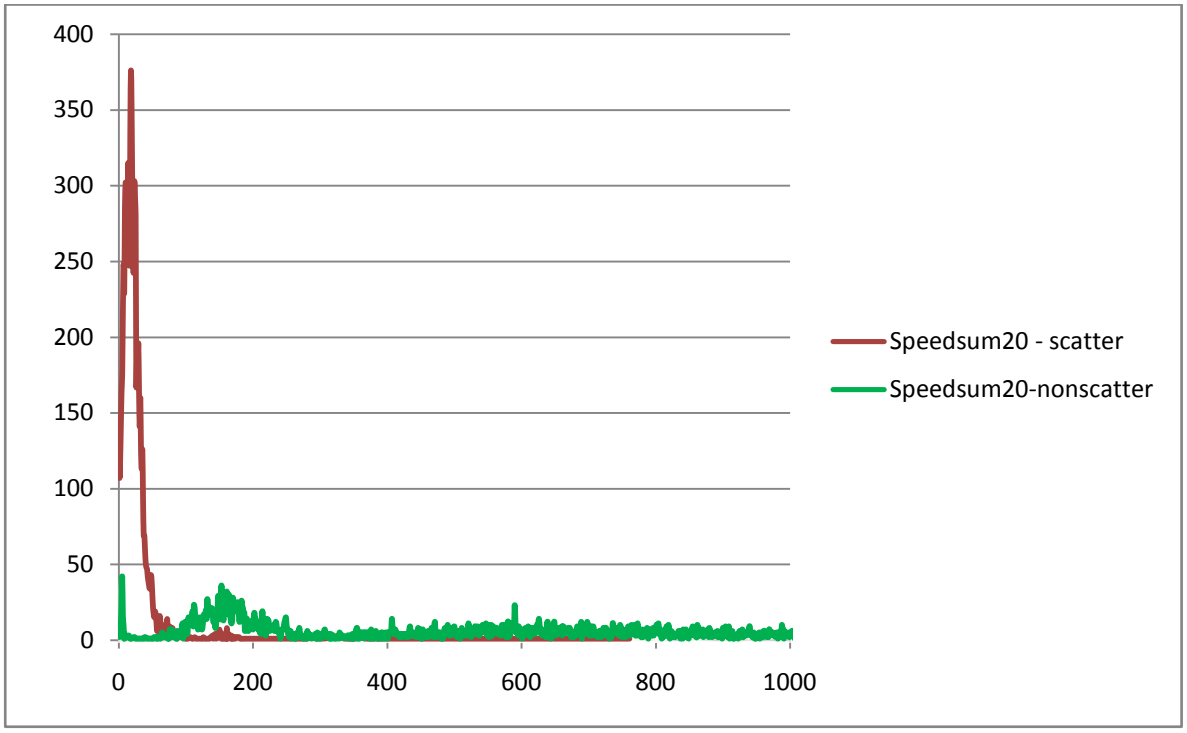


Statistical values of Scatter points	
Min. of Speedsum10	0
Max. of Speedsum10	222
Mean of Speedsum10	11,52
Stddev of Speedsum10	12,24

Statistical values of Non-scatter points	
Min. of Speedsum10	0
Max. of Speedsum10	2541
Mean of Speedsum10	473,03
Stddev of Speedsum10	427,10

Correlations between scatter and non-scatter points	
Number of scatter points	9346
Number of non-scatter points	8658
Approximate X-value of intersection	45
Scatter points with value lower than point of intersection	9176
Scatter points with value higher than or equal to point of intersection	170
Non-scatter points with value lower than point of intersection	340
Non-scatter points with value higher than or equal to point of intersection	8318

Speedsum20



Statistical values of Scatter points	
Min. of Speedsum20	0
Max. of Speedsum20	761
Mean of Speedsum20	23,82
Stddev of Speedsum20	30,13

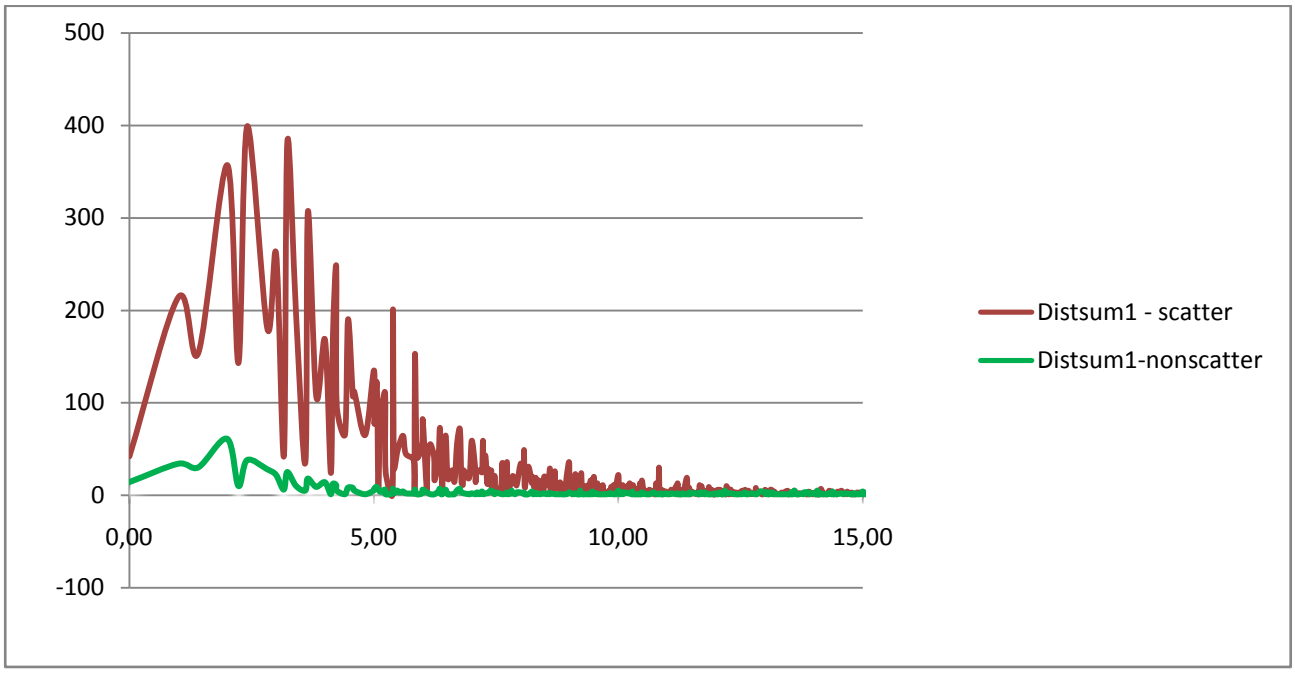
Statistical values of Non-scatter points	
Min. of Speedsum20	1
Max. of Speedsum20	4513
Mean of Speedsum20	861,46
Stddev of Speedsum20	758,88

Correlations between scatter and non-scatter points

Number of scatter points	9346
Number of non-scatter points	8658
Approximate X-value of intersection	85
Scatter points with value lower than point of intersection	9144
Scatter points with value higher than or equal to point of intersection	202
Non-scatter points with value lower than point of intersection	218
Non-scatter points with value higher than or equal to point of intersection	8440

Distsum

Distsum1



Statistical values of Scatter points	
Min. of Distsum1	0
Max. of Distsum1	3362,74
Mean of Distsum1	7,74
Stddev of Distsum1	55,76

Statistical values of Non-scatter points	
Min. of Distsum1	0
Max. of Distsum1	25510,46
Mean of Distsum1	133,69
Stddev of Distsum1	414,42

Correlations between scatter and non-scatter points	
Number of scatter points	9346
Number of non-scatter points	8658
Approximate maximum X-value of scatter	13
Scatter points with value lower maximum X-value	8611
Scatter points with value higher than or equal to maximum X-value	735
Non-scatter points with value lower than maximum X-value	791
Non-scatter points with value higher than or equal to maximum X-value	7867

Distsum5

Statistical values of Scatter points	
Min. of Distsum5	7
Max. of Distsum5	3456,74
Mean of Distsum5	40,90
Stddev of Distsum5	146,06

Statistical values of Non-scatter points	
Min. of Distsum5	5,41
Max. of Distsum5	25878,88
Mean of Distsum5	661,39
Stddev of Distsum5	1043,54

Correlations between scatter and non-scatter points	
Number of scatter points	9346

Number of non-scatter points	8658
Approximate maximum X-value of scatter	80
Scatter points with value lower maximum X-value	9012
Scatter points with value higher than or equal to maximum X-value	334
Non-scatter points with value lower than maximum X-value	439
Non-scatter points with value higher than or equal to maximum X-value	8219

Distsum10

Statistical values of Scatter points	
Min. of Distsum10	19,07
Max. of Distsum10	3618,59
Mean of Distsum10	81,10
Stddev of Distsum10	196,17

Statistical values of Non-scatter points	
Min. of Distsum10	17,48
Max. of Distsum10	50496,76
Mean of Distsum10	1338,15
Stddev of Distsum10	1958,12

Correlations between scatter and non-scatter points	
Number of scatter points	9346
Number of non-scatter points	8658

Approximate maximum X-value of scatter	200
Scatter points with value lower maximum X-value	9160
Scatter points with value higher than or equal to maximum X-value	186
Non-scatter points with value lower than maximum X-value	708
Non-scatter points with value higher than or equal to maximum X-value	7950

Distsum20

Statistical values of Scatter points	
Min. of Distsum20	43,63
Max. of Distsum20	3760,88
Mean of Distsum20	166,36
Stddev of Distsum20	295,59

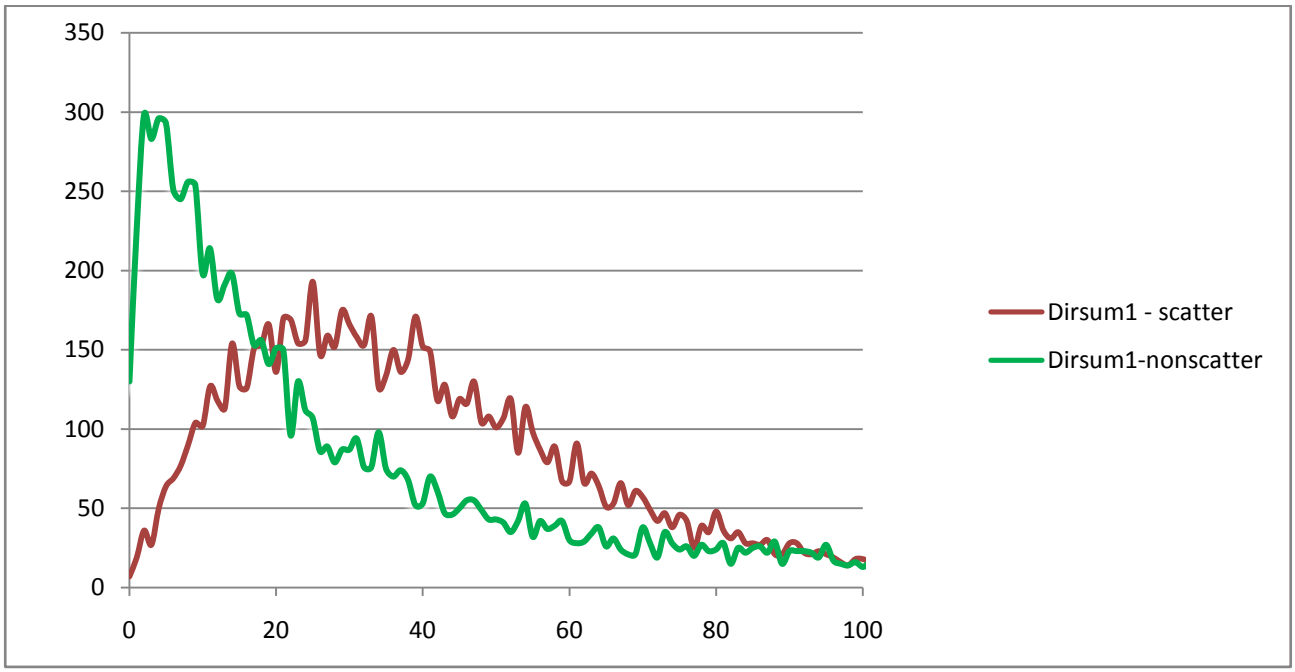
Statistical values of Non-scatter points	
Min. of Distsum20	53,04
Max. of Distsum20	94887,95
Mean of Distsum20	2801,56
Stddev of Distsum20	4231,30

Correlations between scatter and non-scatter points	
Number of scatter points	9346
Number of non-scatter points	8658
Approximate maximum X-value of scatter	380

Scatter points with value lower maximum X-value	9120
Scatter points with value higher than or equal to maximum X-value	226
Non-scatter points with value lower than maximum X-value	530
Non-scatter points with value higher than or equal to maximum X-value	8128

Dirsum

Dirsum1

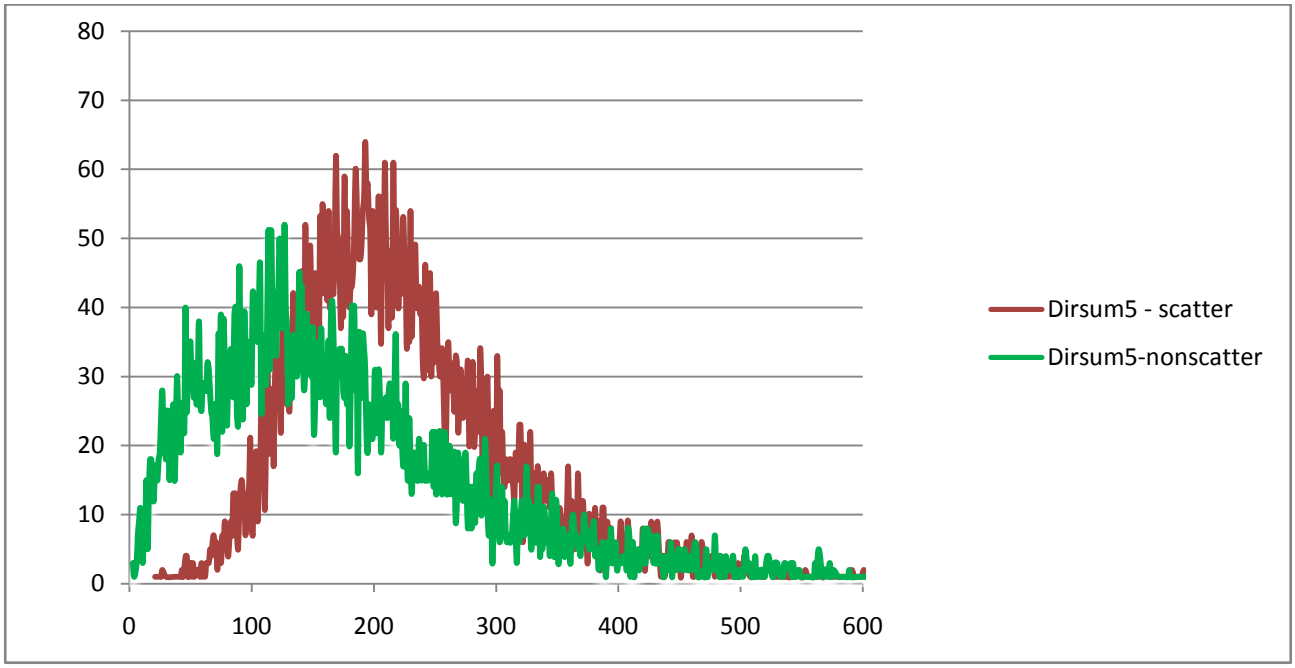


Statistical values of Scatter points	
Min. of Dirsum1	0
Max. of Dirsum1	349
Mean of Dirsum1	45,27
Stddev of Dirsum1	35,31

Statistical values of Non-scatter points	
Min. of Dirsum1	0
Max. of Dirsum1	347
Mean of Dirsum1	34,26
Stddev of Dirsum1	40,57

Correlations between scatter and non-scatter points	
Number of scatter points	9346
Number of non-scatter points	8658

Dirsum5

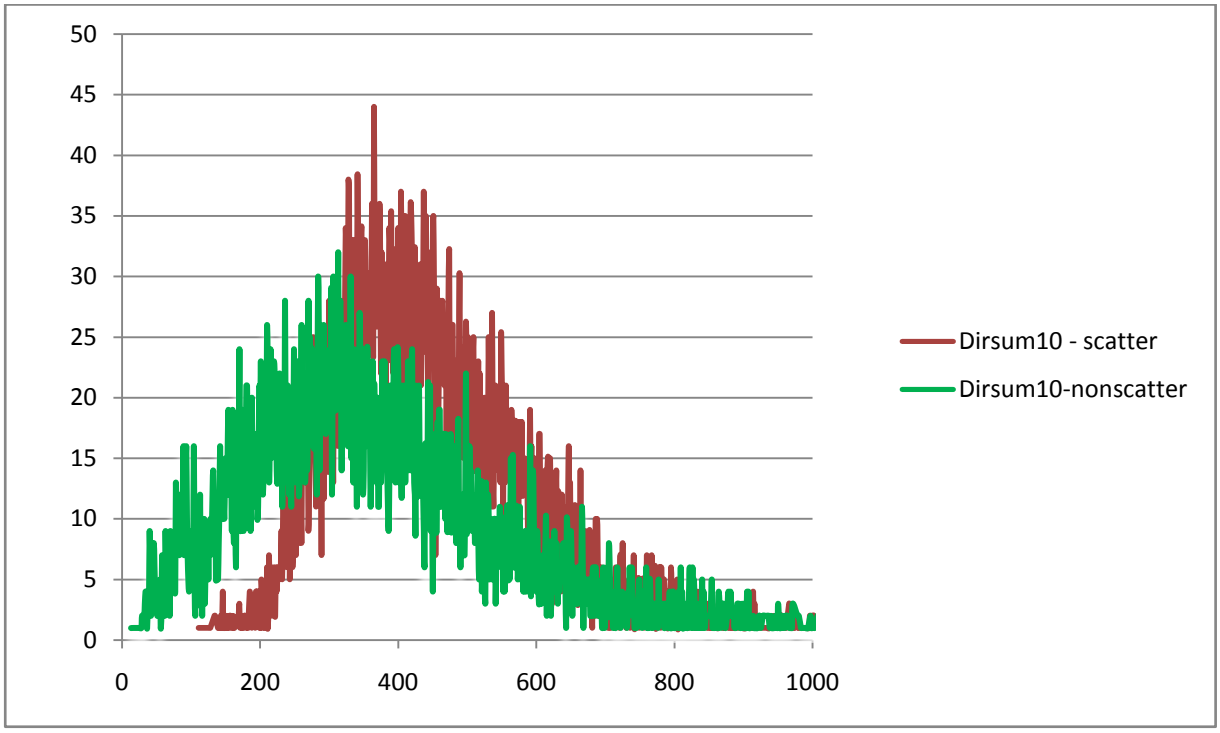


Statistical values of Scatter points	
Min. of Dirsum5	21
Max. of Dirsum5	1210
Mean of Dirsum5	226,43
Stddev of Dirsum5	95,69

Statistical values of Non-scatter points	
Min. of Dirsum5	3
Max. of Dirsum5	1234
Mean of Dirsum5	178,56
Stddev of Dirsum5	119,81

Correlations between scatter and non-scatter points	
Number of scatter points	9346
Number of non-scatter points	8658

Dirsum10

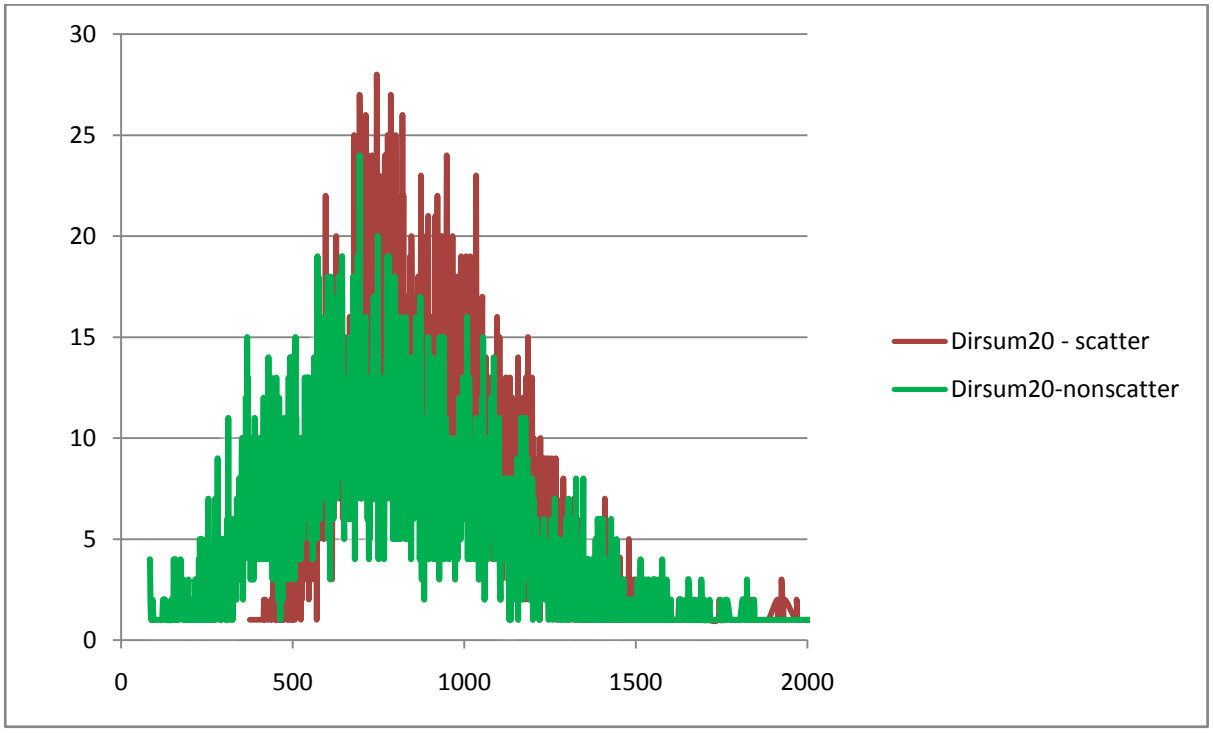


Statistical values of Scatter points	
Min. of Dirsum10	111
Max. of Dirsum10	1678
Mean of Dirsum10	452,87
Stddev of Dirsum10	152,50

Statistical values of Non-scatter points	
Min. of Dirsum10	13
Max. of Dirsum10	1738
Mean of Dirsum10	375,66
Stddev of Dirsum10	200,63

Correlations between scatter and non-scatter points	
Number of scatter points	9346
Number of non-scatter points	8658

Dirsum20



Statistical values of Scatter points	
Min. of Dirsum20	375
Max. of Dirsum20	2272
Mean of Dirsum20	907,53
Stddev of Dirsum20	249,77

Statistical values of Non-scatter points	
Min. of Dirsum20	84
Max. of Dirsum20	2658
Mean of Dirsum20	807,83
Stddev of Dirsum20	344,56

Correlations between scatter and non-scatter points	
Number of scatter points	9346
Number of non-scatter points	8658