Wioleta Błaszczak, Waldemar Kamiński

Institute of Geodesy

University of Warmia and Mazury in Olsztyn

Email: wioleta.blaszczak@uwm.edu.pl

waldemar.kaminski@uwm.edu.pl

LARGE DATA SETS AND THEIR STUDY USING DOUGLAS-PEUCKER METHOD

8th Bilateral Geodetic Meeting Poland-Italy

22-24 June 2006

Problem

Geodetic surveys may produce large amount of data which results in large data sets.

Large data sets – such data sets which may cause problems for users with loading them to suitable software or making next stage of processing more difficult *(megadates, cloud of points).* Their uploading to suitable software and subsequent processing is very time-consuming.

e.g. the survey using laser scanning,
bathymetric made by digital echosonder.

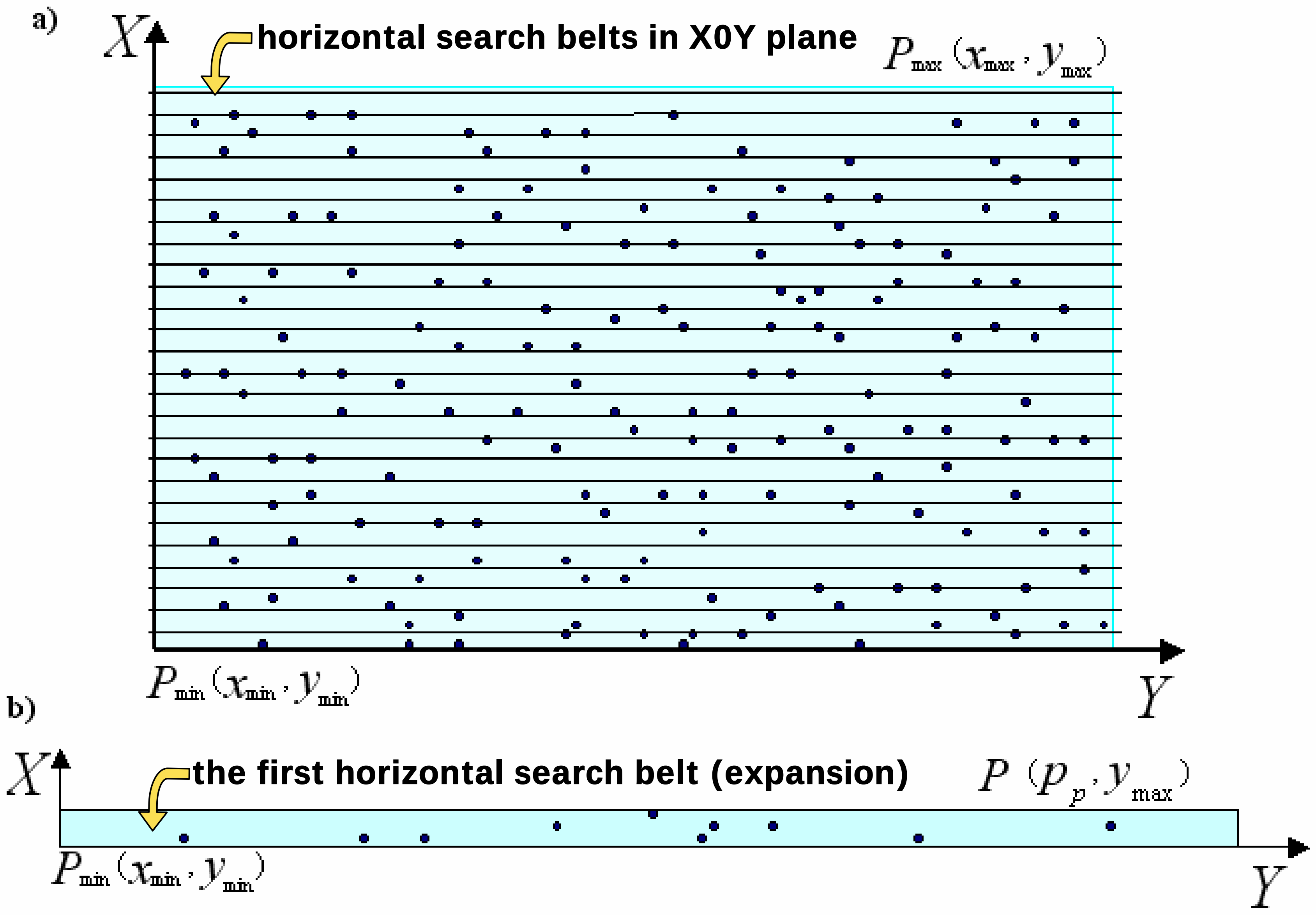
Main goal

There is a need to resample and reorganise the data sets before processing, in order to considerable reduction their amount without loosing essential information.

For reduction of data numbers **’algorithm for reduction of the number of measurement data using generalization method’** has been used.

In this paper the authors present the theoretical and empirical analyses of results obtained from data processing applying the Douglas-Peucker method (further referred to as D-P)

Algorithm for reduction of the number of measurement data



Algorithm for reduction of the number of measurement data

4. Selection of the method for elimination of points in search belts. The elimination process is carried out in Y0Z plane. The method choice depends, among others, on the purpose of processing and character of the processed object. At that stage the tolerance range must also be determined. The tolerance range is necessary in the number reduction process and depends, among others, on the precision of observations.

5. In this study the one following method was used for reduction of points:
   - Douglas-Peucker method.

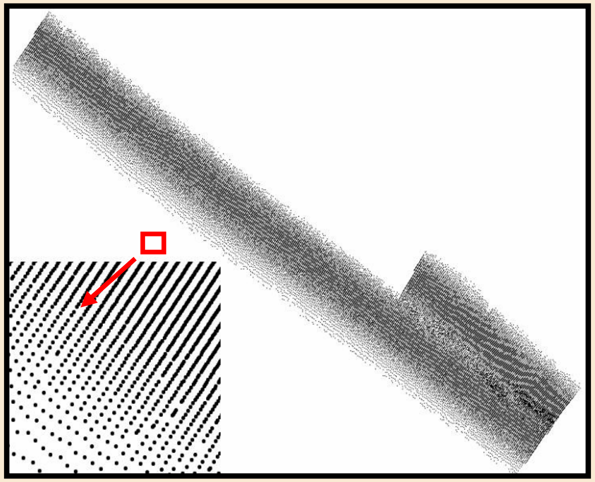
Algorithm for reduction of the number of measurement data

6. Application of the method selected for reduction of points in all search belts in Y0Z plane. The elimination process stops when all search belts have been tested.
7. Establishment of vertical search belts in X0Y plane (belts parallel of 0X axis).
8. Application of generalization method for reduction of points in all search belts on X0Z plane.
9. Completion of initial processing step and writing the resulting file to the local software.

Example of practical application

The proposed algorithm for reduction of the number of measurement data was applied for processing a fragment of the set containing results of Świnoujście-Szczecin channel bottom measurements.

The processing involved the set of 60857 pairs of coordinates X and Y as well as depths Z determined using multiple beam sounder integrated with the DGPS.



Example of practical application

The elimination process using D-P method is carried out in Y0Z plane. We have used following tolerance ranges and variants of algorithm.

TABLE 1

Tolerance range O_i [m]													
O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}	O_{11}	O_{12}	O_{13}	O_{14}
0,0005	0,001	0,002	0,003	0,004	0,006	0,008	0,012	0,017	0,024	0,034	0,048	0,059	0,068

TABLE 2

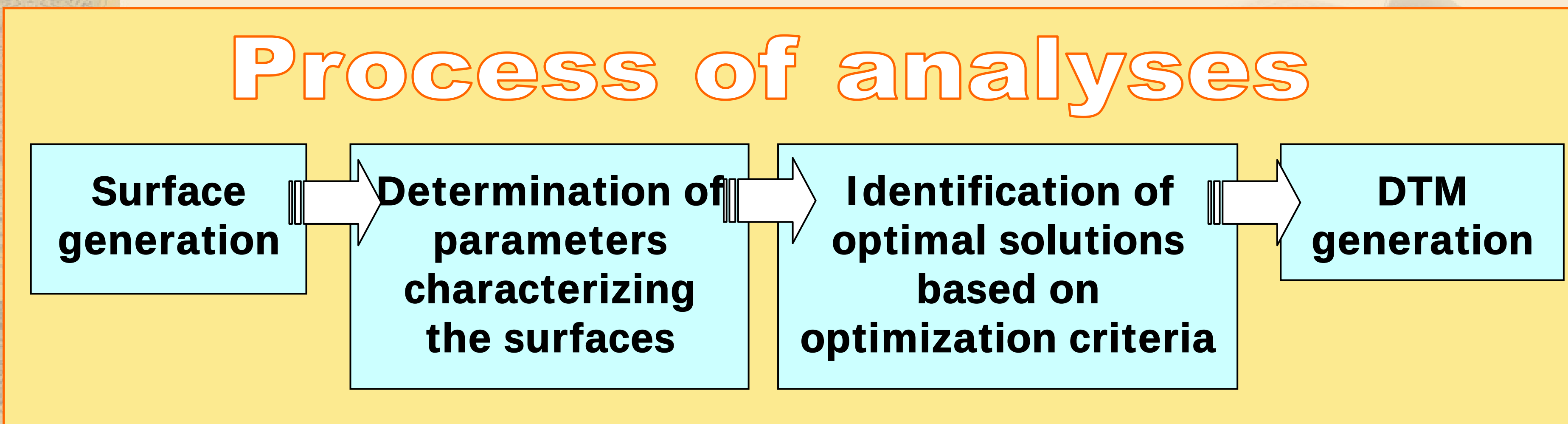
Tolerance range	*variant I* 'D-P'	*variant II* 'D-P'	*variant III* 'D-P'	*variant IV* 'D-P'	*variant V* 'D-P'	*variant VI* 'D-P'
O_1	58625	59264	59218	58907	59055	58878
O_2	57791	59107	59122	58889	59019	58871
O_3	55752	58767	59008	58842	58983	58854
O_4	53561	58396	58896	58834	58913	58838
O_5	51049	57890	58782	58755	58858	58826
O_6	45126	56862	58455	58631	58736	58784
O_7	39148	55314	58085	58474	58569	58732
O_8	29852	49113	54017	54735	54767	54729
O_9	24858	43047	52202	54114	54054	54701
O_{10}	21158	31636	44903	49791	49485	50707
O_{11}	18703	22779	34842	43593	43644	47114
O_{12}	17077	16507	23833	33567	33935	43693
O_{13}	16143	14080	18288	26210	26485	40385
O_{14}	15621	12680	14445	21415	21605	36955

Width of the search belt

variant I 0,012m
variant II 0,024m
variant III 0,037m
variant IV 0,049m
variant V 0,050m
variant VI 0,100m

Carried out analyses of reduced data sets

Generally, the process of analyses of reduced data sets can be presented in the form of the following chart:



Surface generation

Generation of GRID:

GRID P_w - standard surface

GRID P_i - surfaces generated from points of reduced sets obtained after applying the optimization algorithm based on the D-P method (where *i=1,2,...14*)

GRID P_w - ***GRID P_i*** = ***GRID P_w- P_i***

GRID P_w- P_i - differential surfaces represents a difference between the standard surface ***GRID P_w*** generated using all points of measurement set and the surface generated using points of a reduced set ***GRID P_i***

Determination of parameters characterizing the surfaces

We selected parametrs:

$$m = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} \left(z_i - z_{śr}\right)},$$

Average surface error defining the mean deviation of individual values of a feature and the arithmetic average.

$$V = \frac{m}{z_{śr}},$$

Surface variability index that is the quotient of the absolute measure of variability of a feature and the average value of that feature is an abstract value.

Identification of optimal solutions based on optimization criteria

The optimum solution is the reduced set of measurement results that generate the ***GRID*** surface that as a result of the difference from the standard surface ***GRID*** would generate the ***differential surface*** - offering the lowest average error ***m*** and the lowest variability index ***V***.

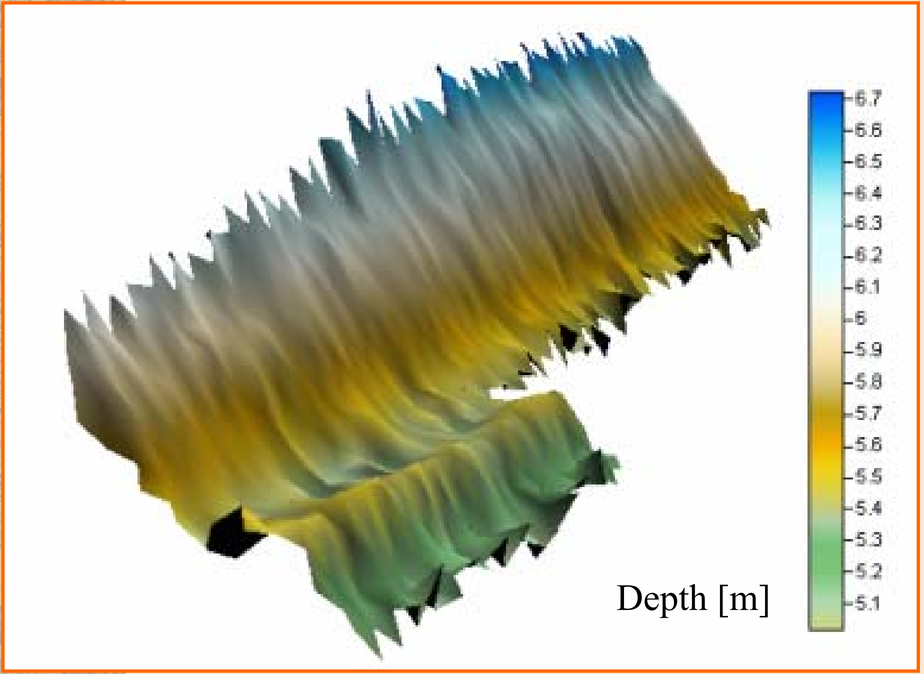
$$m = m_{min} \text{ and } V = V_{min}$$

TABLE 3

D-P METHOD						
Surface	**Number of data set points**	**Search belt width**	**Tolerance range**	**Differential surface**	m [m]	V [m]
P_2^{D-P}	57791	*wariant I*	O_2	$P_w - P_2^{D-P}$	0,0155	19,8580
P_4^{D-P}	58396	*wariant II*	O_4	$P_w - P_4^{D-P}$	0,0155	17,9677
P_7^{D-P}	58085	*wariant III*	O_7	$P_w - P_7^{D-P}$	0,0155	18,6642
P_7^{D-P}	58474	*wariant IV*	O_7	$P_w - P_7^{D-P}$	0,0155	17,7922
P_7^{D-P}	58569	*wariant V*	O_7	$P_w - P_7^{D-P}$	0,0155	18,2309
P_7^{D-P}	58732	*wariant VI*	O_7	$P_w - P_7^{D-P}$	0,0155	18,9378

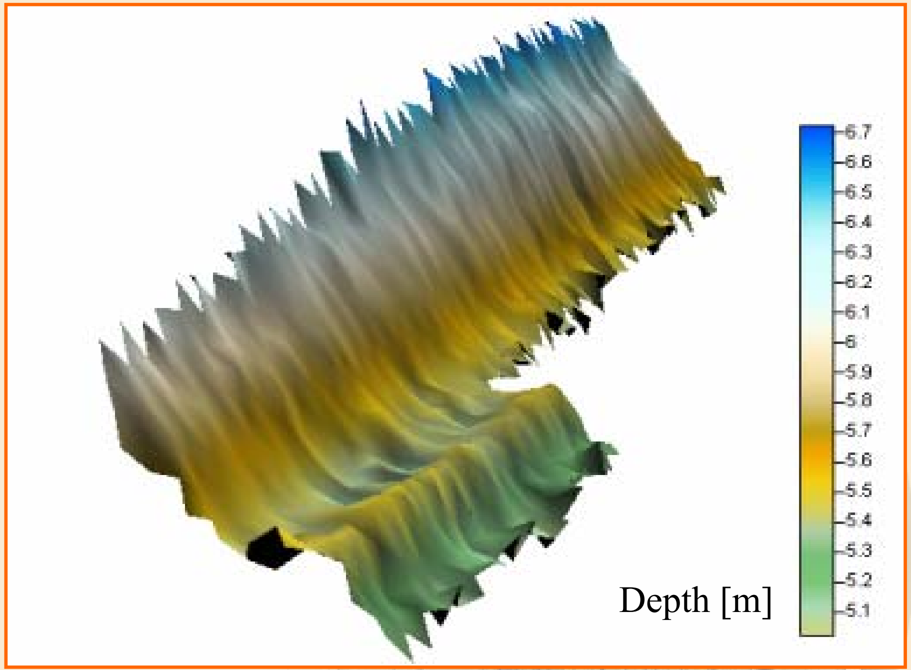
DTM generation

Fragment of Swinoujscie-Szczecin channel



DTM'R'

generated on the basis of the real number of measurements set points (60857 points)



DTM'D-P'

generated of 58474 points of the set obtained after applying the optimization algorithm based on the D-P method

Conclusions

1. The initial data processing stage using the proposed methods allowed obtaining optimally reduced data sets.
2. Identification of the optimum solution was possible by applying assessment parameters for the generated surfaces. The choice of those parameters depends mainly on the goal of the work.
3. In the presented drawings we see very similar digital models of the bottom. We do not observe distortions in the courses of bathymetric contour lines in two presented variants. We did not lose the essential data, so we can use reducing set in main proccesing.
4. It can be concluded that the selected reduced data set can be used in further studies and analyses, particularly when the large data set resulting from the direct measurement is inconvenient for the main processing.

Thank you for attention