TECHNICAL SCIENCES
Abbrev.: Techn. Sc., No 7, Y. 2004

SOME CONCEPTS OF PROCESSING BIG SPATIAL DATA SETS

Krzysztof Bojarowski, Dariusz Gościewski
Institute of Geodesy
University of Warmia and Mazury in Olsztyn

Key words: big data set, digital terrain model, spatial information system.

Abstract

Observation sets generated by continuous-recording devices should be adjusted to the spatial data model included in the system before they are used for digital map compilation or creating databases of spatial information systems. The article presents some concepts and general assumptions of processing big sets of observations, applied for generation of digital terrain model and selecting points representing the shape of line features.

WYBRANE KONCEPCJE PRZETWARZANIA DUŻYCH ZBIORÓW DANYCH PRZESTRZENNYCH

Krzysztof Bojarowski, Dariusz Gościewski
Instytut Geodezji
UWM w Olsztynie

Słowa kluczowe: duże zbiory obserwacji, numeryczny model terenu, system informacji przestrzennej.

Streszczenie

Zbiory obserwacji generowane przez urządzenia o działaniu ciągłym wymagają wstępnego przetworzenia w celu dostosowania ich struktury i wielkości do tworzenia mapy numerycznej i planowanych funkcji systemów przestrzennych. W artykule przedstawiono koncepcję oraz ogólne założenia przetwarzania dużych zbiorów obserwacji, wykorzystywanych do generowania numerycznego modelu terenu i wyboru punktów reprezentujących kształt obiektów liniowych.

Introduction

Computer technology makes it much easier to perform various tasks connected with spatial information processing. Increased computing speed and capacity of primary and disk storage allow to process bigger and bigger data sets. At the same time, the use of modern technologies of data acquisition, where the operator is no longer needed for point positioning and recording, makes it necessary to search for new methods of data evaluation (KISTOWSKI, IWAŃSKA 1997, PARKER 1996). This type of observations is time-saving, enables high accuracy and data updating. The need to modify the methods of spatial data processing results also from the application of automatic procedures during graphical-numerical processing. Automatic vectorization of line features generates similar set structures as those obtained for some GPS measurement methods (ACKERMANN 1996, BOJAROWSKI, GOŚCIEWSKI, SZACHERSKA 2000, BOJAROWSKI, GOŚCIEWSKI, WOLAK 2002). In both cases sequences of pairs of coordinates of points representing the structure shape are obtained. Also the fact that data sets are bigger and bigger makes it necessary to use procedures in which the role of the operator is as limited as possible. The optimization of the size of data sets included in spatial system databases and designed for archiving is aimed at efficient performance of system functions and – if possible – real-time data processing (BOJAROWSKI, GOŚCIEWSKI, WOLAK 2002).

Data set characteristics

The size and structure of observation sets generated by continuous-recording devices should be adjusted to the spatial data model included in the system before they are used for digital map compilation or creating databases of spatial information systems. Processing can be divided into four main stages:

– initial data processing,
– spatial and statistical analysis of the data set,
– data processing,
– conversion of data sets to the needs of spatial system structure.

Preliminary data evaluation is aimed at dividing sets into elements whose spatial scope and capacity make them convenient for processing. At this stage the effects of random errors are also eliminated by introducing the necessary corrections.

Verified and corrected measurement results provide the basis for determining the indices of spatial and statistical characteristics of sets which are used for determining the parameters and criteria of data processing at the next stages.

The selection of algorithms for the processing of measurement results generated by continuous-recording devices depends, to a great extent, on the measurement method and the method of data recording in the set. Taking into account a spatial structure of observation results, they can be divided into (Fig. 1):
– regular model,
– profile system,
– dispersed model.

Measuring techniques with fully automated result recording are efficient and time-saving tools for spatial information acquisition, especially in the case of area features, e.g. while generating digital terrain models (Bojarowski 2001, Bojarowski, Gościewski, Wolak 2002).

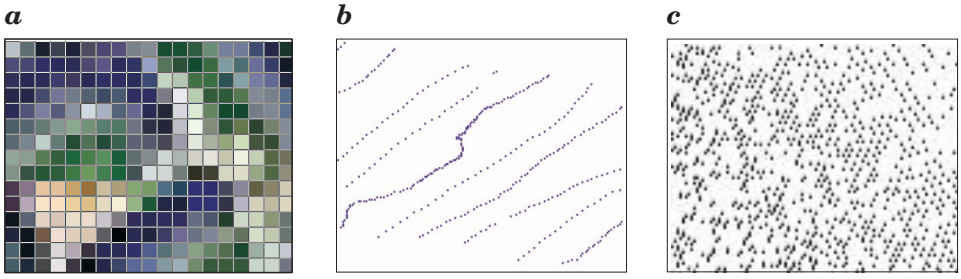

Fig. 1. Spatial structure of observation results: *a* – regular model, *b* – profile system, *c* – dispersed model

The conversion of data sets format according to the specification of the spatial system structure is aimed at reducing the number of their elements, maintaining high accuracy (quality parameters) of information. The structuring of data, enabling their use for the purposes of Land Information Systems (set ordering according to the criteria established and set saving according to the requirements of LIS), must be adjusted to the spatial data model included in the system.

General principles of data processing

The data sets discussed require special processing due to a high number of observation results. The following factors should be taken into consideration:
– full automation of processing,
– the possibility of making a qualitative and quantitative evaluation of processing results,
– the possibility of adjusting the size of data sets and data record structure according to the user's requirements.

The general diagram of processing (Fig. 2) presents the principles of establishing of information processing criteria. The main factor determining measurement method selection and principles of data acquisition and processing are the user's requirements. They result first of all from technical and economic conditions of task performance. In some cases (especially while designing land information systems) also legal aspects and engineering standards should be taken into account.

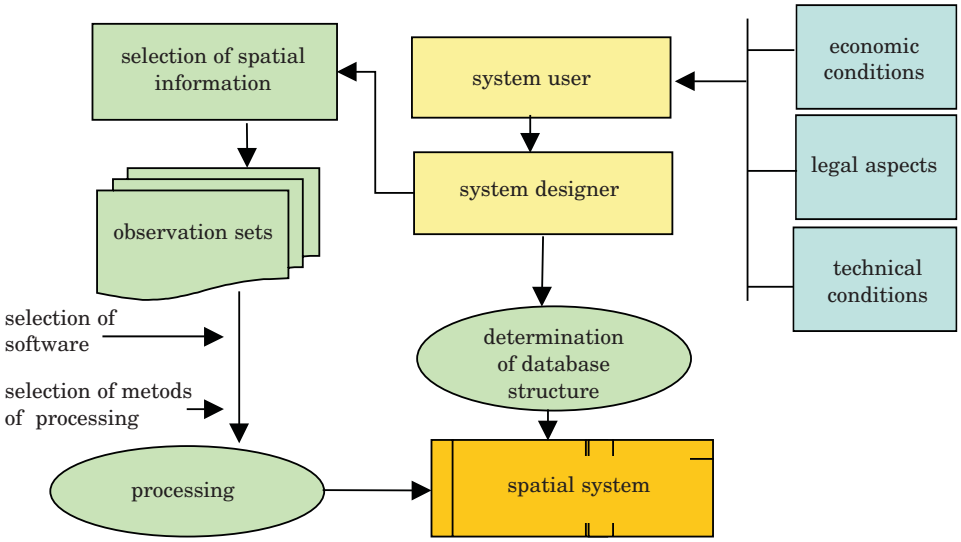

Fig. 2. General diagram of processing

Algorithms for spatial data processing should be formulated taking into consideration the planned spatial data model, especially the dimension and graphical representation of features. Fig. 3 presents methods of data modelling depending on the dimension of a structure recorded in the spatial system. It is natural that observations should allow to determine the position of a point in the system. That is why data set processing most often consists in analysis of geometric relationships between the points recorded. The result of such analyses are sets representing shapes of line features or boundaries of area features. Sometimes the above analysis concerns also other attributes whose value affect the processing results, like in the case of digital terrain models or profile courses.

The next part of article presents some concepts and general assumptions of processing sets of observations, applied for generation of digital terrain model and selecting points representing the shape of line features, resulting from the measurements by a modern systems.

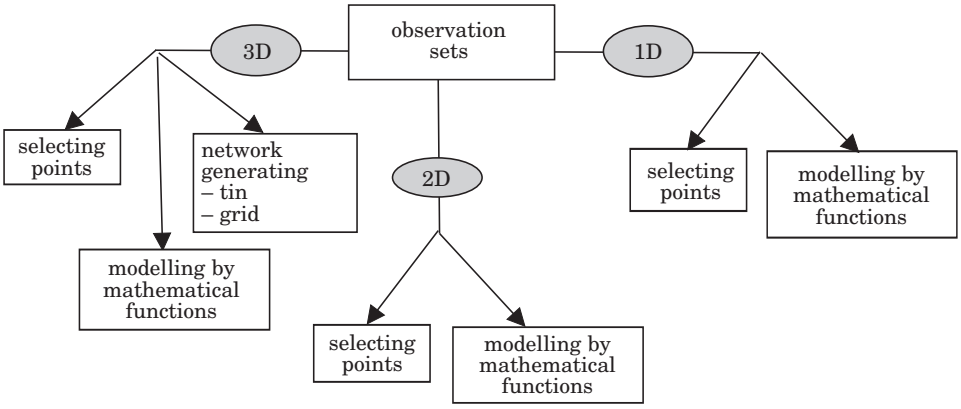

Fig. 3. Method of data modelling

Digital terrain model (DTM)

Digital terrain model (DTM) constitutes the basic information layer for systems describing spatial phenomena (ADAMCZEWSKI 1998, PARKER 1996). Due to its fundamental importance and common use in many spheres of social and economic life, it has to satisfy special requirements, i.e.: accuracy of data acquisition and processing, the possibility of data updating, completing and dynamic use. All these factors make specialists search for better and better methods of information acquisition and processing. The approach to data evaluation is changing due to the latest technological achievements.

Modern measurement systems, based on devices enabling continuous and fully automated recording of observation results, allow to obtain – within a relatively short time – a large amount of information about area features, including terrain surface (ACKERMANN 1996, PARKER 1996). Data representing the shape of area features can be acquired, among others, by and with:

– GPS measurements,
– laser aerial scanning,
– multiple-beam echo sounder,
– laser measurement stations.

All of the above measurement systems allow to record the position (spatial coordinates) and attributes (e.g. height of points) of millions of points during a single session. Evaluation of such a high number of observations, usually characterized by irregular spatial distribution, requires the application of special methods and properly selected processing algorithms.

When the scope of processing had been selected, sets are divided into sub-sets, in order to reduce the amount of data being processed simultaneously (BOJAROWSKI, GOŚCIEWSKI, WOLAK 2002, GOŚCIEWSKI 2002). Then the method of DTM representation is determined. The numerical terrain model

is usually generated on the basis of points distributed on the terrain surface and organized into certain structures (ACKERMANN 1996, ADAMCZEWSKI 1998, KISTOWSKI, IWAŃSKA 1997). The most common structures are:
– irregular triangle network (TIN),
– regular square network (GRID).

The decision if the model is to be represented by TIN or GRID depends first of all on the purpose and scope of processing, as well as the requirements of spatial information systems.

While generating DTM it is important that points (nodes) representing the approximated terrain surface are uniformly distributed. This allows to reduce the number of points recorded in the database to the necessary minimum, maintaining maximum accuracy of terrain model representation. Determining the right distance between nodes is closely correlated with the terrain model accuracy and the scope of further data evaluations based on it. Both aspects are usually determined at the beginning by the client (investor). The right densification of nodal points, which are the only ones to be recorded in the database, constitutes an important element of terrain model structure. There is a close correlation between the length of sides of triangles or squares and densification of measurement points. If it is high, TIN and GRID structure design and uniform area coverage do not pose a problem. The situation is different when densification of terrain points is insufficient or non-uniform.

The right length of sides of triangles or squares is then determined on the basis of prior statistical analysis concerning point dispersion on the area to be elaborated (FOSTER, KESSELMAN 1998, KURCZYŃSKI 1999). The area analyzed is divided into sub-areas with different densification of measurement points (*P*1, *P*2, *P*3, ...) and the coefficients of direct distance between TIN (Fig. 4a, b) or GRID (Fig. 4c) nodes are determined. The distances between TIN or GRID nodes can differ depending not only on measurement point densification, but also topographic and morphological features of terrain (GOŚCIEWSKI 2002, KURCZYŃSKI 1999, SAMET 1990). Smaller distances between nodes are usually typical of terrains showing higher morphological diversity,

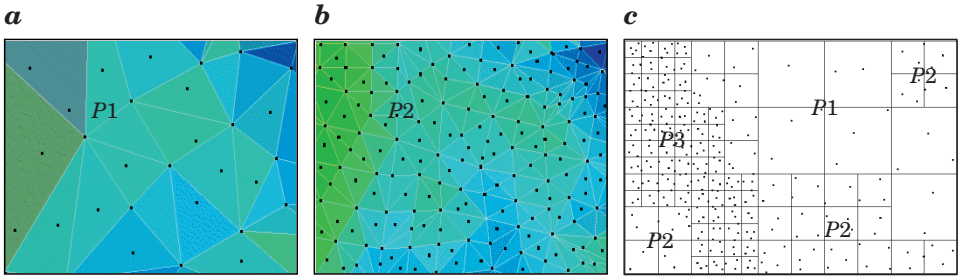

Fig. 4. Determination of density of nodes in the TIN (a,b) and GRID (c) models
*P*1, *P*2, *P*3 – densification of measurement points

whereas greater – of those with minor local height differences. The coefficients (*W*1, *W*2, ...) determining distances between nodes can be determined also in this case (Fig. 5). These distances are supposed to characterized a given area with the highest possible accuracy.

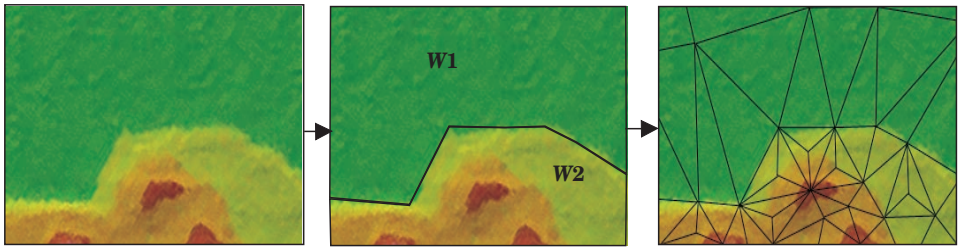

Fig. 5. Determining of distances between nodes: *W*1, *W*2 – The coefficients determining distances between nodes can be determined also in this case

The TIN model assumes uniform terrain coverage by points distributed in the vertexes of triangles. It is based on points that are directly measured, so there is no need for height interpolation or extrapolation at nodes. After covering a set of observations with a network of triangles, the points situated near triangulation nodes are to be found. The range of search around nodes is closely connected with the accuracy of observations and should not exceed the values resulting from the ellipse of errors of point positioning (Fig. 6). The triangulation node (triangle vertex) is a measurement point found in the course of search. This method of DTM generation allows to preserve natural relation between morphological features of terrain and proper topological relation between measurement points.

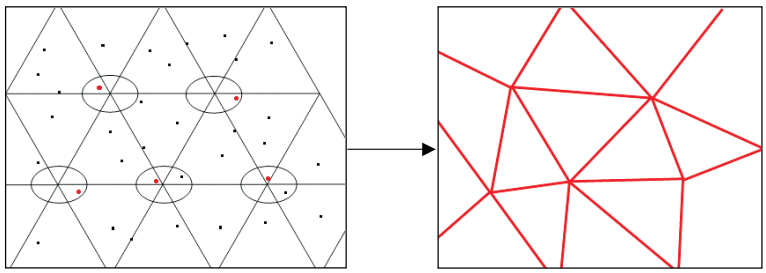

Fig. 6. Principles of generating of the triangle network

The GRID model is based on a regular square network, uniformly covering the measurement area. The distances between nodes (lengths of square sides) are determined in a similar way as in the TIN model. This method is not based on natural terrain points, so the height at nodes must be determined with interpolation algorithms. In special cases, at high densification of measured points, it may be assumed that the height at particular node

can be determined by assigning to its height of point situated within a given distance. This distance, similarly as in the TIN model, is directly correlated with the ellipses of errors of point positioning (Fig. 7). However, this approach does not allow to determine the height at nodes in whose vicinity there are no measured points.

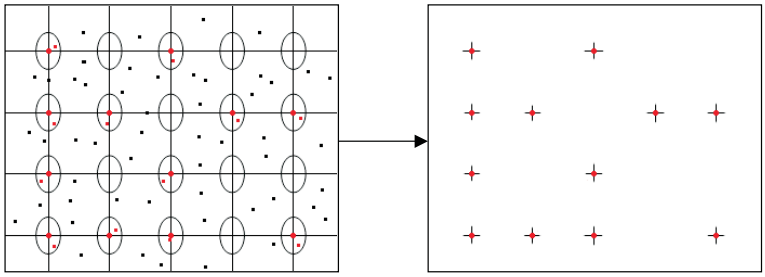

Fig. 7. Principles of generating of square network

The range of search around nodes should be determined taking into account measurement point densification (Bojarowski, Gościewski, Wolak 2002, Gościewski 2002), to reduce the number of operations during set searching. If all measured points were to be used, the radius of search would be equal to half-diagonal of the GRID-forming square.

The operating speed of interpolation algorithms is especially important in the case of a large amount of information. If the number of operations is to be as low as possible, the number of points used for calculations must be also reduced to the necessary minimum. Moreover, to achieve the required accuracy and reliability while determining the height at particular nodes, the calculations must be based only on points characterized by strictly defined spatial structure. Therefore, the points which fulfill the interpolation condition (are located in the vertexes of a circumscribed triangle, forming an approximating plane) and are situated nearest the node (*W*1 in Fig. 8) must be found in the set of points surrounding the node. Extrapolation should be avoided, especially in the case of high point densification. Choosing three points situated near one straight line is connected with the risk of a too

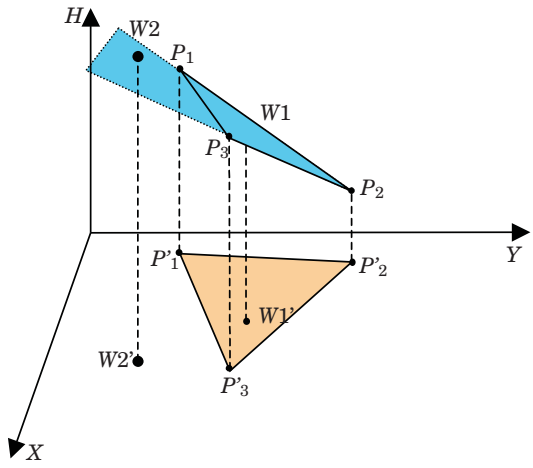

Fig. 8. Determining of the height of points: P_1, P_2, P_3 – The points of triangle, $W1$,$W2$ – The node

sharp slope of the plane passing through them, which in the case of extrapolation results in considerable errors in the values determined (*W*2 in Fig. 8).

While processing big data sets it is important to find properly distributed points, as this enables the application of simplified, quick interpolation algorithms, maintaining the required accuracy of results. One of the methods allowing to find the necessary points around nodes within a short time is division of the research area R into sections and placing the node in the middle. If the area is divided into three sections the points determined not always fully meet the assumed condition, but a triangle formed in this way enables height interpolation characterized by a minor error, because this triangle is situated near the node (Fig. 9a). According to the second method the area is divided into six sections, and points are searched for in sectors M or N. This allows to find two triangles, and then choose that one in which the sum of squares of distances from the node is the smallest (Fig. 9b). The third method consists in forming a circumscribed quadrilateral from the points surrounding the node. From the triangles formed by dividing the quadrilateral with diagonals the two which contain nodes are chosen. Then, as in the previous method, the triangle whose vertexes are situated within the shortest distances from the node is used for further calculations (Fig. 9c).

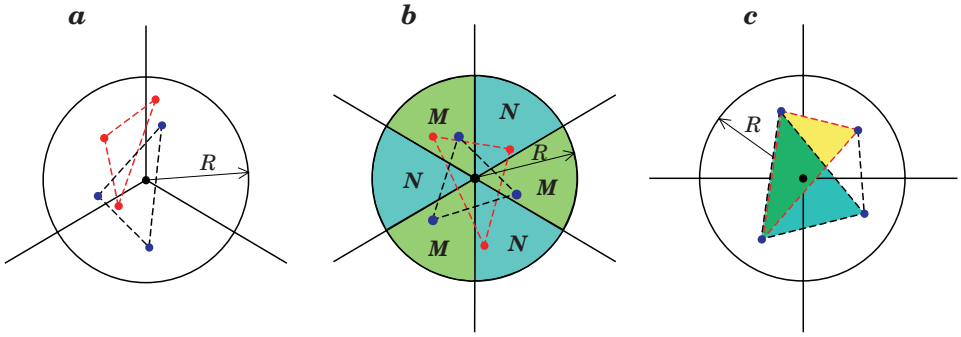

Fig. 9. Methods of the search of points to height interpolation: *R* – The research area, *M* , *N* – The sectors of search *a* – three sections, *b* – six sections, *c* – four sections

Interpolation can be carried out on the basis of points found near the node, by one of commonly applied methods (Douglas, Parker 1973, Samet 1990). Due to a considerable amount of data to be processed, the quickest method is considered to be the best. Proper point distribution around the node allows to choose the simplest method. Taking these factors into account, it seems advisable to employ the method of weighting by reciprocal of squares of distances (Fig. 10a). It enables very quick height calculation for particular nodes, but the value determined depends on the nearest points to a too high degree. The advantage of this method is that the calculations can be done using only two points located near the node. Another method of interpolation, equally quick and having no such disadvantages, is the

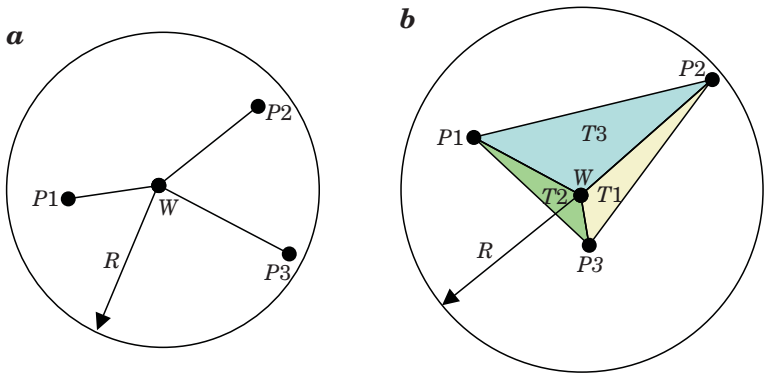

Fig. 10. Principles of the interpolation of nodes: R – The research area: $T1$, $T2$, $T3$ – The areas of opposite triangles $P1$, $P2$, $P3$ – The points of triangle, W – The node; a – The method of weighting by reciprocal of squares of distances, b – The method of weighting by areas of opposite triangles

method of weighting by areas of opposite triangles (Fig. 10b). This method enables proportional equalization of the resultant value, but does not provide the required accuracy for nodes situated outside the circumscribed triangle.

If there are no measurement points around the node, interpolation may be based on neighboring nodes. Then the mean value for linear interpolation carried out in both directions of the square network should be adopted.

Other interpolation methods can also be used for height determination at particular nodes. Special attention should be paid to the least square method, neural networks, algorithms of differential or radial equations. However, all these methods quite time-consuming, so they cannot be rationally used for real-time processing of very big data sets.

Generalization of line feature shape

Algorithms for processing data representing the shape of line features or the boundaries of area features may concern observation results acquired by or with:

– GPS method,
– vertical echo sounder,
– continuous digitization,
– automated vectorization of raster images.

Some proposals concerning processing of this type of observations have been presented in previous publications (Bojarowski 1995, Bojarowski 2002). The new methods discussed in this paper can be employed for processing measurement results obtained in various ways. Their main assumption is that the points selected from a set must fulfill certain geometric conditions.

Data processing begins with fixing the starting point, which together with the next one forms base b1-2 (Fig. 11). A local system of coordinates is determined on the basis of the points forming the base. The system origin is at point 1, and the direction determined by 1-2 is the direction of y axis. The coordinates of the points analyzed are transformed to this system of co-ordinates by the following formulas:

$$x' = x_i - x_p + x_i \cdot \cos\alpha + y_i \cdot \sin\alpha ,$$

$$y' = y_i - y_p - x_i \sin\alpha + y_i \cos\alpha ,$$

where:

x', y' – co-ordinates of points in the local system,

x_p, y_p – coordinates of the starting point,

x_i, y_i – coordinates in the coordinate system of a digital map,

α – angle of system rotation.

In this system co-ordinates x of successive points recorded in the set – 3,4,5 ….n are analyzed. If during the analysis co-ordinate x of the point analyzed (8) exceeds the admissible value dx of point deviation from the line being generalized, the previous point (7) is recorded and used for forming a new base with the already analyzed one (b_{78}). The process of point selection continues until the last point in the sequence of coordinate pairs is analyzed. The broken line shows the shape of a line feature or the boundaries of an area structure.

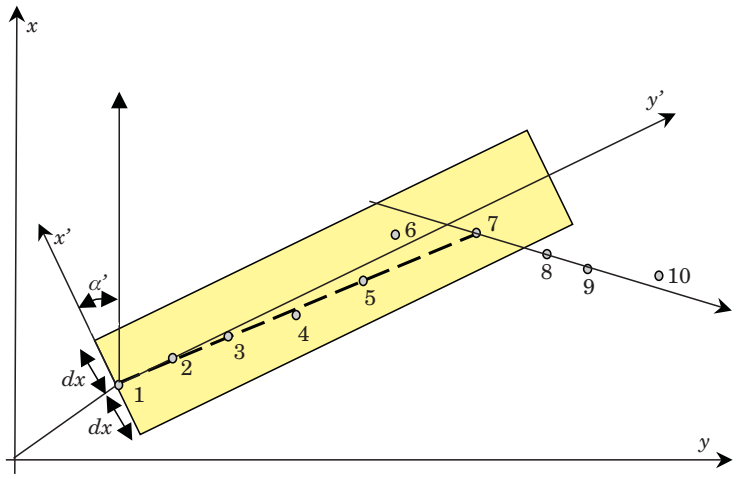

Fig. 11. Generalization of the line feature

Profile feature processing

Profile structure processing is similar to the already described method of selecting points representing the shape of line features. In this case the geometric condition constituting the criterion of point recording is formulated on the basis of point height at the profile.

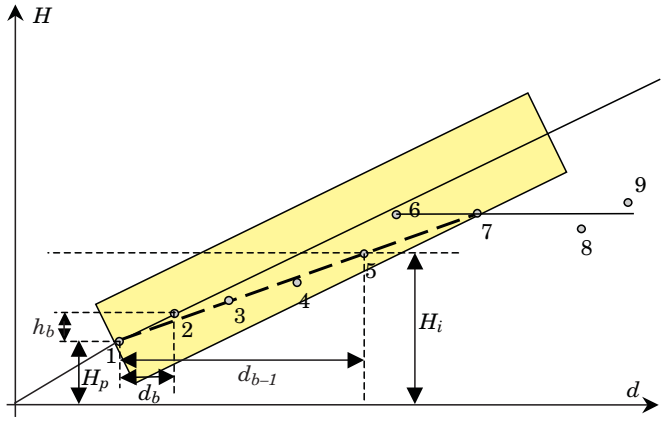

Fig. 12. Processing of the profile structure

As shown in Fig. 12, the first two analyzed points form the first base. The condition being the recording criterion is determined from the dependence:

$$\left| \left(H_i - H_p\right) - \frac{d_{p-i} \cdot \Delta h_b}{d_b} \right| \leq \Delta p,$$

where:

H_i – height of the point analyzed,
H_p – height of the starting point,
d_{p-i} – distance between the starting point and the point analyzed,
dh_b – height difference between base points,
d_b – base length.

Conclusions

The directions of measuring technique development suggest that the information contained in spatial system databases will be more and more frequently acquired with the help of automatic recording devices. The studies conducted so far indicate the necessity to apply special algorithms in the

process of data preparation for the creation of spatial system databases, mainly due to a large number of observations. The main aim of data preparation is to adjust the size and structure of sets to the planned system functions. The choice of a processing algorithm depends first of all on the type of structure (point, line, area, 3D), and the processing criteria are determined on the basis of geometric or topological relationships between measurement points. Moreover, the structure and size of sets should correspond with the assumptions made during spatial data modelling in the system.

References

ACKERMANN F. 1996. *Techniques and Strategies for DEM Generation*. An Addendum to the Manual of Photogrammetry ASPRS.

ADAMCZEWSKI Z. 1988. *Wprowadzenie do numerycznego modelowania terenu*. VIII Konferencja Naukowo-Techniczna *Systemy informacji Przestrzennej*. Warszawa.

BOJAROWSKI K. 2001. *Graficzna analiza rozmieszczenia obiektów o wybranych atrybutach w zarejestrowanych systemach przestrzennych*. Materiały XI Konferencji Naukowo-Technicznej *Systemy Informacji Przestrzennej*. PTIP, Warszawa.

BOJAROWSKI K., GOŚCIEWSKI D., SZACHERSKA M. K. 2000. *Wizualizacja zmian ukształtowania dna morskiego jako etap modelowania procesów w systemach przestrzennych*. Materiały XII Międzynarodowej Konferencji Naukowo-Technicznej *Rola nawigacji w zabezpieczeniu działalności ludzkiej na morzu*. Gdynia-Oksywie.

BOJAROWSKI K. 2002. *Opracowanie wyników pomiaru echosondą pionową*. Materiały XIII Międzynarodowej Konferencji Naukowo-Technicznej *Rola nawigacji w zabezpieczeniu działalności ludzkiej na morzu*. Gdynia-Oksywie.

BOJAROWSKI K., GOŚCIEWSKI D., WOLAK B. 2002. *Technologia przetwarzania wyników pomiaru ukształtowania dna rejestrowanych przez urządzenia o działaniu ciągłym*. MKN-T EXPLO-SHIP 2002 *Problemy eksploatacji obiektów pływających i urządzeń portowych*. Zeszyty Naukowe WSM w Szczecinie, Szczecin.

DOUGLAS D.M., PARKER T.K. 1973. *Algorithms for the reduction of the number of points required to represent a digitized line or its caricature*. Canadian Cartographer.

FOSTER I., KESSELMAN C., 1998. *Blueprint for a Future Computing Infrastructure*. The Grid Morgan Kaufmann Publishers, San Francisco, Calif.

GOŚCIEWSKI D. 2002. *Optymalizacja struktury i wielkości zbiorów obserwacji wykorzystywanych do tworzenia numerycznego modelu dna*. Materiały XIII Konferencji Naukowo-Technicznej *Rola nawigacji w zabezpieczaniu działalności ludzkiej na morzu*. Gdynia – Oksywie.

KISTOWSKI M., IWAŃSKA M. 1997. *Systemy informacji geograficznej*. Wydawnictwo Naukowe, Poznań.

KURCZYŃSKI Z. 1999. *Technologiczne uwarunkowania budowy numerycznego modelu rzeźby terenu*. Prace IGiK, Warszawa.

PARKER D. 1996. *Innovations in GIS. University of Newcastle*. Newcastle.

SAMET H. 1990. *The design and analysis of spatial data structures*. Reading, MA. Addison-Wesley.

SZACHERSKA M.K., BOJAROWSKI K. 1997. *GIS – a Goal or a Tool*? International Archives of Photogrammetry and Remote Sensing, 31, part GW1.

Accepted for print 2003.11.06