River Network Selection based on Structure and Pattern Recognition

Guillaume Touya

River Network Selection based on Structure and Pattern Recognition

Guillaume Touya

COGIT Laboratory, Institut Géographique National,
2 av. Pasteur, 94160 St-Mandé, France
guillaume.touya@ign.fr

1. Introduction

Generalisation aims at reducing the level of detail of a database in order to meet new specifications. A lot of work have dealt with automated generalisation for many years. Most of them concerned cartographic generalisation that aims at producing maps but this paper focuses on purely database (or model) generalisation. This paper addresses a specific and important problem of database generalisation, the generalisation of river network data. The paper emphasizes on the problem of selection: which stream is important enough to be part of the generalized dataset? River network selection has already been treated in the past by [12,9]. The solution presented here particularly employs the principle of "good continuation" to enrich the database with "river strokes" [15].

The second part of the paper presents a state of the art of river network generalisation. The third part explains the problems, the hypotheses and the pre-processes necessary for selection. The fourth part describes the generalisation process proposed to solve the problem. The fifth part presents the results obtained with this process on actual French datasets. Finally, this paper is concluded with leads for further research.

2. River Network Generalisation

2.1 Generalisation

As generalisation is a key issue in data presentation and integration, a lot of work deal with the topic. It is called cartographic generalisation when it aims at producing maps but this paper deals with database generalisation [19]. The purpose of database generalisation is to reduce the level of detail of a whole database transforming both the schema and the data, without considering a specific symbolisation of geographic objects.

The geo-database generalisation process may be decomposed in several steps [8]: selection, attribute generalisation and geometric generalisation. Only selection is detailed

in this paper but a full process has been developed for river network database generalisation.

[7] also raises the importance of spatial analysis to detect actual patterns and enrich the database introducing the detected structures. It allows to guide and improve the generalisation process. [3] and [18] highlight the importance of pattern recognition and generalisation in road networks and it can be applied to river networks although patterns in natural networks such as river networks are different from man-made networks such as road networks.

2.2 Perceptual Grouping and "Strokes"

The principles of perceptual grouping were firstly enunciated by the Gestalt psychologists [20]. Perceptual grouping describes the phenomena whereby the human brain organizes all the elements of his visual field. The laws of perceptual organisation play a key role in understanding the two-dimensional images of three-dimensional landscapes so as a consequence in understanding maps and semiology [6].

These laws also play a key role in map generalisation and they have been used for a long time [8]. In the case of linear networks, an interesting law is the principle of "good continuation". It allows to group the segments of a network in "strokes", sets of arcs that appear to be continuous (as in a straight line or curve) [16]. In maps or geo-databases, road or river strokes have a real meaning as roads and rivers are quite continuous phenomena. For example, in a road network, a ring road is a stroke. [15] applied the principle of good continuation to build strokes in road and river networks in order to perform their selection for generalisation.

2.3 River Network Selection

River network generalisation has already been tackled for many years. [12] enunciated that this problem is mainly a selection problem. It could be solved using an organisation into hierarchy of the streams coupled with their length. The main classification methods for streams are Strahler [14], Horton [4] (Fig.1) and Shreve [13].

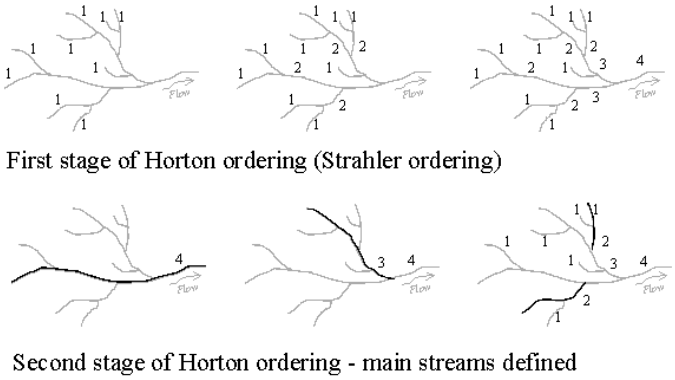

Fig. 1. Description of Strahler and Horton ordering used for the selection. The first step is the Strahler ordering of each stream segment. Then the Horton order of a stream is the maximum of the Strahler orders of its segments. Picture from [15].

[10] compared the different methods of organisation into hierarchy and ways to determine the main stream. He also highlighted the problems raised by these different methods. For example, determining the main stream using the longest path [4] on clipped river network (the streams are clipped excluding their source or their sink) causes big mistakes.

[12] and [9] propose a generalisation model using hierarchy abstraction and aggregation. [15] used the principle of good continuation to build "river strokes" on a river network. The selection finally consisted in thresholding the strokes using their Horton order and their length. In figure 1, the streams highlighted for Horton ordering are ordered "river strokes".

In this paper, we present a method for river network selection, already introduced in [17], that relies on the organisation of river strokes in hierarchy. The additional contributions are a pre-processing step, the structure detection step that helps dealing with islands or irrigation areas and the management of clipped drainage basins.

3. Hypotheses on the Proposed River Network Selection Process

3.1 Background and Data

The problem we deal with in this paper is the selection of linear river networks in the context of database generalisation. The main application case is related to two IGN (French National Mapping Agency) databases : BD TOPO® database which resolution is 1 m and level of detail is approximately 1:15000 and BD CARTO® database which resolution is 10 m and level of detail approximately 1:100000 (Fig.2). But, the aim is to conceive a generic selection process for linear river network in model generalisation whose parameters could

allow to meet different kinds of target database specifications. In general and particularly with BD CARTO®, map or geo-database specification are quite fuzzy as specification do not exactly detail which geographic objects should be present in the database. For example, a specification could mention: "cul-de-sacs should not be in the database except if they are very long". Getting generalisation parameters from such specifications generates fuzziness. So the choice of parameters for such selection is essential for having a correct generalisation.

The hypotheses made on data for this work correspond to BD TOPO® specifications to allow the application case. In BD TOPO®, the river network is modelled as complete linear network. So the process will only deal with linear features. Besides, in BD TOPO® the information related to the flow direction of the stream segments is carried by the geometry, not the attribute data. According to the specifications, the direction of the geometry must be the flow direction. To use the proposed process with different kind of initial data, pre-processing will be necessary to obtain a complete linear and planar network with flow direction information.

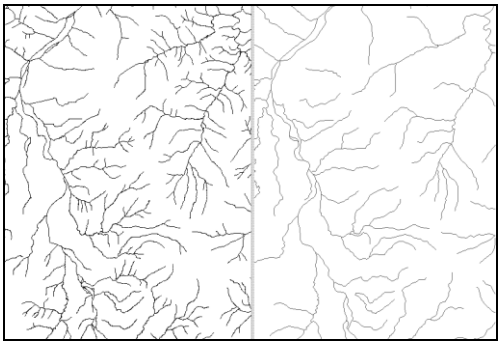

Fig. 2. Two extracts of the hydrologic data used: on the left, an extract from BD TOPO® and on the right, an extract of the same zone from BD CARTO®.

Even if a generic algorithm should not take into account varying parameters like attributes, river network selection is obviously better when attributes like river name are used. So attributes of the reference database may be considered as parameters of the overall process.

3.2 Pre-processing

The information related to the flow direction of the river segments is not always completely reliable like in BD TOPO®. So, the initial dataset has often to be cleaned to contain the fewest errors possible in flow direction.

We can use topology and geometry like in [11] to compute flow directions for all streams of the dataset. But this algorithm produces some errors and we need the data to be

as correct as possible. Thus, two processes are successively applied to infer flow directions in the river network. First, a process adapted to initial data from BD TOPO® is carried out. It consists in using elevation data that is attribute data of stream segments. Each stream segment has an elevation value for initial and final vertex. Obviously, if elevation value of final vertex is bigger than elevation value of initial vertex, stream segment geometry has to be reversed to be consistent with flow direction. The second process is more independent to initial data as it consists in a neighbourhood analysis. Locally analysing the inflow and outflow streams may allow to infer flow direction for simple cases like in figure 5a. At sources or sinks, neighbourhood analysis may correct other errors of flow direction (fig.5b).

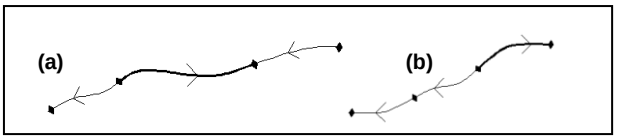

Fig. 3. (a): the lightened segment has a wrong flow direction as it is inconsistent with its neighbours. (b): the lightened segment has a wrong flow direction, inconsistent with its neighbours, and it is the first river segment from a source.

For our initial data, the two processes are sufficient. The first one corrects almost all errors provided an elevation consistency checking process is used before. The second one corrects almost all remaining errors. Very few flow direction problems remain after the corrections (less than 0.1% of segments).

Furthermore, generalisation processes often need topologically correct network data to provide meaningful results. Most real data contain errors and especially topological errors and so does BD TOPO® but very few. So the network is corrected making it a real planar graph.

After these pre-processing steps, the selection process itself can take place.

4. Proposed Selection Process

The first part of this section deals with the first step of this selection process, database enrichment, and particularly with the building of "river strokes" and islands. The second part of the section explains how the selection is made using the enriched dataset.

4.1 Data Enrichment Step

4.1.1 Enriched Data Schema for the Generalisation

Automatic generalisation, like river network selection, is a complex process that often needs to enrich the raw dataset to be generalised to recognise implicit structures and

patterns [7]. The data model below (Figure 4) summarizes all the add-ons made to allow generalisation as the initial database only contained the "River segment" class.

Like in [15], a "River stroke " class is added to store the strokes that will be used for selection. The classes "Source" and "Sink" are added to store the beginning and the end of the strokes so as to use them as database objects in further processes.

Two classes are also added in order to manage the selection of river islands. Finally, a class to store irrigation zones is added. Indeed, the strokes building algorithm does not work in these zones so they have to be previously detected.

Sections 4.1.2 to 4.1.5 explain how these classes are automatically instantiated.

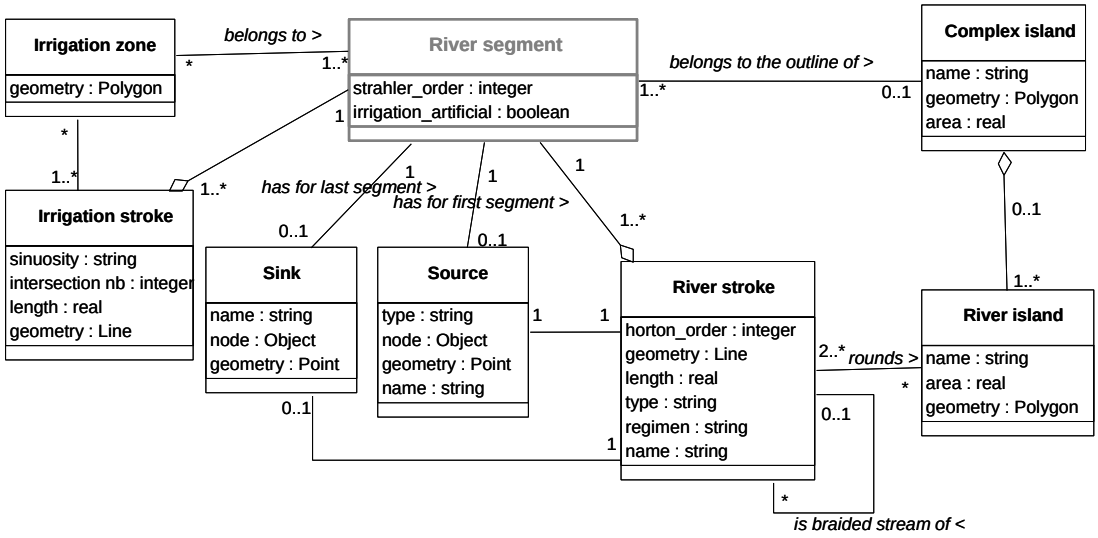

Fig. 4. Data model for hydrological network generalisation

4.1.2 Sources and Sinks

In order to build the "river strokes" and the irrigation zones, it is first necessary to build the sources and sinks. A source is a node of the network that has a single output link and no input link. A sink is the opposite: a node with a single input link and no output link. [15] already used sources and sinks but on a full drainage basin. The data model presented above allows to work on an extract of the drainage basin thanks to the attribute "type" added on the classes Source and Sink. In the studied application case, the network is clipped to the test zone, a French county (7600 km²) that does not cover the full drainage basin of area. The clipped streams have the source or sink type "zone limit", whereas the other sources or sinks are labelled as "natural" (Figure 5).

As the length of the streams may be used to determine the main stream at a confluence point, these cases have to be distinguished. Otherwise, the process would give wrong

results as the real length of the clipped streams is unknown. With this model, clipped streams may be treated differently avoiding to cause big mistakes.

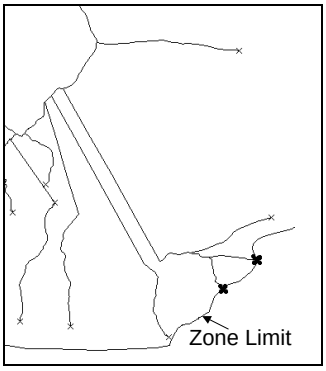

Fig. 5. The two types of sources: the simple crosses are natural sources whereas the highlighted ones are "zone limit" sources.

4.1.3 "River strokes" Building

The strokes building algorithm is a downstream pass: it begins at the sources of the network and ends when all the nodes (sources, sinks and confluences points) have been treated. It covers the network using flow directions. Wrong flow directions due to errors in the database would cause river strokes wrongly created. That is why the pre-processing presented in section 3.2 is necessary. Like in [15], "river strokes" are built to correspond to the classic ordering of rivers: a river stroke begins at a source and ends at a sink or at a confluence point with a more important river stroke. The tricky part of the creation of such strokes is to define the main stream at a confluence point. To have more realistic results, the Gestalt principle of "good continuation" is constrained by rules and the river strokes are different from the purely geometric perceptual strokes. Many rules can be used to define the main path. [4] used the longest and straightest path, [15] used the river name, the longest and straightest path, others used the largest drainage basin. The rules used here are a bit more complex:

- Strokes always follow a named stream.
- All other things being equal, a "permanent" regimen river has priority on an "intermittent" regimen river.
- All other things being equal and the sources of the streams being "natural", the longest and straightest path has priority. The length is the major parameter as [10] showed it was more relevant than straightness. Straightness is rather used when the confluence angle is greater than a threshold (60°) and if the length difference is less than a threshold (500 m).

- All other things being equal and one of the sources of the streams being "zone limit", the straightest path has the priority.

Besides, to follow the good continuation principle, straightness means here curvature continuity. The straightest path is the one that leads to more curvature continuity with the upstream stroke. The length of a river is measured from the source to the confluence point.

Now river networks are sometimes composed of deltas and islands where a stream splits into two or more channels in the downstream direction. So, two types of river strokes are distinguished: "main channels" and "braided streams" like in [15]. When a stream splits into two or more channels, one outflow is a considered as the main channel and continues the "main channel" stroke and the others are used as the first segment of new "braided" river strokes. The chosen main channel is the straightest path (Figure 6). The shortest path to the sink could have been used instead but the process would have slowed down as shortest path is a computation time consuming algorithm and the contribution is not sufficient. But this rule could be used for small datasets.

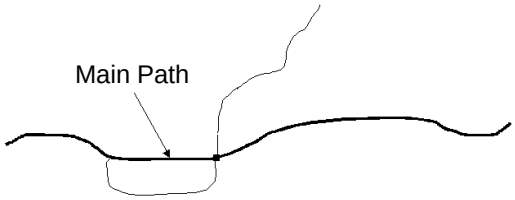

Fig. 6. Choice of the main stream on a confluence point with two outflow and two inflow streams.

4.1.4 Islands and Complex Islands

In river networks, braided streams often correspond to islands on the rivers. Moreover, there are often several adjacent islands on the river (Figure 7). These islands are significant data and it is interesting to select the outline of some islands in the generalised dataset. During the data enrichment step, the aim is to build islands and complex islands as database objects to make the selection of the outline easier. A complex island is an aggregation of islands. This idea of aggregating elements to raise the level of abstraction has been introduced in generalisation by [12]. In our enriched dataset, each island is linked to the stream segments of its outline (Figure 7), even for complex islands. If selection specifications require a different level of abstraction for islands and especially a smaller level, adjacent islands can be aggregated to several adjacent complex islands in order to keep the information of adjacent islands after the selection. Then, the enriched data model would be a bit modified to allow adjacent complex islands.

To build such islands the topologic faces of the network, seen as a graph, have to be built. Only the small faces, the ones that really represent islands, are kept to give the geometry to the new objects: a threshold is determined based on the test dataset. Complex islands are built by clustering the simple islands. In order to build adjacent complex islands for a smaller level of abstraction, a hierarchical clustering would rather be used.

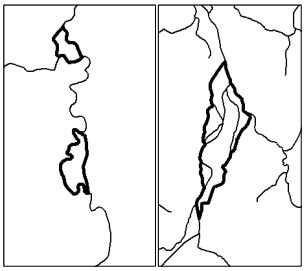

Fig. 7. Simple islands on the left picture and a complex island composed of adjacent simple islands on the right picture. Their outline is automatically created by the process.

4.1.5 Irrigation Zones

The process also allows the detection of irrigation zones. The detection is necessary because irrigation zones are sources of errors during the creation of river strokes. As explained above, the stroke creation algorithm is based on flow directions and in irrigation zones, most streams are artificial and flat and have no real flow direction. In the database, the segments in irrigation zone are given an arbitrary flow direction in the original data. As inconsistent flow directions on a river forbid the algorithm to go further on this stream, it is important to know where the algorithm may not work properly.

Irrigation zones are characterized by a flat ground, a strong density of straight and short streams and many sources and sinks. These characteristics are used to automatically detect irrigation areas and to build them as objects in the dataset. Then, compactness is used to remove over-detected areas. Figure 8 shows a correct result of the automatic detection of irrigation zones.

Fig. 8. Results of irrigation zone automatic creation : the bold polygon contours the irrigation area.

The next step is the automatic recognition of artificial streams in regards to natural ones within irrigation zones in order to remove them from the selection process. Artificial streams are rather short and straight whereas natural ones are rather long, sinuous and have many intersections with other streams. To translate the difference into geo-computational measures, the strokes are used once again but this time with geometrical strokes (only curvature variation is used). So strokes are computed in each irrigation area and then characterised by measures like sinuosity and number of intersections with other strokes in order to be differentiated (Fig. 9).

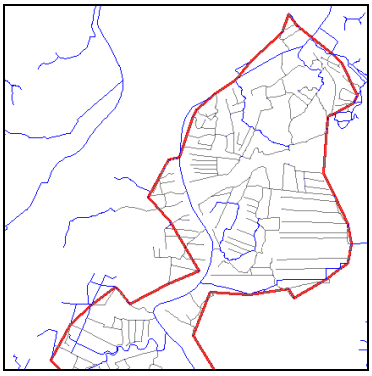

Fig. 9. Results of automatic detection of artificial streams in irrigation areas. In blue, natural streams and in grey, artificial streams

The detection of natural and artificial streams inside irrigation zones ends up the data enrichment step of the proposed river network selection process. The next step deals with the organization of strokes in hierarchy in order to select the more important ones.

4.2 Selection Step

Once the data enrichment step has been carried out, the selection step itself can take place. The selection step is based on the data enrichments described above. The major part is the selection of the river strokes. As strokes are built to represent whole rivers, selection rather concerns stokes than segments. All segments of a stroke are selected or none is. In order to determine which strokes are to be selected, the main criterion used is a hierarchical organisation of the strokes. We use the Horton ordering (Figure 1) that was initially dealing with rivers but that is applied here on strokes [15]. Indeed, the information of real rivers is not present in the initial data and is approximated by the strokes. To compute the Horton ordering, a Strahler ordering of the river segments is needed and used.

The Strahler order is assigned to the river segments during the strokes building algorithm. When a segment extends a river stroke, its Strahler order is computed like in figure 1. The segments belonging to a "braided stream" stroke are not used to compute the

Strahler order and no Strahler order is assigned to them as their selection is not based on their own hierarchy. The Horton order of each stroke is finally computed after the building of all strokes in the network: the Horton order of a stroke stream is the maximum of the Strahler orders of the river segments that compose it. The selection process is based on the river strokes using the Horton order to organise them in hierarchy. The less important ones in the hierarchy are not selected.

The selection criteria are a threshold on the Horton order of the strokes and a threshold on the stroke length. The river strokes are selected according to their type: "primary" or "braided". High Horton order primary strokes are always selected and low ordered primary strokes are selected if they are long enough. The braided strokes long enough are selected only when their primary stroke is selected. In our test case, the generalisation of BD CARTO® from BD TOPO®, the thresholds used are derived from the target database specifications. All the primary strokes with an Horton order of 3 or more are selected. The strokes with an order of 2 are selected if the length is bigger than 700 m. The strokes with an order of 1 are selected when they are longer than 1000 m. The braided strokes are selected if their primary stroke is selected and if their length is bigger than 1000 m. Obviously, parameters may be tuned for other cases.

The first experiments showed that generally, the selection criterion on the braided strokes was not sufficient to correctly select the outline of most islands and meet the target database specifications. That is why islands and complex islands are introduced in the process. A threshold is used on islands area and the river segments belonging to the island (or complex island) outline are selected. This selection does not depend on strokes. Only additional river segments are eventually selected. So, this process is complementary to the braided strokes selection.

Finally, in irrigation zones, only the natural river segments can be selected. Artificial segments that were taken away from strokes building algorithm are obviously also dropped during selection. Natural segments are grouped according to the good continuation principle just like the other segments and the strokes thus created are selected using the same criteria as the other ones.

5. Results

The complete process described in this paper has been tested for the generalisation of the river network of a French county (7600 km²) with heterogeneous landscapes (mountains,

plains, seaside, irrigation areas). It as been implemented in the GIS Clarity™ from 1Spatial (former Laser-Scan). As explained in section 3, the process has been parameterised for the generalisation of BD CARTO® (10 m resolution) from BD TOPO® (1 m resolution). So the results of the selection process can be compared to the actual BD CARTO®. Some results obtained automatically are shown in figure 10.

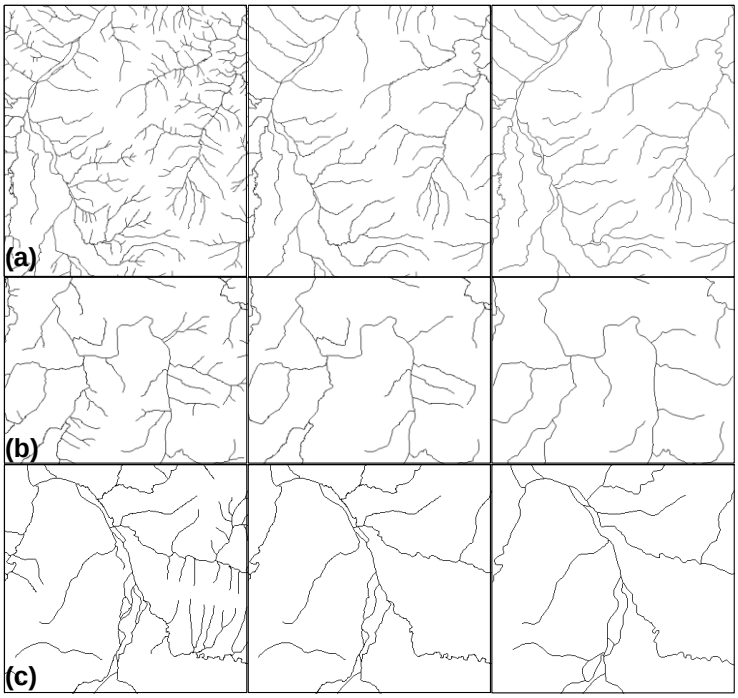

Fig. 10. Results of the generalisation process in three different areas: on the left, the initial data from BD TOPO®, in the middle, the generalized data and on the right, the target data from BD CARTO® (used to compare).

The results allow to explain what can be called "equivalence" regarding selection. As target and initial databases are totally independent in their management and their creation, up-to-dateness differences and biases in survey specifications forbid to select exactly the same real world objects during generalisation as in target database. An equivalent selection consists in selecting exactly the same important features and the same quantity of features but accepting small differences in secondary objects (Fig.10(b)).

The figure above shows that the generalized dataset is quite close to the target dataset (Fig.10(a)). The generalized dataset and the target dataset have the same density of objects and the main streams of the network are the same. Fig.10(b) illustrates the equivalent selection and moreover the generalized dataset is considered to have a higher quality than the target one from BD CARTO® (better up-to-dateness). Fig.10(c) shows some good results on the generalisation of islands.

On a perfect network, results are very good. But the corrective pre-process presented in section 3.2 is essential to avoid the remaining errors to be propagated to selection results.

6. Conclusion and Further Research

In this paper, a river network selection process for database generalisation has been proposed. This process is based on previous work on river network generalisation, especially [15] that introduces the notion of "strokes" but also [10] for the organisation in hierarchy of the streams of the network. This work adds the management of river islands, irrigation zones and allows the building of strokes on a clipped area where some sources are not natural. A data correction step dealing with network topology and flow direction has also been developed. The complete selection process is composed of three steps: data correction, enrichment and selection. The implementation of the process on Clarity™ GIS gives encouraging results on real and large datasets.

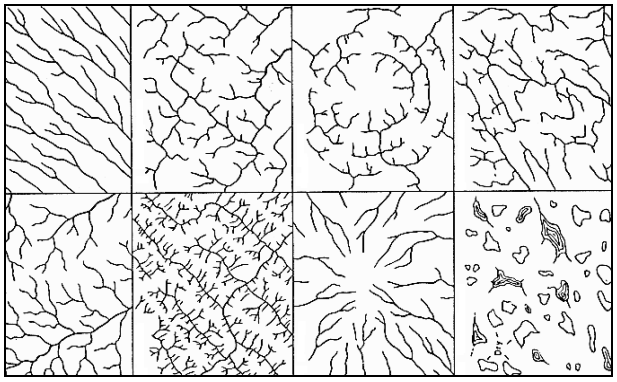

Fig. 11. Classification of drainage basin structures in relation to geomorphologic structures [5].

Further research could be done to improve selection. First, a major topic is drainage basins structures (Fig.11). Knowing rivers belong to a particular drainage structure would allow to constrain selection to maintain the characteristics of the structure. COGIT lab is now trying to develop measures and methods to automatically detect the structures. It would thus interesting to apply this work in river network selection to provide selections that maintain geomorphologic structures.

Then, structure recognition could be more focused on as other typical structures of river networks than islands and irrigation zones could be taken into account. For example, large and typical meanders could be detected to be sure the stroke main path follows correctly the meanders. Furthermore, it would be interesting to improve the contextual correction of flow direction to be able to correct all errors without using elevation data. To improve the results with clipped areas, the strokes creation algorithm could also be improved by

introducing more complex rules. Finally, when geometric transformations are important, it would be interesting to study the use complex river line simplification algorithms like [1].

References

1. Christensen, A.H.J.: Two Experiments on Stream Network Generalisation. Proceedings of the 21th ICC. ICA, Durban, South Africa (2003)
2. Deffontaines B., Chorowicz J.: Principles of Drainage Basin Analysis from Multisource Data: Application to the Structural Analysis of the Zaire Basin. Tectonophysics, 194 (1991) 237-263.
3. Heinzle F., Anders K.-H., Sester M.: Pattern Recognition in Road Networks on the Example of Circular Road Detection. in proceedings of the 4th International Conference GIScience 2006, Münster, Germany, September 2006. Full paper, pp 153-167.
4. Horton, R.A: Erosional development of Streams and their Drainage Basins: Hydrophisical approach to Quantitative Morphology .Geo. Soc. America Bull, Vol. 56 (1945) 275-370
5. Howard A.D.: Drainage Analysis in Geologic Interpretation: a summation. Bull. Am. Assoc. Pet. Geol., 51(11) 2246-2259
6. MacEachren, A.M.: How Maps Work. The Guilford Press (1995) 51-150
7. Mackaness, W., Edwards, G.: The Importance of Modelling Pattern and Structure in Automated Map Generalisation. In Proceedings of Joint Workshop on Multi-scale Representations of Spatial Data, Ottawa, Canada (2002)
8. MacMaster, R.B., Shea, K.S.: Generalisation in Digital Cartography. Association of American Cartographers, Washington D.C. (1992)
9. Martinez Casasnovas, J.A., Molenaar, M.: An Aggregation Hierarchy for Multiple Representation of Hydrographic Data Within the Context of Conceptual Generalisation. Proceedings of the 17th ICC. ICA, Barcelona. Vol. 1 (1995).
10. Mauger, F.: Hiérarchisation d'un réseau de talwegs en vue de sa généralisation. Training course dissertation Master AIST, COGIT Laboratory (1997)
11. Paiva J., Egenhofer M.: Robust Inference of the Flow Direction in River Networks. In Algorithmica 26 (2). (2000).
12. Richardson, D.: Generalisation of Spatial and Thematic Data Using Inheritance and Classification and Aggregation Hierarchies. Advances in GIS Research 2. Taylor and Francis, London (1994) 957-972
13. Shreve, R.L: Statistical low of stream numbers. Journal of Geology , Vol.74 (1966) 17-37
14. Strahler, A.N: Quantitative Analysis of Watershed Geomorphology.
American Geophys. Union Trans. Vol. 38, (6) (1957) 913-920
15. Thomson, R., Brooks, R.: Efficient generalisation and abstraction of network data using perceptual grouping. In proceedings of the 5th GeoComputation. University of Greenwich, Kent U.K. (2000)
16. Thomson, R., Richardson, D.: The "Good Continuation" Principle of Perceptual Organization Applied to the Generalisation of Road Networks. In proceedings of the 19th ICC. ICA, Ottawa, Canada (1999)
17. Touya G. 2006.: A Method for Generalisation of River Networks Based on "Strokes" and Database Enrichment. Extended Abstracts Proceedings of 4th International Conference GIScience 2006, Münster, Germany. pp 191-194.
18. Touya G. : A Road Network Selection Process Based on Data Enrichment and Structure Detection. In proceedings of the 10th ICA Workshop on Generalization and Multiple Representation, Moscow (Russia), 2007.
19. Weibel, R., Dutton, G.: Generalising Spatial Data and Dealing with Multiple Representations. Geographical Information Systems, Vol. 1, Principles and Technical Issues. (1999) 125-150
20. Wertheimer, M.: Laws of organization in perceptual forms. First published as Untersuchungen zur Lehre von der Gestalt II, in Psycologische Forschung, 4, 301-350. Translation published in Ellis, W. (1938). A source book of Gestalt psychology (1923) 71-88