AUTOMATING GENERALIZATION – TOOLS AND MODELS

Dan Lee and Paul Hardy

ESRI, Inc.
380 New York St., Redlands CA 92373, USA
Telephone: (909) 793-2853; Fax: (909) 793-5953
E-mail: dlee@esri.com; phardy@esri.com

ABSTRACT

Many national mapping agencies (NMAs) are pursuing the idea of building a master database and deriving multiple-scale products from it. To support this production goal, GIS-based generalization is a necessity. The solution for generalization involves data modeling, process automation, multiple representations, updating, and more. This paper focuses on the automation of generalization processes in ArcGIS (the GIS software created by ESRI, Inc.).

To automate generalization requires translating the cartographer's knowledge into computer logic and algorithms in order to derive desired results. Our starting point is the Generalization toolset in ArcToolbox, the powerful geoprocessing framework containing hundreds of data analysis and management tools and a ModelBuilder for process chaining. Existing and forthcoming tools along with on-going research cases will be used to illustrate the automation challenges, such as defining rules, recognizing certain patterns and contexts, and producing topologically correct output with feedback for evaluation and post-processing. Sample generalization models will also be presented.

1   INTRODUCTION

It is a strategic aim of many national mapping agencies (NMAs) to build a high resolution, high accuracy digital landscape model (DLM), from which new DLMs at reduced scales are to be derived. The digital cartographic models (DCM) and target cartographic products are then compiled from the corresponding scale-band DLM. At the heart of such a production strategy is generalization – the intelligent abstraction of data to a smaller scale. One of the essential tasks of implementing a GIS-based solution for generalization is the automation of generalization processes that used to be done manually by experienced cartographers.

With the experience of developing generalization commands for the coverage data model [ESRI, 2000] our first step towards the integration of generalization capability in ArcGIS was to start and grow a set of core generalization tools for geodatabase features. These tools reside in the Generalization Toolset of the Data Management Toolbox in ArcToolbox. Some have been released with ArcGIS 9.0 or under development for upcoming ArcGIS releases, as indicated in Figure 1; others are in prototype.
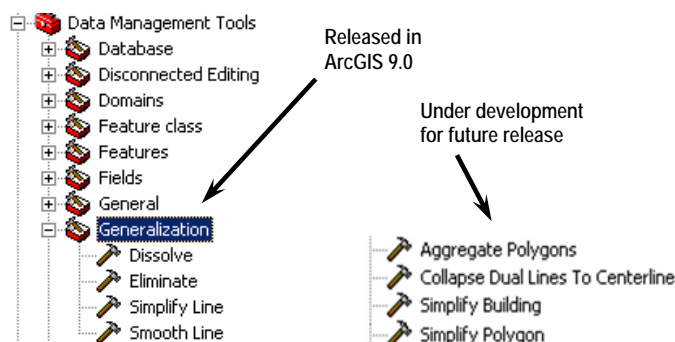

Figure 1:   Geodatabase feature generalization tools

The geoprocessing framework allows you to execute any tool through one of four ways: <u>a tool dialog</u> invoked from the tool icon, <u>a command line</u> entered in the Command window, <u>a model</u> containing chained tools, or <u>a script</u> for more complex processes. See a previous paper [Lee, 2003] for more details. Currently each generalization tool is designed to take a certain type of input features, process a particular operation, such as simplification or aggregation, and store the generalized output.

Generalization in traditional mapping relies on a cartographer's judgment and skills. Developing automated generalization tools is a process of reverse-engineering, meaning to be able to deduce human analysis and decisions from existing maps and general specifications, define explicit rules, compose logical sequences of steps, develop computational techniques to achieve desired results in each step, preserve geographic characteristics and spatial relationships, and provide feedback for evaluation and post-processing. The following sections present what it takes to develop some of the existing and forthcoming generalization tools.

2   DEFINING GENERALIZATION RULES

Very few manual generalization guidelines exist in textbooks. NMAs may have specifications, but they are usually too general and incomplete to support automation. It is critical to be able to clearly define generalization rules that can be coded in computer language and lead to acceptable result. These rules should precisely describe the measurable conditions where certain generalization is needed, the various forms or options for a shape or pattern to be generalized and the prioritized order of considering these options, the extents to which problems are solved, the exceptions and limitations, and so on.

2.1   Rules for building simplification

In large-scale databases buildings are commonly represented individually as detailed footprint polygons. Within the relatively large-scale range, they should remain individual polygons, but with less and less details. Some example instructions for building footprint simplification state: "The measured area of the simplified outline should remain roughly the same as the area of the original", "General form should be maintained", and "If possible, draw rectangles" [Swiss Society of Cartography, 1987]. To implement the Simplify Building tool, detailed rules have been defined. Here are some example rules, followed by illustrations of simplified individual buildings and building groups.

- A building must be simplified if it contains one or more sides shorter than a specified length.
- The orthogonal shape should be preserved or enhanced, that is, making near-90-degree corners exactly 90 degrees.
- A building can be simplified by filling up small corners, close off or widening isolated small intrusions or extrusions, or by straightening or reducing various stair patterns (see discussion in the next section), while keeping the measured area gain and loss roughly balanced.
- A building that does not retain a specified minimum area will be excluded unless the user chooses to keep it.
- Under relatively large reduction, a building can be turned into a rectangle taking the shape of the bounding box oriented along the longest side and an area equal to the original.
- For a group of adjacent buildings connected by near-parallel boundaries (walls), only the outer boundary of the group is simplified and the buildings should remain connected after simplification.
- For a group of adjacent buildings connected in more complicated ways, they will not be simplified, but flagged in the output. This is a limitation and will remain for future research.
- Grouped buildings will be assigned a unique group id. Simplification status should be flagged including, for example, properly simplified, grouped and simplified, grouped and not simplified, too small but not excluded, potentially in conflicts, and so on. This information will facilitate quality assessment and post-editing.
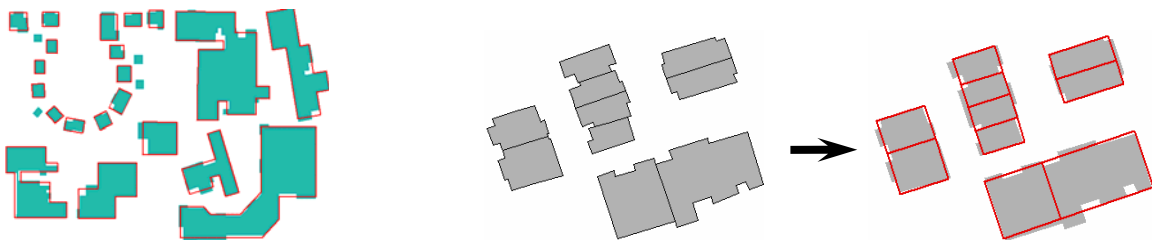

Figure 2:   Simplified individual buildings (left) and building groups (two on right)

2.2   Rules for collapse of dual lines (road casings) to centerlines

The task of creating road centerlines from dual-line casings is an important generalization operation for large-scale to medium-scale mapping. It may seem not very difficult to calculate a line centered between a pair of parallel lines. Some researchers have explored different techniques to derive centerlines, or in a more general sense the skeleton or axis of features [Christensen, 1996]. However making a complete solution (a tool) for this task goes beyond where centerlines need to be created. The more difficult part is to deal with intersections and data stored in various ways in the input. It is not necessarily acceptable that all road casings become centerlines in the output; some large intersection areas are actually town squares or plazas where simply connecting all incoming lines to a central junction point may not be the right representation. It's even harder to be able to recognize and process interchange roads with overpasses, ramps, etc. accordingly when they are all mixed in the input and may or may not be classified. Here are examples rules that guide the implementation of the Collapse Dual Lines To Centerline tool, followed with sample results shown in Figure 3.

- The tool is intended to create centerlines for open-ended, generally parallel road casings within the specified range of width.
- No lines should be created outside of the casings or intersecting the casings.
- At relatively smaller and simpler intersections, the two lines closest to be on a straight line (or collinear) should be connected at a junction point first; the remaining lines may connect to the same junction point if the connecting angle is acceptable, or projected to the already connected line if the projected angle is not extremely sharp to cause the intersecting point too far away from the intersection; otherwise the lines may need to be reshaped (or bent) to form acceptable connections.
- Normally one junction point in each intersection area should be created. For wider and more complex intersections, separate junction points may be considered.
- If an intersection area is relatively large, the incoming centerlines should stop before entering the area and an outline of the area will be created, connected properly to all the incoming centerlines, and flagged.
- For cul-de-sac, the centerline should end at a near-centroid location of the ending shape, usually rounded.
- Where the dual lines form a continuous network, the resulting centerlines and intersection outlines together should represent the same continuous network. Any unused lines (single lines or open lines) should be included in the output and flagged.

Figure 3:   Examples of resulting centerlines with intersections and cul-de-sac

In both subsections above, further rules and additional details for building simplification and collapse of dual-lines (road casings) to centerlines are not described due to the limited length of the paper. The rules are being enhanced and perfected over time in order to continuously improve the output quality. The closer these guidelines are to the cartographer's thinking, the more successful the automation will be.

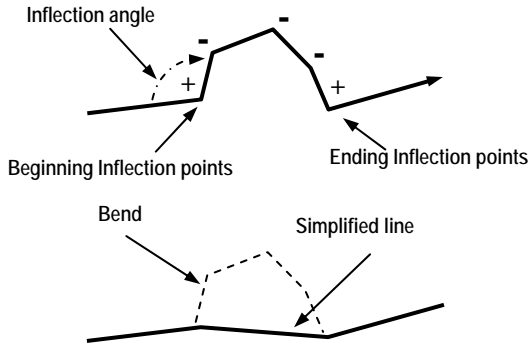## 3   CONSIDERING GEOGRAPHIC CONTEXTS

The challenge in developing generalization solutions roots from the complexity of generalization tasks itself – no features should be generalized in isolation. A previous paper [Lee, 2004] discussed that for model generalization it is important to consider geographic contexts and that geographical patterns and feature spatial relationships are the main considerations of the geographic context. Geographic patterns can be at a detailed level, such as in a feature's shape, or much wider in scope, such as the alignment of a group of features or a special landscape region like an urban area or a hilly area. Generalization analysis, options, and strategies should be determined accordingly.

3.1   Patterns embedded in feature shapes

Many geographic features are stored in linear and polygonal shapes. Generalization should reduce the level of detail in a geometric shape while preserving the characteristics of the feature. Simplification of building footprints, coastlines, or boundaries is one example that deals with patterns embedded in feature shapes.

3.1.1 Bends in linear features or polygon boundaries

In Simplify Line or Simplify Polygon tools, one of the available algorithms is BEND_SIMPLIFY. This algorithm searches for bend, the basic pattern in linear and boundary shapes, by the inflection points; where the inflection angle changes the sign is the beginning or ending of a bend [Wang, 1996], as shown in Figure 4. Based on several geometrical properties of the bend, as described in the ArcGIS online tool help, a bend is either kept or eliminated, meaning replaced by its baseline (the line connecting the beginning and end points of the bend). The simplification takes place iteratively such that the smaller bends may "disappear" in the early rounds and bigger bends later. The resulting line follows the main shape of the original line quite faithfully and shows good cartographic quality.

1.1.1.1   Figure 4:   Eliminating a bend in line

Figure 5:   Example of polygon simplification by BEND_SIMPLIFY

3.1.2 Patterns in building footprints

In order to reduce the details in building footprints, many patters and their simplification options were defined and implemented for the coverage command BUILDINGINGSIMPLFY [ESRI, 2000]. One of the common patters is stair-like shape. Figure 6 shows the various styles of stairs and their simplified choices.

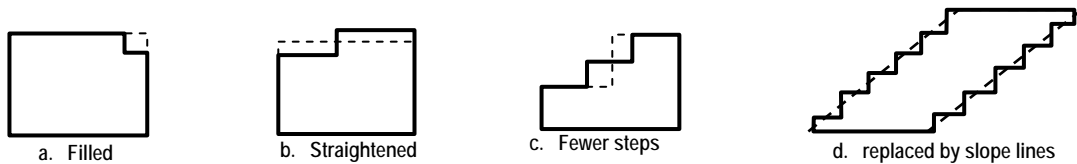a. Filled   b. Straightened   c. Fewer steps   d. replaced by slope lines

Figure 6:   Stair patterns and simplification options

Table 1 gives a summary of the defined patterns and simplification options. More details and illustrations were presented in the previous paper [Wang and Lee, 2000]. The upcoming Simplify Building tool for geodatabase features will adopt and enhance these patterns and simplification options.

Table 1.   Patterns in building footprints and simplification options

PATTERNS		SIMPLIFICATION OPTIONS
Arc	Part of a circle	Eliminate the arc based on span
	Full circle	Delete or keep the circle based on size
Dull corner formed by two long sides		Project two edges
Small spike with two short-sides and their two long neighboring, aligned sides		Remove spike
Stair	Multi-step stair	Long and uniform	Transform to a line along slope
		Others	Reduce the number of steps
	Single-step stair		Fill or equal-area straighten
Low undulate with intrusion and extrusion		Equal-area straighten
High undulate	Single extrusion	Eliminate the extrusion with area control
	Single intrusion	Fill or widen the intrusion
	Multiple-extrusion/intrusion	Equal-area straighten

3.2 Patterns of wider scope

Another major part of geographic characteristics of a mapped area is the wide range of geographic patterns. A geographic pattern can be a unique natural formation (a mountain range) or a cultural phenomenon (an urban or a rural area). A geographic pattern can cover a very large region (a hydrographic watershed or network) or a relatively small area (a residential block or a group of similar features). Such geographic patterns are most often not explicitly defined and stored as features in a database and are difficult to model and to generalize automatically.

3.2.1 Patterns among features

The medium level of patterns are those among features, that is, the way they are positioned to each other, for example aligned in one direction, forming a circular or other regular shape, separated by a similar distance, appearing in certain combination or configuration of features, bounded by other features, contained inside other features, and so on. Figure 7 shows a typical residential area, where some buildings are aligned to a straight road; some follow a smoothly curved road. Also buildings are partitioned by the roads. These patterns and partitions can be used to guide more advanced generalization processes, such as typification, conflict resolution, and ultimately contextual generalization.

Figure 7: Patterns among features

3.2.2 Other geographic patterns

During our continuous investigation and research on generalization requirements and issues, a good variety of generalization specifications that make references to geographic patterns have been found among NMAs mapping guidelines. The following examples illustrate just a few of such specifications.

Spot height selection in terrain context (specifications from ICC, the Institut Cartogràfic de Catalunya, Pla, 1999):
> Example specification 1 – "In <u>mountain passes</u>, always preserve one or more spot height with the first consideration of the lowest ones and the second consideration of the most centered ones". Figure 8-a shows the digital data (Topographic Database at 1:5.000) with a very high density of spot heights and the map of 1:10,000.
> Example specification 2 – "In <u>open area, raised areas, leveled areas</u>, and rustic parcels, consider keeping the most centered ones. Figure 8-b shows the same digital data and the generalized map.

a: Mountain pass spot heights in database (left); selected spot heights on the map (right)

b: Open area spot heights in database (left) ; selected spot height on the map (right)

Figure 8: Spot height selection in terrain context - (thanks to ICC for the data and specifications)

Feature importance and representation in natural or cultural context [NIMA, 1990]:

     Example specification 3 – In <u>arid and undeveloped areas</u>, depict as many drains as possible.

     Example specification 4 – In areas where numerous tanks exist, a <u>representative pattern</u> is used which will retain the <u>general layout</u> of the entire tank area.

The geographic patterns mentioned and underlined in the above specifications may not have a clear boundary on the ground and therefore not collected and stored as geographic features, but they are the keywords in the specifications and set the scope of each particular requirement.

3.2.3   Challenge in automation

It is already not easy for a human cartographer to visually recognize the geographic patterns on a base map and portray them at a reduced scale. Developing a digital solution such that the geographic patterns could be "perceived" automatically is even harder. A very essential task is to be able to recognize the "invisible" spatial extent of a geographic pattern. The extent of a geographic pattern can be seen as a generalization solution space, within which uniquely structured features reside and are related. Certain generalization actions and rules may only apply to features within the extent, and the alteration of feature locations or shapes as in typification or displacement should only consider features in context within the extent and should not propagate to beyond the extent.

In order to find the digital extent of a geographic pattern, such as the "open area" or the "arid area" stated in the above examples, terrain analysis, probably combined with the help of geographic attributes of features, and even interactive decision-making are needed. There are no existing clearly defined guidelines and techniques that could lead to solid implementation; this is one of the areas where more questions may remain than answers. Our research is underway.

3.3   Topological relationships

In developing generalization tools, one of the most important aspects of output quality is the topological relationship among features. There are many topological relationships. This section focuses on a few of them.

3.3.1   Resolving topological errors introduced by simplification

It is clear that topological errors can be introduced by simplification processes. The Simplify Line and Simplify Polygon tools deal with three types of topological errors, line-crossing, coincident lines (lines fallen on top of each other), and collapsed zero-length lines as shown in Figure 9.

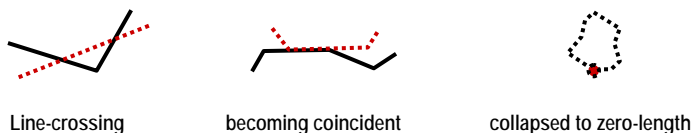Line-crossing   becoming coincident   collapsed to zero-length

Figure 9:   topological errors introduced by simplification

An iterative strategy has been implemented. The input features are first simplified using the specified tolerance. Then, a detection routine will look for these topological errors, and, if any, locate the involved line segments. A reduced tolerance (half of the original) will be applied to re-simplify these segments. This detection and re-simplification with a reduced tolerance (half of the last used) will repeat until no more errors are found. Figure 10 shows a comparison between an input line and its simplified form. The bend where the arrow points at is much smaller than those in the left circle, but can not be removed as those were in the right circle without causing line-crossing; so it was "under-simplified" and kept in the result. The output will contain two new attributes, MaxSimpTol and MinSimpTol, which show the range of tolerances actually used in simplifying each feature [Lee, 2004].
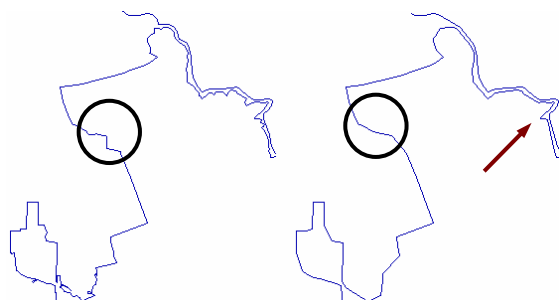

Figure 10:   Before (left) and after (right) simplification: where the arrow points at is obviously less-simplified compared to the shape change in the circled area; it's the result of resolving line-crossing errors. (Thanks to the US Census Bureau for the test data.)

Currently the simplification tools operate on single input feature classes. It is possible to create a model that combines different feature classes into one, applies a simplification tool, and then separates the feature classes later by a common attribute. As the development advances to contextual generalization, multiple feature classes will be handled together.

3.3.2 Ensuring shared geometry

In certain linear features, such as a road network, or in any polygon data, there could be shared geometry representing multiple features, for example multiple bus routes sharing the same road or adjacent counties sharing a boundary. In Simplify Line, Simplify Polygon, and Simplify Building tools, shared geometry is identified through the use of geoprocessing functions that access the underlying topology engine of ArcGIS. Shared geometry in lines and polygons will be simplified consistently and the routes and boundaries properly preserved. An example of simplified polygons with shared boundaries can be seen in Figure 5.
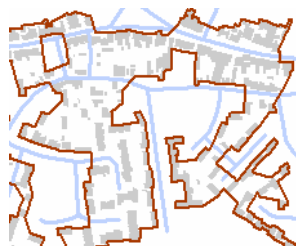
Shared geometry may also exist between different feature classes, as shown in Figure 11 where buildings are inside residential blocks with shared edges. It is necessary to take into account of both feature classes and ensure the same simplified shape. This will be addressed in future development.
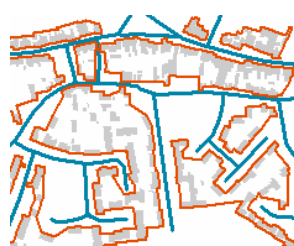

Figure 11: Shared edges between buildings and residential blocks (Thanks to ICC for the data)

3.3.3 Considering interfering features

Interfering features are constraints to generalization actions, in other words, a certain generalization cannot happen as it would normally do because of the existence of an interfering feature. One typical example is the so called "constrained aggregation, that is, to aggregate features within a given distance, but not to cross another feature. The Aggregate Polygon tool combines features within the specified distance as shown in Figure 12-a, aggregating buildings without roads. The constrained aggregation tool (prototype) aggregates buildings avoiding the roads, as shown in Figure 12-b.

a. Aggregating buildings without roads   b. Aggregating buildings avoiding roads (in prototype)

Figure 12: Aggregation with and without constraints (Thanks to Ordnance Survey for the data)

Another case of interfering features in generalization is where a feature is on one side of a line or boundary, and simplification of the line or boundary should not cause the feature to be on the opposite side, as shown in Figure 13. The existing simplification tools can be enhanced to take into account of the constraint features and preserve the relative positions. A potential approach is to put features in a TIN structure so that they can "see" neighbors easily. If a line goes to the wrong side, inversed triangles would occur, thus indicating the invalid spatial relationship.

Figure 13: Simplification of line (solid to long-dashed) may cause another feature (in the circle) to be on the opposite side

4 CREATING GENERALIZATION MODELS

Generalization processes are not straightforward; to model the process is always a challenge. The ModelBuilder mentioned earlier helps us to experiment with different procedures, adjust the workflow according to different themes and target maps, and make the generalization processes easy to manage. You can create and edit a model diagram in ModelBuilder to put the generalization steps in a desired sequence. The diagram can be saved as a model in a user-specified toolbox and modified easily to repeat the same or similar processes for different datasets or for the same data with different parameters and options.

4.1 Sample model 1 – contour generalization:

The model diagram in Figure 14 illustrates a hypothetical scenario of contour generalization sequence. The goal is relatively simple:

- Select contours of 50-meter interval.
- Connect "broken" segments so that longer lines are formed for better result.
- Simplify the contours with the Resolve Topological Errors option.
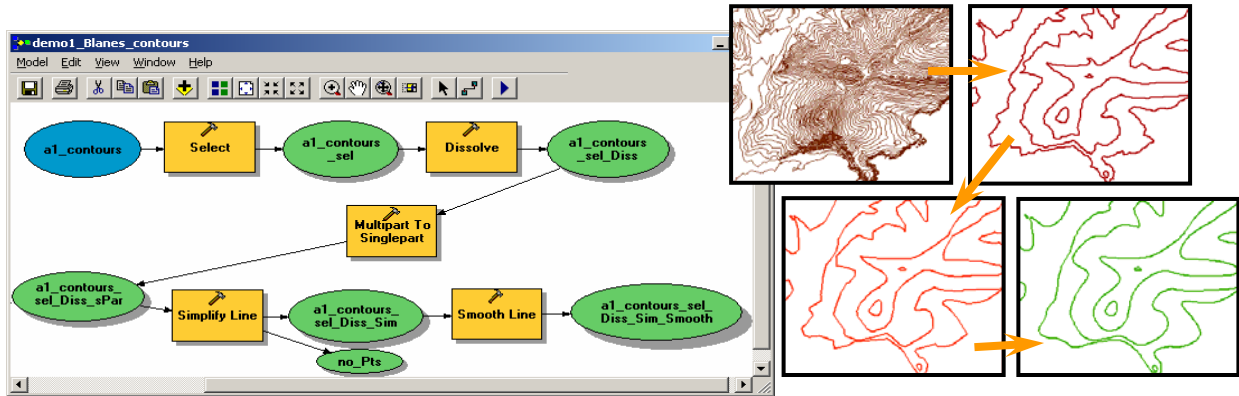- Smooth the contours for aesthetical quality.

Figure 14: Contour generalization model and data in the process (input contours, selected contours, simplified, and smoothed contours)

4.2 Sample model 2 - Building generalization in urban and rural areas

The model diagram in Figure 15 illustrates a hypothetical sequence of building generalization. The idea is to generalize urban buildings and rural buildings separately with different parameters and operations. The main steps are:

- Aggregate neighboring street blocks to obtain urban areas.
- Find urban buildings inside the urban areas by overlapping all input buildings with the urban areas.
- Simplify the urban buildings with 5-meter tolerance and exclude buildings smaller than 25 square meters.
- Find buildings outside the urban areas as the rural buildings.
- Simplify the rural buildings that are larger than 100 square meters by 10-meter tolerance.
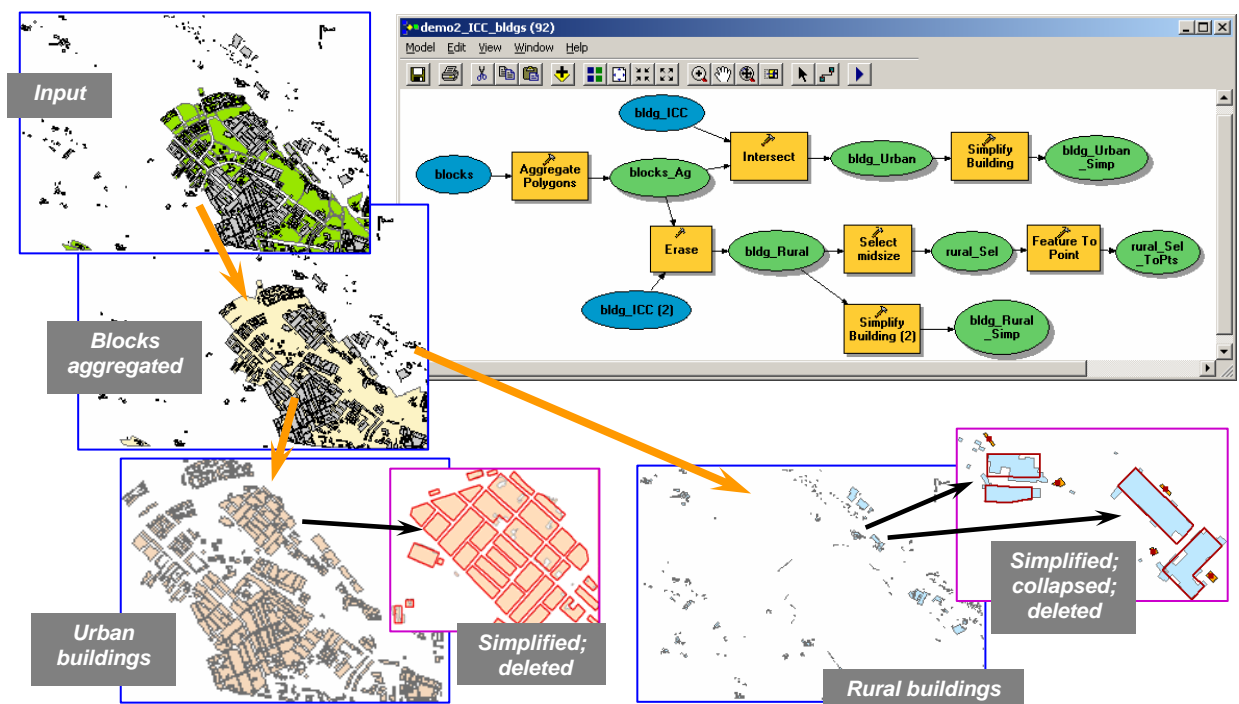- Collapse the rural buildings smaller than 100 square meters to points.

Figure 15: Building generalization model and results – urban and rural buildings are generalized differently (thanks to ICC for the data)

5   FUTURE DIRECTIONS

The implementation of automating generalization depends on the successful translation of human knowledge from manual generalization into explicit rules and logic that guide the development of computational approaches, tool design, and model creation. ESRI plans that the existing and forthcoming generalization tools that operate on single input feature classes will be extended to handle multiple feature classes and feature types for contextual generalization; that is, generalizing features that are related and interfere with each other. These tools will then be embedded in geoprocessing models that understand the iterative and adaptive work flows of generalization.

The contextual generalization solution will exercise the following concepts [Hardy and Lee, 2005]:
- partitioning mapped areas into geographic zones based on cultural patterns, terrain patterns, distribution and neighborhood patterns; only when these higher level entities have been recognized and constructed, will the appropriate generalization process be modeled and underlying generalization tools be determined and applied to features, in order to derive the appropriate abstracted forms.
- creating an adaptive system that automatically derives scale-based settings for generalization operators, parameters, priorities, and other preferences that govern the generalization processes and decisions, and allows the freedom to adjust these settings to optimize the output and to suit for multiple products.
- facilitating quality assessment, post-editing, and representation refinement.
- logging generalization procedures, maintaining links between source features and generalized features, and tracking and propagating changes for product updates.

In the meantime, model generalization (involving true-to ground data without symbolization) will feed into cartographic generalization, which operates on features in fully symbolized contexts [Lee, 2004], and multiple representations, to support the DLM to DCM strategy. Cartographic generalization and representation tools are under development in both geoprocessing and editing environments for the upcoming releases of ArcGIS as part of the ESRI database cartography solution [Hardy and Kressmann, 2005].

6   CONCLUSIONS

- Many NMAs and commercial map publishers have expressed a need for generalization in order to derive multiple products from a master database. ESRI has a set of generalization tools targeted at this requirement, and is extending the set in future releases of ArcGIS to meet these needs.
- The ArcGIS geoprocessing framework and the ModelBuilder provide the appropriate framework for such bulk intelligent generalization, and ongoing development is focused on establishing and using spatial context during generalization.
- Recent work has implemented intelligent algorithms that detect and use topological and spatial patterns, and these will be combined with partitioning and adaptive processing models to create efficient work flows for automated contextual generalization.

NOTES

1. This paper is a forward-looking document, and some of the capabilities it describes are still under development. As such, it is intended to give guidance as to likely future direction and should not be interpreted as a commitment by ESRI to provide precise capabilities in specific releases.

REFERENCES

Christensen, Albert H., 1996, "Street Centerlines by a Fully Automated Medial-Axis Transformation", Proceedings of GIS/LIS 96 Conference, Denver, CO, p.107-115.

ESRI, 2000, "Map Generalization in GIS: Practical Solutions with Workstation ArcInfo Software", http://downloads.esri.com/support/whitepapers/ao_/Map_Generalization.pdf.

Lee, Dan, 2003, "Generalization within a Geoprocessing Framework", GEOPRO conference proceedings, Mexico City, p.82-91.

Lee, Dan, 2004, "Geographic and Cartographic Contexts in Generalization", ICA Workshop on Generalization and Multiple Representation, Leicester, UK, http://ica.ign.fr/Leicester/paper/Lee-v2-ICAWorkshop.pdf.

NIMA, 1990, "Military Specifications – 1:100,000 Scale Topographic Maps", Mil-T-89306 (DMA).

Hardy, Paul, & Lee, Dan, 2005, "Multiple Representations with Overrides, and Their Relationship to DLM/DCM Generalization", to be presented at the ICA Workshop on Generalisation and Multiple Representation, 7-8 July 2005, A Coruña, Spain.

Hardy, Paul, & Kressmann, Thierry, 2005, "Cartography, Database and GIS: Not Enemies, but Allies!", to be included in the 22nd ICA Conference Proceedings, A Coruña, Spain.

Pla, Maria, Dec. 17, 1999, "Spot Height Selection", Automated Cartography, Institut Cartogràfic de Catalunya.

Swiss Society of Cartography, 1987, *Cartographic Generalization – Topographic Maps*, Cartographic Publication Series, No.2 (Zurich: Swiss Society of Cartography, 1987), p.22.

Wang, Zeshen, 1996, "Manual versus Automated Line Generalization", GIS/LIS '96 Proceedings, p.94-106.

Wang, Zeshen, and Lee, Dan, 2000, "Building Simplification Based on Pattern Recognition and Shape Analysis", Proceedings of the 9th International Symposium on Spatial Data Handling, Beijing, China, p.58-72.

BIOGRAPHY

Mrs. Dan Lee has been a Cartographic Product Researcher and Specialist in the Software Development Department at ESRI, Inc. since 1995, heading the research, design, prototype, and implementation of map generalization solutions. She has recently joined the Multiple Representation project for advanced cartography.

She was a Cartographic Systems Consultant for four and a half years in the Mapping Division at Intergraph, defining and marketing generalization and other mapping products.

She has been a corresponding member (from the U.S.) and actively involved in the ICA Map Generalization and Multiple Representation Commission, previously the ICA Map Generalization Commission, and before that the Map Generalization Working Group, since 1992; a member of the ICA Commission on Incremental Updating and Versioning since 2001, and a member of the American Congress on Surveying and Mapping since 1989.

Mrs. Lee holds a BS degree in Physical Geography from Peking University in China, an MA degree in Geography/Cartography from Syracuse University in the U.S., and an MB degree in Geodetic Science and Surveying from Ohio State University in the U.S.