Scalability of Contextual Generalization Processing using Partitioning and Parallelization

Marc-Olivier Briat, Jean-Luc Monnot, Edith Punt

(processing large seamless datasets)
Partitioning
Handling large volume of data

- At ArcGIS 10.0, contextual generalization tools are limited to a map sheet worth of data
  - 100,000 features
- Large seamless datasets are commonly available and need to be generalized
- Workflows to overcome those limits are complex and require additional database management steps. Sometimes the tools are simply not used.
Handling large volume of data

- A natural approach is to consider partitioning the dataset spatial extent. Each partition is defined as a polygon feature and isolates a subset of data to process.

- Partitions should:
  - Provide control over the volume of data
  - Be available for all tools used in the workflow
  - Not have any impact on the result
Handling large volume of data

- We want partitions to be freely defined by the user
Dealing with boundaries

- **Two main goals**
  - Provide seamless processing
  - Avoid post processing of boundaries

- **Contextual tools**
  - Cannot arbitrarily stop at boundary
  - Need surrounding features, but up to what extent?
Contextual tools

- Can we predict what extent around a partition has an impact on processing the content of the partition?
Contextual tools

- For most of our tools, we can derive this maximum area of influence
  - Aggregation distance (aggregate polygons)
  - Merge distance (merge divided roads)
  - Minimum length (thin road network)
  - Symbol width (resolve road conflicts)
  - Etc.
Adding a buffer

- Contextual aspect addressed by buffer
- Load all features inside the buffer
- Modify only features inside the partition
Thin Road Network
Thin Road Network

- Buffer value
  - Notion of how much a feature contributes to the network using its position inside multiple itineraries
  - Itineraries need to start at least from ‘Minimum Length’ outside the partition
  - Buffer = 1.5 x Minimum Length
- Features processed by one partition are considered “locked” for adjacent partitions
Thin Road Network

- Entire streets network from California
  - 2,860,000 features
  - 157 partitions
  - 15,000 features overlapping boundaries
  - 75 visibility mismatch
Resolve Road Conflicts
Resolve Road Conflicts

- **Buffer**
  - This tool resolves symbol overlaps
  - Distance is given by symbol width
  - Buffer = 10 x symbol width
  - Modifications extend outside the partition
Contextual tools that would not work

- Extent is not predictable
  - Distance of influence is supported by features
  - Case for the Propagate Displacement tool

```
DISP_X / DISP_Y
```
Contextual tools that would not work

- Features identify a larger structure
  - Lines forming a closed polygon
  - Case for the Propagate Displacement tool
Controlling the buffer value

- Large buffer values
  - Impact the volume of data to load
  - Create additional neighbor partitions

- Worst case in our California test was +20% for the Thin Road Network tool (x10 scale jump)

- Favors a ladder approach (vs star)
Parallel Processing
Goals

- Prototype work
  - No plan to release this functionality
  - Experiment and learn

- Validate
  - This partitioning approach is suitable for parallel processing
  - No impact on workflow aspects

- Make our testing framework more efficient
Database centric

- Concurrent access to data (input + partitions)
- The database synchronizes the work
  - Using locks on datasets
  - Processes wait for dataset availability
- Allows multiple clients
  - On same machine
  - On remote machines
Prototype

- Prototype uses a file geodatabase

- Setup requires
  - Defining a shared folder
  - Adding an exe into ArcGIS/bin
  - Enable parallel processing with some registry keys
Transparent for the user

- User runs the geoprocessing tool as usual
  - A task file is added to the shared folder
  - Additional processes are started to work on the same task
Processing partitions

- Locks to assign partitions to processes
Concurrent access to data

- Typical tool execution profile
  - In memory processing takes a lot more time than DB access
  - Makes DB locks acceptable
Concurrent access to data

- Other tools have a more complex pattern
  - Deal with more datasets
  - Have a lower ratio of pure processing compared to processing + DB access

**Merge Divided Roads**
1. Acquire partition
2. Read input (geometry + merge field)
3. Read output (for connectivity to existing)
4. Processing
5. Read all input attributes / insert in output
6. Update input to add QC values
7. Update partition
Concurrent access to data

- Understanding those DB access patterns is important to decide how many parallel processes could work efficiently
- Potential improvements by creating output tables instead of qualifying the input
- Increasing the size of partitions improves the ratio
Adjacent partitions

- Cannot process adjacent partitions simultaneously
  - Seamless database => Features will overlap multiple partitions
  - Some tools have to adapt to existing results (continue the work – example of RRC)

- Plan to prevent this to happen
  - Defined by the partition status

<table>
<thead>
<tr>
<th>partition processing - roads</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;all other values&gt;</td>
</tr>
<tr>
<td>STATUS</td>
</tr>
<tr>
<td>Not Processed</td>
</tr>
<tr>
<td>Being Processed</td>
</tr>
<tr>
<td>Successfully Processed</td>
</tr>
<tr>
<td>Out of Memory; Error</td>
</tr>
</tbody>
</table>
Adjacent partitions

- Process partitions with a lot of unprocessed neighbors first
Adjacent partitions

- Avoids neighbor conflicts when ending processing
### Some Results

- Entire street network for the state of California
  - 2,860,000 features / 157 partitions
  - 50,000 features max per partitions

<table>
<thead>
<tr>
<th>Tool</th>
<th>PC</th>
<th>With PP</th>
<th>Without PP</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thin Road Network</td>
<td>4 core 4 processes</td>
<td>3h30min</td>
<td>13h30min</td>
<td>3.85</td>
</tr>
<tr>
<td>Check Network Connectivity</td>
<td>4 core 4 processes</td>
<td>6min30s</td>
<td>24min</td>
<td>3.7</td>
</tr>
<tr>
<td>Thin Road Network</td>
<td>4 core HT 8 processes</td>
<td>2h30min</td>
<td>10h30min</td>
<td>4.2</td>
</tr>
<tr>
<td>Merge Divided Roads</td>
<td>4 core HT 4 processes</td>
<td>45min</td>
<td>2h45min</td>
<td>3.7</td>
</tr>
<tr>
<td>Resolve Road Conflicts</td>
<td>4 core HT 8 processes</td>
<td>3h35min</td>
<td>12h30min</td>
<td>3.5</td>
</tr>
</tbody>
</table>
Future work

- Make more tools work with partitions
- Continue improving results quality
- Adapt prototype to new pieces of technology
  - Cloud computing
  - Geoprocessing services