Conflation in Geoprocessing Framework – Case Studies

Recent development and looking ahead

Dan Lee, Weiping Yang, and Nobbir Ahmed
Geoprocessing
Esri, Inc
Redlands, USA
e-mail: dlee@esri.com; wyang@esri.com; nahmed@esri.com

Abstract—**Multiple sources of geographic information systems (GIS) data have been more easily and frequently produced and updated than ever. Yet, the traditional problem of data inconsistency spatially and in attribution, as the result of different ways of collecting and modeling data over time, remains obstacles in using the data for analysis and mapping. Efficient tools for data conflation have become a necessity for GIS users. Our recent work on conflation tools for the 10.2.1 desktop release of ArcGIS (the commercial GIS software by Esri Inc.) focused on linear feature matching techniques for identifying matching and no-match features. The initial results have proven time-saving in reconciling data for better positional and attribute quality and harmonization. Future challenges lay in formalizing data preparation, handling other feature types, and optimizing feature matching and workflows.**

Keywords-conflation; geoprocessing, feature matching; change detection; spatial adjustment; attribute transfer; workflow.

I. INTRODUCTION

GIS data maintained by many organizations and government agencies or obtained from data providers often need to be used together for multiple purposes of analysis and mapping. However, you may find that when displaying spatially overlapping or adjacent data, features representing the same ground locations or objects don't line up even in the same map projection; or that a spatially up-to-date data lacks the desired attributes that only exist in another data source. Conflation is the process of matching corresponding features and making spatial adjustment or attribute transfer between them to improve data quality and consistency.

Within the geoprocessing framework, six new tools, shown in Fig. 1, have been developed for the 10.2.1 desktop release of ArcGIS. This development is an advance from the legacy technology used in Esri's earlier product [1].

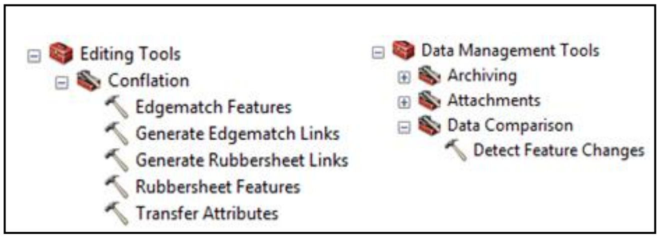


Figure 1. Conflation tools (by the tool icon) inside Editing and Data Management toolboxes in ArcGIS.

Good conflation outcome is achievable as a result of high feature matching accuracy; inspection and editing may be necessary as part of the workflows. The automation and the significantly reduced manual work enable GIS users to move away from living with imperfect data and to reach higher standards in geographic data integration, analysis, and mapping more efficiently.

This paper is organized as follows: Section II briefly reviews the feature matching processes and associated tools. Our initial efforts have focused on linear features. Section III presents a few conflation scenarios and workflows used to accomplish the tasks with success and efficiency. Section IV gives conclusions and thoughts on future work.

II. FEATURE MATCHING IN CONFLATION TOOLS

At the core of conflation is feature matching, either between overlapping datasets or between adjacent datasets. Feature matching accuracy relies highly on data quality, similarity, and complexity. The feature matching techniques used in the conflation tools are briefly described below.

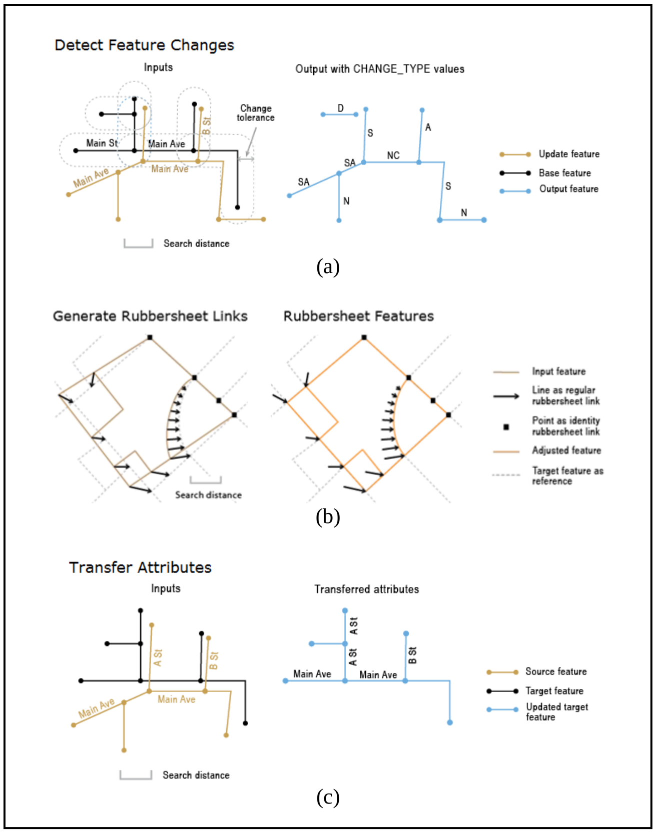
A. *Feature matching of overlapping datasets*

There have been many research papers and implementations on feature matching of overlapping datasets. Some examples include: a five-step statistical approach with the use of a merit function to compute unique combinations of matching pairs among potential but ambiguous matching pairs [2]; a delimited stroke oriented algorithm consisting of four processes for the matching of road networks [3], and an optimization model for linear feature matching which takes into account all potential matched pairs simultaneously by maximizing the total similarity of all matched features [4].

The feature matching technique we choose to use is based on the fundamental analysis of the topological structures and feature pattern recognition. The key processes are: (1) analyzing feature topology, i.e., to find nodes and joining lines in linear features, (2) building structures (paths and patterns), (3) matching structures, and (4) matching features within structures. More details on this feature matching approach are given in a separate paper [5].

The matching information can be written out to a match table with five fields: SRC_FID (source feature ID), TGT_FID (target feature ID), FM_GRP (feature match group ID), FM_MN (matching relationship in the form of m:n, where m and n represent the numbers of source features and target features respectively in a match group and can be greater or equal to 1), and FM_CONF (feature matching confidence level with values between 0 and 100). This feature matching process is the basis for the following tools which help perform conflation tasks on overlapping datasets that cover the same geographic areas:

- Detect Feature Changes (DFC) identifies spatial and attributes changes between update and base features. The output change types include: S for spatial change, A for attribute change, SA for spatial and attribute changes, NC for no change, N for new update feature, and D for potentially to-be-deleted base feature. See the illustration in Fig. 2–(a).
- Generate Rubbersheet Links (GRL) generates rubbersheet links, including regular links (lines going from matching source to target locations) and identity links (points where source and target locations are identical and not being moved in rubbersheeting adjustment). The tool Rubbersheet Features (RF) does rubbersheeting adjustment using the generated links to align source features with target. See the illustration in Fig. 2–(b).
- Transfer Attributes (TA) transfers feature attributes from source to matching target features. See the illustration in Fig. 2–(c). By design when multiple source features match one or more target features, attributes from the first picked source feature are transferred to all matched target features.


Detect Feature Changes

(a)

Generate Rubbersheet Links

Rubbersheet Features
Search distance
Input feature
Line as regular rubbersheet link
Point as identity rubbersheet link
Adjusted feature
Target feature as reference
(b)

Transfer Attributes

(c)

Figure 2. Feature matching based tools for conflation of overlapping datasets.

B. Edgematching - matching features of adjacent data areas

Edgematching is the process of identifying corresponding features along the edge (meeting locations) of adjacent (side-by-side) datasets. The key processes are: (1) finding features within the specified search distance to each other along their meeting areas, (2) evaluating the geometric characteristics and continuity from input to adjacent features and vice versa, and (3) determining the best fit pairs of corresponding features. The tools for edgematching are:

- Generate Edgematch Links (GEL), which generates edgematch links, followed by Edgematch Features (EF) to adjust features to new connecting locations guided by the links. See the illustrations in Fig. 3.

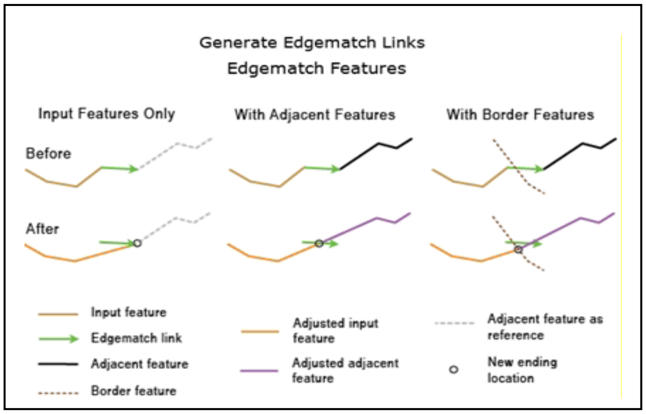

Generate Edgematch Links
Edgematch Features

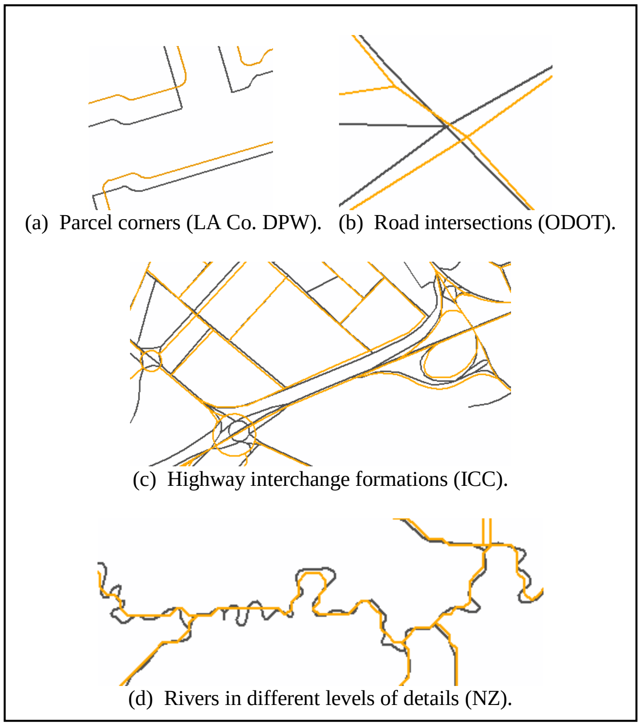
Figure 3. Edgematching by moving endpoints (one of the available options) of features to new connecting locations.

C. Challenges in feature matching

No matter how sophisticated the feature matching techniques are, the reality of geographic data is often more challenging than the automatic analysis can handle. The main factors causing difficulties and errors in feature matching include:

- Invalid feature topology, such as gaps, overshoots, undershoots, overlaps, and duplicates.
- Differences in feature representations and data modeling of the same ground objects, especially between overlapping data sources, ranging from variations in their geometric characteristics to distinctions in their structural formations. For example: round vs. squared corners of parcel boundary lines in Fig. 4–(a), one vs. separate road intersections in Fig. 4–(b), and different collections of road merging or splitting around complex highway interchange areas in Fig. 4–(c).
- Features with different levels of details for multiple map scales. See the illustrations in Fig. 4–(c) and (d).

The more dissimilar the corresponding features the harder to make the right feature matching decisions. The conflation tools can take into account common attributes between input datasets to help determine the right match, but the common attributes are often either unavailable or incomplete.


(a) Parcel corners (LA Co. DPW). (b) Road intersections (ODOT).

(c) Highway interchange formations (ICC).

(d) Rivers in different levels of details (NZ).

Figure 4. Examples of dissimilar overlapping features.

Given the highly automated conflation tools and the possibility of some mistakes in the results, the question we need to address is what it will take to find and correct errors and to complete the conflation tasks. That leads to the discussion on conflation workflows. Due to the length limitation of the paper, the discussion focuses on conflation of overlapping data sources.

III. CONFLATION WORKFLOWS

Conflation tasks may be as simple as to make spatial adjustment or attribute transfer from one data source to another for better positional accuracy and attribute consistency, or as comprehensive as to unify information from multiple data sources for the best combined result. In general a conflation workflow may consist of three components: (1) preprocessing to eliminate input data issues and to exclude irrelevant features from participating in the conflation process; (2) automated processes using the conflation tools to produce mostly correct results and conflation evaluation tools to derive information that help identify potential mismatched features and find locations that need attention; and (3) interactive review and editing based on the automatically derived information, as well as visual inspection, to improve the result to satisfaction.

A few workflows are examined below using real world data. But before getting into that, it is necessary to briefly explain the preprocessing and conflation evaluation tools mentioned above.

Preprocessing is common to all conflation tasks. A few generic guidelines and possible geoprocessing tools to use are given below and are not repeated for every workflow scenario:

- Fix invalid geometry (Repair Geometry tool)
- Validate feature topology (Topology Tools)
- Remove overshoots and undershoots (Trim Line and Extend Line tools)
- Delete unwanted duplicates (Delete Identical tool)
- Break unintended long-running features at intersections (Feature To Line tool)
- Exclude irrelevant features from participating in conflation processes. (Select By Attributes or Select By Location)

Other data specific preprocessing may also be necessary; it is important to identify data issues and use appropriate tools to resolve them.

The conflation evaluation tools mentioned in workflow component (2) above have been built either by Python scripting or by chaining together existing geoprocessing tools. They produce information to help understand the conflation results, identify potential errors, and facilitate the interactive review and editing processes. They are supplementary tools and do not come with the release. Here are the main evaluation tools:

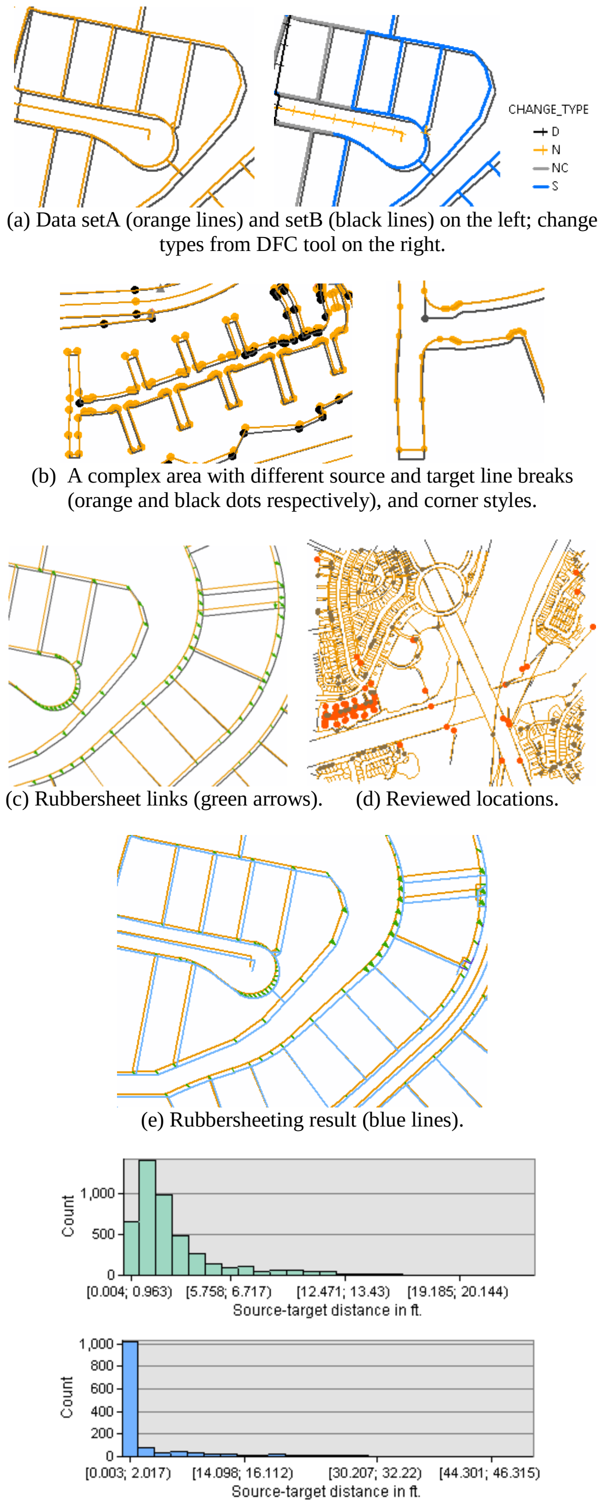
- Check Feature Matching (CFM) – analyzes the feature matching information produced by DFC, GRL, or TA tools and flags questionable matched conditions, for instance, the multiple source or target features in a m:n match group don’t belong to the same line or the matched source and target features may be too far apart to be the right match.
- DFC and Evaluation – runs DFC tool and checks for potential change type errors caused by mismatches; it is especially helpful to verify change types D and N so their source features can be excluded from participating in GRL and TA processes as needed. The tool also makes a bar graph for change types.
- GRL and Evaluation – runs the GRL tool and produces point features at locations where the generate rubbersheet links intersect or where no links are generated. It also adds source (from-point) and target (to-point) vertex types to the links. The vertex types are simply: 0 for in-line vertex, 1 for dangle end, 2 for pseudo node, 3 for T-node, 4 for cross-node, 5 for node with 5 joining lines, and so on. This information is intended to facilitate the inspection, especially at major intersections, for example if a link starts at type 4 and ends at 1, it may not be linking corresponding locations.
- RF and Assessment – It runs RF tool to perform rubbersheeting adjustment and produces additional data and information to help compare the source data and its adjusted result and to assess the location accuracy improvement.
- TA and Evaluation – It runs TA tool and produces additional data and information to help inspect no transfer and potentially mis-transferred cases.

Using real world test data, two conflation scenarios and workflows were examined: A. rubbersheeting spatial adjustment workflow; B. a more comprehensive workflow requiring both spatial and attribute unification.

A. Rubbersheeting spatial adjustment workflow

The data used to demonstrate this workflow are two sets of parcel lines (provided by LA Co. DPW); let's name them setA (3779 lines) and setB (3840 lines) as shown in the left image of Fig. 5-(a); preprocessing details are omitted here. The goal was to spatially adjust setA towards the more accurate setB. The workflow steps, actions, and results are:

- Step 1: Ran the DFC and Evaluation tool – see change types in the right image of Fig. 5-(a). Notice that both setA and setB contain lines that were not parcel lines and didn't have corresponding features. They ended up being N and D change types. Actions were taken to verify them and exclude them from GRL process in Step 2 for better result.
  - Through flagged information and visual inspection, 86% of the Ns (not matched in setA) were confirmed; others corrected. Most of the errors occurred in one large area with not only complex feature shapes but also a big contrast in the number of line breaks (orange dots for source line breaks; black dots for target line breaks) and corner styles, shown in Fig. 5-(b).
  - Through flagged information and visual inspection, 99% of the Ds (not matched in setB) were confirmed; others corrected.
  - Selected 3056 matched lines from setA and 2915 from setB, excluding the verified Ns and Ds respectively, as inputs for Step 2 below.
- Step 2: Ran the GRL and Evaluation tool - total 4413 regular rubbersheet links (see Fig. 5-(c) for a close up) and 0 identity link were generated.
  - Reviewed the flagged no link locations (red dots in Fig. 5-(d), mostly concentrated in the southwest area, i.e., the complex area shown in Fig. 5-(b)), and added 65 critical links.
  - Through flagged intersecting links (brown dots in Fig. 5-(d)) and other hints and inspections, total 104 links were modified and 29 deleted.
  - Analyzed the feature matching result; the estimated accuracy value breakdowns are presented in Table I. A 98.34% high accuracy was reached among matched features, while the overall accuracy 93.84% was largely affected by the no match cases in the complex area.
- Step 3: Ran the RF and Assessment tool – setA was adjusted, as shown in Fig. 5-(e), using total 4449 rubbersheet links. Among many possible ways of measuring positional alignment improvement, the following two are simple and effective:
  - Compared rubbersheet link counts from Step 2 and from rerun of GRL after rubbersheeting: regular link count reduced from 4413 to 1200; identity link count increased from 0 to 3102. This indicates over 70% of source locations are perfectly adjusted to target locations.
  - Compared source to target (source-target) distance distributions through the lengths of the regular links: the distances were obviously more concentrated in the shorter range after rubbersheeting adjustment; see Fig. 6-(f).



(a) Data setA (orange lines) and setB (black lines) on the left; change types from DFC tool on the right.

(b) A complex area with different source and target line breaks (orange and black dots respectively), and corner styles.

(c) Rubbersheet links (green arrows). (d) Reviewed locations.

(e) Rubbersheeting result (blue lines).

(f) Comparison of source-target distance distributions before (upper) and after (lower) rubbersheeting adjustment.

Figure 5. Rubbersheeting workflow (LA Co. DPW parcel data).

TABLE I. ESTIMATED FEATURE MATCHING ACCURACY IN STEP 2

Matched Feature Groups				
Match Relation -ship	***Group Count (Gc)***	***Correct group count (Cgc)***	***Error group count (Egc)***	***Accuracy (percentage of Cgc/Gc)***
1:1	2267	2244	7	98.99%
m:n	384	363	21	94.53%
Total	**2651**	**2607**	**28**	**98.34%**
Unmatched Features				
Match Relation -ship	***Feature count (Fc)***	***Correct feature count (Cfc)***	***Error feature count (Efc)***	***Accuracy (percentage of Cfc/Fc)***
1:0[a]	116	5	111	4.31%
0:1[b]	26	9	17	34.62%
Total	**142**	**14**	**128**	**9.86%**
Grand total	**2793**	**2621**	**156**	**93.84%**

a. Change type N features of DFC output. b. Change type D features in DFC output.

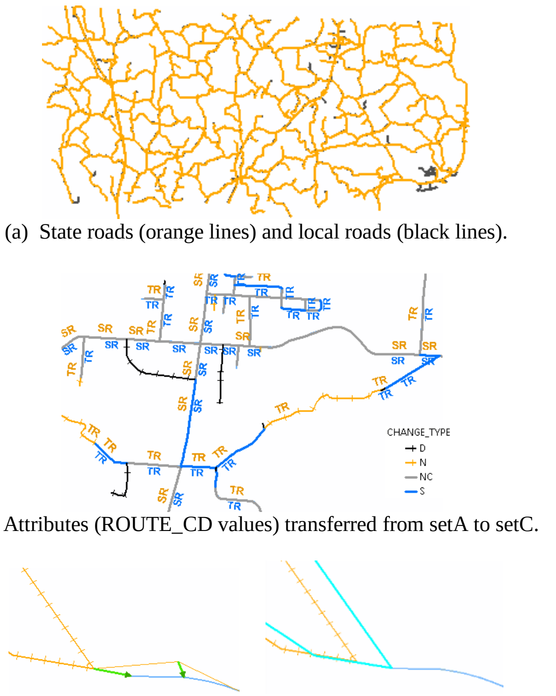
This test data was intentionally chosen for its challenging conditions including the seemingly CAD-imported features with no attributes to separate road centerlines and other features from the parcel lines and the inconsistent data modeling. Also on purpose, only minor preprocessing was done so the strengths and weaknesses of the conflation tools could be tested using near raw data. Although unaccounted errors may exist and would slightly lower the estimated accuracy levels, this exercise produced encouraging result.

B. *Workflow of unifying datasets for the best outcome*

The data used to demonstrate this workflow are two subsets of state and local roads in northeast area of Meigs County (provided by Ohio DOT) as shown in Fig. 6-(a). Let's name them setA (775 lines) and setB (827 lines), knowing that setB is spatially more up-to-date and accurate than setA. The goal was to produce a unified output with the spatial accuracy of setB, the uncommon attributes from both sets for matched features, and all unmatched features of both sets properly positioned keeping their original attributes.

There could be various ways to get there; all would be quite comprehensive. Below is one of the possible workflows attempted in this study. Good preprocessing was done to break lines where necessary, especially for route features in setA, and to improve the topological consistency between the two sets; details are omitted here. The conflation workflow strategy was to: (1) make a setC by copying setB so it has the spatial accuracy and attributes of setB intact, (2) transfer desired attributes from setA to setC for matched features, (3) identify unmatched features in setA, (4) spatially adjust the unmatched features of setA towards setC, and (5) merge the adjusted unmatched features of setA into setC. Here are the details:

- Step 1: Copied setB to setC and ran DFC and Evaluation tool – Actions were taken to verify N and D change types and to exclude them from TA process in Step 2 for better result.
  - Through flagged information and visual inspection, 58 of the 61 Ns (unmatched in setA) were confirmed, 3 corrected; 109 of the 117 Ds (unmatched in setC) were confirmed; 8 corrected.
  - Selected 717 matched lines from setA and 718 from setC, excluding the verified Ns and Ds respectively, as inputs for Step 2 below.
- Step 2: Ran TA and Evaluation tool – see the example result of attribute transfer (ROUT_CD) in Fig. 6-(b), which is superimposed with change types from Step 1. Actions were taken to verify TA result:
  - Reviewed the only 4 no transfer records in setC. Manual transfer was needed.
  - For each of the 39 m:n match cases, attributes from one of the m features were transferred to all n features in the match group by design. Review of the flagged cases and corrections would be needed, if the default transfers were undesired.
  - Analyzed the feature matching result; the estimated accuracy value breakdowns are presented in Table II. A 100% accuracy was reached among matched features, while the slightly lower overall accuracy 98.74% was mainly affected by the few mistakenly unmatched cases.
- Step 3: Ran GRL tool - total 12322 regular rubbersheet links and 19 identity links were generated.



(a) State roads (orange lines) and local roads (black lines).

(b) Attributes (ROUTE_CD values) transferred from setA to setC.

(c) Rubbersheet links & Ns of setA. (d) Adjusted Ns merged in setC.

Figure 6. Unifying multiple data sources for best outcome.

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-326-1

TABLE II. ESTIMATED FEATURE MATCHING ACCURACY IN STEP 2

Matched Feature Groups				
Match Relation -ship	***Group Count (Gc)***	***Correct group count (Cgc)***	***Error group count (Egc)***	***Accuracy (percentage of Cgc/Gc)***
1:1	656	656	0	100%
m:n	39	39	0	100%
Total	**695**	**695**	**0**	**100%**
Unmatched Features				
Match Relation -ship	***Feature count (Fc)***	***Correct feature count (Cfc)***	***Error feature count (Efc)***	***Accuracy (percentage of Cfc/Fc)***
1:0[a]	61	58	3	95.08%
0:1[b]	117	109	8	93.16%
Total	**178**	**167**	**11**	**93.83%**
Grand total	**873**	**862**	**11**	**98.74%**

a. Change type N features of DFC output. b. Change type D features in DFC output.

- Reviewed the links focusing on where the Ns of setA were to be adjusted to connect with features in setC, see the example in Fig. 6-(c). All 58 Ns had links to target locations.

- Step 4: Ran RF tool to adjust the N features using the generated rubbersheet links.
- Step 5: Append the adjusted Ns (highlighted) of setA onto setC (blue lines), as shown in Fig. 6-(d).

This test data was chosen for its clean topology and relatively high similarity. The test results indeed proved a close correlation between high quality starting data and accurate conflation result. Subsequently, very little manual work was needed in this case study.

IV. CONCLUSIONS AND THOUGHTS ON FUTURE WORK

Through the two case studies using real world data, the new conflation tools developed for ArcGIS were successfully tested and produced high quality results. The first case study proved that the very costly and nearly impossible task of generating thousands of rubbersheeting links one-by-one manually could be done mostly automatically with small amount of interactive editing. The second case study gave a good consensus on how the ultimate goal of conflation, i.e., the fusion of multiple source information for the best unified single outcome can be achieved efficiently.

The development of highly automated conflation tools for linear features is a major step forward in supporting the reconciliation of multiple data sources – an important process for data integration and data sharing demanded and embarked on by GIS and mapping agencies [6][7]. Our future efforts will focus on the following areas:

- Enhancements on feature matching with additional pattern recognitions and richer output information.
- New tools for other feature types (point and polygon) and contextual conflation.
- Integrated interactive conflation inspection and finishing environment.
- Streamlining of workflows by testing more real world scenarios and making conflation support tools available at ArcGIS Resources Center: http://resources.arcgis.com/en/home/.
- Investigations on harmonizing spatially related features, such as utility lines and other boundaries spatially associated with parcel lines.
- Extending the use of conflation tools in other areas, such as data quality checking, linking multi-scale geospatial databases and cartographic representations for incremental updating.
- Potential of using image and other information sources in feature matching [8].

ACKNOWLEDGMENT

Thanks to the many Geoprocessing team members for their support and to the following organizations for sharing their data for testing and creating illustrations in this paper:

Parcel datasets, sourced from the Department of Public Works (DPW), Los Angeles County, USA.

Road datasets, sourced from Institut Cartogràfic de Catalunya, Barcelona, Spain.

River datasets, sourced from National Institute for Water and Atmospheric Research (NIWA) and Land Information New Zealand (LINZ). Crown Copyright Reserved.

Meigs county road data, sourced from Ohio State Department of Transportation, USA.

Also thanks to Mark Hollingshead, Esri St. Louis Office, USA, for providing well-prepared Meigs road data.

REFERENCES

[1] A. E. Lupien and W. H. Moreland, “A general approach to map conflation”, Proceedings of AutoCarto 8, March 30 - April 2, 1987, Baltimore, Maryland, USA, pp. 630-639.

[2] V. Walter and D. Fritsch, “Matching spatial data sets: a statistical approach”, International Journal of Geographical Information and science, vol. 3(5, 1999), pp. 445-473.

[3] M. Zhang and L. Meng, “Delimited stroke oriented algorithm – working principle and implementation for the matching of road networks”, Journal of Geographic Information Sciences, vol. 14(1), June, 2008, pp. 44-53.

[4] L. Li and M. Goodchild, “An optimization model for linear feature matching in geographical data conflation”, International Journal of Image and Data Fusion, vol. 2(4), 2011, pp. 309-328.

[5] W. Yang, D. Lee, and N. Ahmed, “Pattern Based Feature Matching for Geospatial Data Conflation”, submitted to GEOProcessing, 2014, Barcelona, Spain.

[6] L. Stanislawski, C. Nelson, and M. Hamann, “Automated Conflation of Reach Data for the National Hydrography Dataset”, http://proceedings.esri.com/library/userconf/proc02/pap1207/p1207.htm, [retrieved: Nov. 26, 2013].

[7] Y. Li and C. Liu, “Spatial approaches for conflating GIS roadway datasets”, Sustainable Transportation Systems, 2012, pp. 290-298, doi:10.1061/9780784412299.0035.

[8] C. Chen, C. Knoblock, and C. Shahabi, “Automatically conflating road vector data with orthoimagery”, GeoInformatica, 10(4), 2006, pp. 495-530.