Improving Cross-border Data Reliability Through Edgematching

Dan Lee, Weiping Yang, Nobbir Ahmed

Esri Inc., 380 New York Street, Redlands, CA, USA

Abstract. The demands on consistent geospatial data for cross-border analysis and mapping are growing at local, national, and global levels. Many national mapping agencies and GIS organizations are facing challenges in keeping data harmonized across internal and external boundaries. Discrepancies and misalignments among neighboring datasets must be resolved to ensure seamless coverage. These tasks are critical to making data reliable for use, especially for collaborative work across borders. Edgematching[1] is the process of matching correspondent features from both sides of borders (or edges), establishing links for them, and connecting them through spatial adjustments. Depending on data quality and complexity, matching correspondent features automatically can be challenging. The more congested the features are, the higher the ambiguity; therefore the possibility of making incorrect matches exists.

The two edgematching tools, Generate Edgematch Links and Edgematch Features, available since ArcGIS Desktop 10.2.1, can be used to automatically generate links between matched features and to perform edgematching spatial adjustment respectively. Additional geoprocessing workflow tools have been built to evaluate the edgematch links and to facilitate post-processing. This paper examines some typical edgematching issues with the focus on linear features, explains the edgematching tools and the workflows, and present test results on two real world use cases to improve cross-border data quality. High accuracy of edgematching can be reached through automated processes and minimal interactive work. Cross-border data quality and reliability can be significantly improved.

Keywords: Edgematching, Cross-border, Data quality and reliability

---

[1] The spellings of edgematching and edgematch in this paper are intentional so they are consistent with ArcGIS documentation.

1. Introduction

The world is more connected than ever. Even though many GIS and mapping organizations or agencies maintain data within the boundaries of their interests or ownerships, the analysis, mapping, or collaborative projects they do may go beyond these boundaries and rely on continuous data over borders. For example, in order to compile state roads from county roads, all county roads must properly meet neighboring county roads at their borders. Misaligned or disjoined roads would cause errors in routing or other spatial analysis and result in poor quality maps. Edgematching may be needed between map sheets or across any natural or manmade boundaries. Neighboring data can be commercially available or obtained from neighbors and other sources, but problems often occur along the borders. Features may not always meet properly; overlaps, gaps, misalignments, incompatible levels of detail, and inconsistent attributes stand as obstacles for data harmonization and utilization. As local data are increasingly pieced together for building seamless coverage at regional and global levels, it is a necessity to ensure cross-border data connections and consistency in order to support reliable spatial analysis and quality mapping.

Cross-border data issues can occur in all geometric feature types: point, line, and polygon. Linear features are typically essential in GIS and mapping databases; our initial research and development, therefore, focused on linear feature use cases and solutions. *Figure 1* illustrates two scenarios: in the left scenario two stream datasets meet with gaps; in the right scenario two road datasets misalign along the border (dashed blue line). The real-world situations can be more complex than these examples, as discussed later.

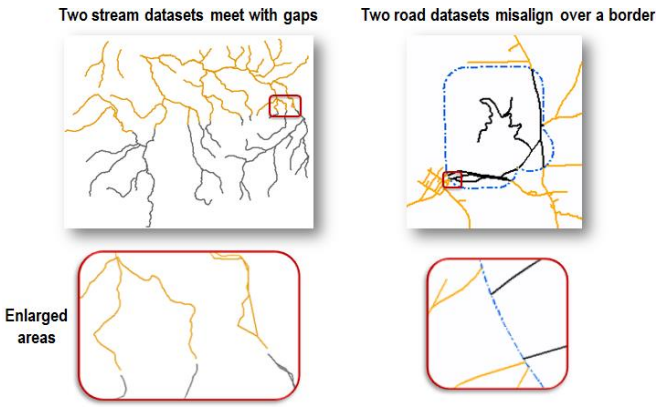

Figure 1. Neighboring linear data issues along their meeting areas.

The solution to cross-border data issues is edgematching. Edge refers to where adjacent data areas meet; it is not necessarily a straight line as traditionally between two map sheets but any shape dividing any types of areas. Administrative, man-made, or natural boundaries are common edges; edges may be explicit (left case of *Figure 1*) or implicit (right case of *Figure 1*). Edgematching is the process of ensuring clean and correct continuation of adjacent datasets at their meeting edges. The correspondent features on both sides of the edge must first be identified, and then their shapes need to be properly adjusted so they are precisely connected. This process used to be done interactively and was time-consuming. To better meet the demands of data harmonization across borders, automated edgematching tools have been developed and added in the 10.2.1 desktop release of ArcGIS (the commercial GIS software by Esri Inc.). The following discussion provides overviews of these tools and presents the workflows and test results in real-world scenarios. Conclusions and future focuses are given at the end.

2. Edgematching Tools and Workflow

The automation of edgematching is under the umbrella of research and development of Conflation tools and solutions at Esri. The edgematching tools, indicated in *Figure 2*, were briefly introduced in our recent paper (Lee et al. 2014) on geoprocessing conflation tools and workflows. More details are given below to help understand the two edgematching tools: Generate Edgematch Links and Edgematch Features, and how they are used in workflows for improving cross-border data quality.

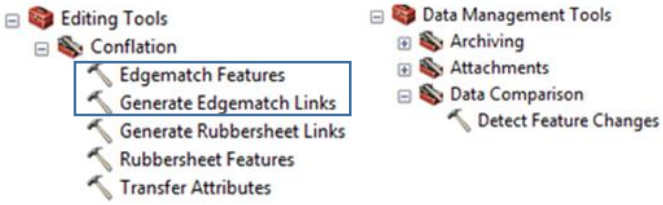

Figure 2. Edgematching tools in Conflation toolset released in ArcGIS 10.2.1.

2.1. Generating Edgematch Links

The Generate Edgematch Links (GEL) tool is designed to automatically create links between two neighboring line inputs, namely Source Features and Adjacent Features. It finds disjoined features near the meeting edges, determines correspondent features, and generates lines from source features to the matched adjacent features. These lines are edgematch links, as illustrated in *Figure 3*.

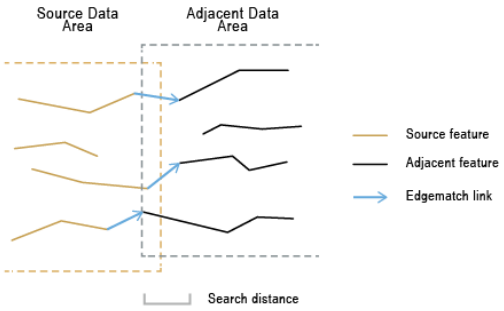

Figure 3. Illustration of Generate Edgematch Links.

The edgematch links are to be used to guide the feature adjustment by the Edgematch Features tool, as explained next. The matching process of the GEL tool is mostly based on proximity, topology, and continuity analysis, as well as optional attributes. The details of the analysis and algorithm are not the focus of this discussion and may be covered in a future paper.

The edgematch links carry the following attributes:

- SRC_FID – The source feature ID at the starting points of the links.
- TGT_FID – The adjacent (target) feature ID at the endpoints of the links.
- EM_CONF – Values representing the level of confidence, ranged from 0 to 100, where 100 as the highest level of confidence.

The EM_CONF values reflect the match conditions, therefore, the quality of links. The matching process largely depends on data quality and complexity. The less ambiguity exists in the data, the stronger match can be made, therefore, higher EM_CONF values. Example a. in *Figure 4* shows a link of EM_CONF value 100 with no ambiguity in the data; Examples b. and c. show increasing ambiguity resulting in decreasing EM_CONF values. More details can be found in ArcGIS Help topic "About edgematching". Post-inspections and editing may be necessary as explained in *Section 2.3*.

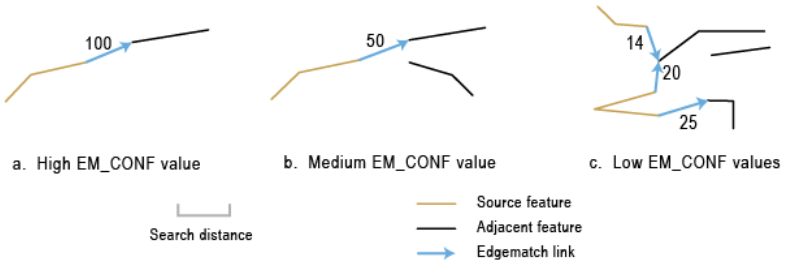

Figure 4. Examples of edgematch links and EM_CONF values

2.2. Edgematch Features

The Edgematch Features (EF) tool is designed to adjust features guided by the edgematch links produced by the GEL tool introduced above. Based on the required and optional feature inputs, lines associated with the links are adjusted accordingly so they end at new locations and connect properly with their matched features. The three available adjustment methods are:

- MOVE_ENDPOINT—Moves the endpoint of an input line to the new ending location.
- ADD_SEGMENT—Adds a straight segment between the endpoint of an input line and the new ending location.
- ADJUST_VERTICES—Moves the endpoint of a line to the new ending location and adjust the remaining vertices so their positional changes gradually reduce toward the opposite end of the line.

The determination of the new ending locations depends on what inputs are specified. Using the MOVE_ENDPOINT method, the examples in *Figure 5* show the following three scenarios (examples for the other two adjustment methods with various inputs can be found in ArcGIS Help, Edgematch Features tool reference):

- When only Input Features is specified, the endpoint of an edgematch link is used as the new ending point.
- When Input Features and Adjacent Features are specified, the midpoint of an edgematch link is used as the new ending point.
- When Input Features, Adjacent Features, and Border Features are specified, the location on a border that is nearest to the midpoint of an edgematch link is used as the new ending point.

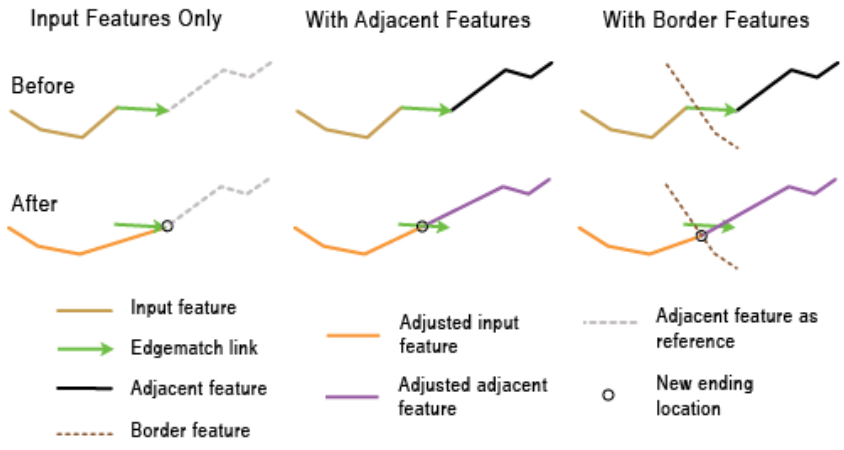

Figure 5. New ending locations and results of MOVE_ENDPOINT.

2.3. Edgematching Workflow

Depending on the data quality, level of detail, and complexity, the automated process may or may not find the right match 100% correctly; post inspection and editing can be expected, as mentioned in *Section 2.1*. The recommended workflow includes preprocessing, edgematch link generation and quality control, and edgematch adjustment.

2.3.1. Preprocessing

Preprocessing prepares the data to a good condition for maching. In general the following areas are important to consider. Most of them are simply common sense or standard best practice for geospatial analysis; details of these steps are not explained in this paper.

- Project the datasets to the same coordinate system.
- Make sure the data are topologically clean.
- Obtain consistent attributes for features continuing across borders.
- Generalize the datasets to similar level of details.
- Clean up overshoots and undershoots at borders (be aware of features coincident with borders).
- Exclude irrelevant features.

2.3.2. Edgematch link generation and quality control

Generating edgematch links is the most challenging and essential step in edgematching. The edgematch links are automatically generated by the GEL tool, but it is necessary to evaluate the result followed by quality improvement before using the links to adjust features. The following steps are necessary:

a. GEL and Evaluation – this is an automated step using the geoprocessing model shown in *Figure 6*. This model runs the GEL tool and additional analysis and produces data and evaluation information to facilitate the interactive quality improvement process in step b. later.

Figure 6. Geoprocessing model - GEL and Evaluation.

The model makes an assessment on the general link quality based on the EM_CONF values. It reports the ratio of the edgematch links with the relatively low confidence level over the total count of links.

The model also produces a point feature class containing points at locations where links intersect or touch each other. Intersecting links can be generated where multiple matching candidates are found in the edge area; they may or may not be correct, therefore, need to be verified.

Expected links may not be generated due to ambiguous conditions. To identify them the model flags for review any "dangles" - locations of the input lines that have no obvious connection and are within the specified search radius to the generated links.

b. Quality improvement – this step takes two parts:

First, the operator performs an interactive review of the evaluation results and makes necessary corrections of link issues. Wrong links need to be modified or deleted; missing links added.

Then, a model, shown in *Figure 7*, is used to automatically update the SRC_FID and TGT_FID values for all modified or added links.

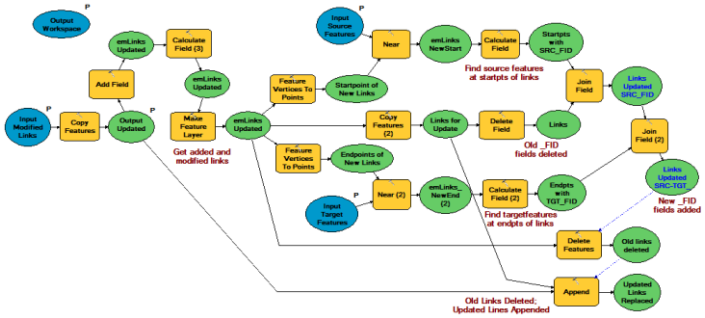

Figure 7. Geoprocessing model: Update Link Info.

2.3.3. Edgematch adjustment

Once the edgematch links are ready to use, the EF tool is used to adjust features. The source features and adjacent features (if specified) associated with the links through the SRC_FID and TGT_FID are modified according to the specified adjustment method so they are precisely connected.

The edgematching workflow described above is quite straightforward. Although some features may escape the automated link generation and the interactive quality improvement processes, the chance is rather slight. The general match accuracy can reach 85 – 95%; the interactive process should improve the links near 100%. Two real world use cases are presented in

Section 3. The edgematched data are properly continuous over the borders and become reliable basis for spatial analysis and high quality mapping.

3. Use Case 1: Road Data Edgematching

This use case requires edgematching between roads maintained by the Resource Management Service, LLC (RMS) and the commercially available TIGER roads (provided by the United States Census Bureau) outside RMS's ownership boundaries. The RMS roads are to be held in position; the TIGER roads must be adjusted to connect with correspondent RMS roads for routing analysis. The sample datasets are: EdgeRoads1km (RMS roads containing 7576 features) and GISRoads1km (TIGER roads containing 3634 features). Both datasets are clipped to the RMS ownership borders. Only features within 1 km to the borders are selected for processing, as shown in *Figure 8*, to minimize unnecessary computation time.

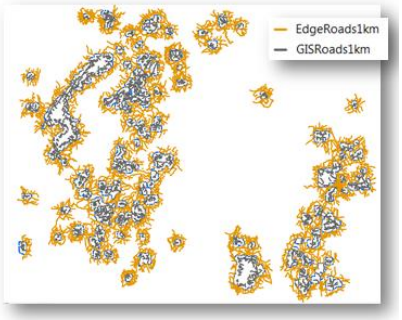

Figure 8. Use case 1: input roads.

3.1. Results of GEL and Evaluation

The process generated 454 links. Since the links are too short to be visible on the overall map, the midpoints of the links are shown in *Figure 9*. The border features are displayed as reference, not involved in the process.

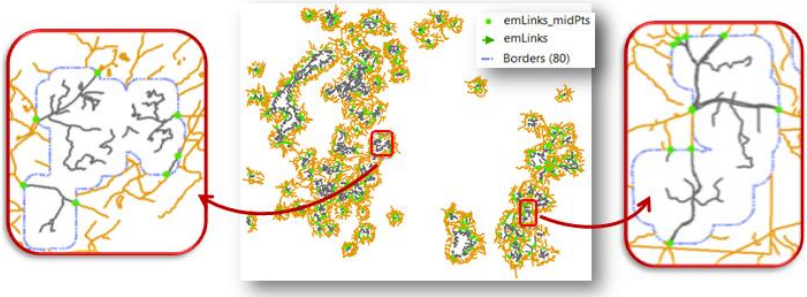

Figure 9. Automatically generated links shown by their midpoints.

As discussed earlier the links carry EM_CONF values that are affected by the level of ambiguity. The value of 100 indicates a match with no spatial and attribute ambiguity; the value decreases when conditions are more complicated or confusing mostly due to the existence of multiple candidate features, as shown in *Figure 10*.

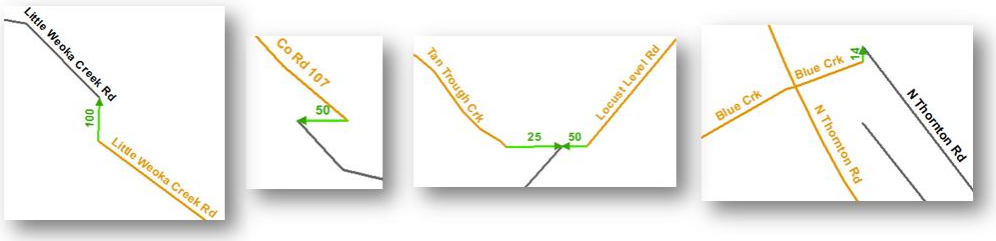

Figure 10. Example links with varying EM_CONF values.

The process produced the following data and information for the quality control process:

- 134 of the 454 links with EM_CONF values lower than 33, indicating a relatively high level of complexity and ambiguity encountered in the matching process.
- 33 points at locations of intersecting links, see red dots *in Figure 11*.
- 62 points at locations of potential missing links, see red X in *Figure 11*.

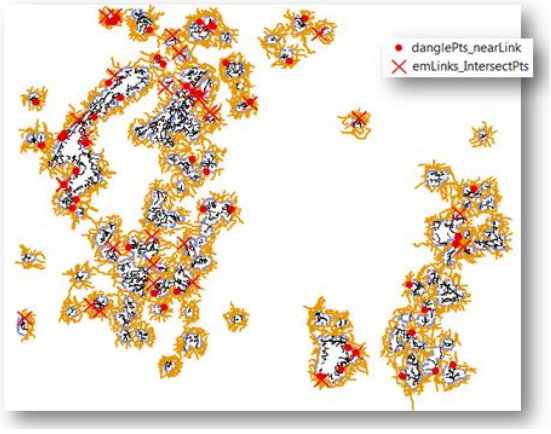

Figure 11. Locations of intersecting and potential missing links.

3.2. Results of Link Quality Improvement

The quality control improvement was done interactively with the help of a customized Add-in toolbar, which allows the operator to go through the features and records one by one easily, zoom to the involved features auto-

matically, make necessary edits, and add review notes. Example (a) in *Figure 12* shows links with low EM_CONF values and intersecting links were modified and noted as "Good" and "Recheck" and (b) one of the intersecting links was wrong and noted to be removed and two locations flagged as potential missing links turned out false alarms (no links added).

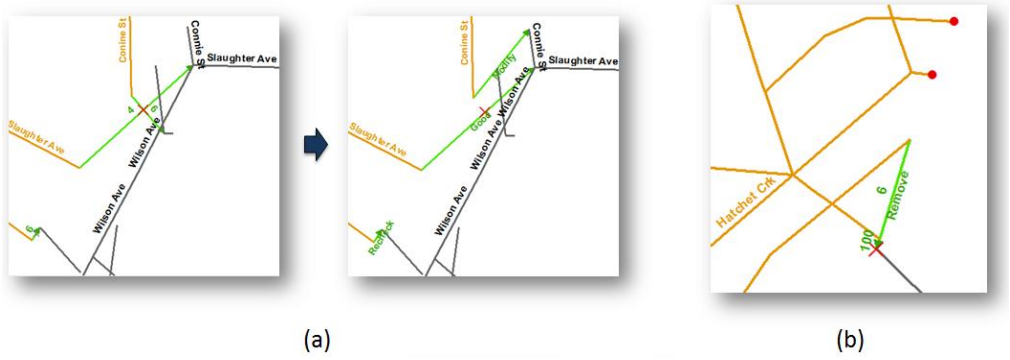

Figure 12. Examples of link quality control.

The review notes were added into a field, REV_FLAG; features that were not flagged for review have Null value. The counts (FREQUENCY values) for all review flag values are shown in the table in *Figure 13*. The estimated accuracy of automatic link generation (based on false alarm verification in quality control) is 85% and error rate 15%; see the calculation details in *Figure 13*. Upon the finish of quality control the link accuracy is near 100%. The geoprocessing model Update Link Info introduced early was used on the modified and added links to update their associated feature IDs. The links are ready to use for the adjustment next.

emLinks_freqREVFLAG

OBJECTID *	FREQUENCY	REV_FLAG
1	267	<Null>
2	5	Added
3	66	Good
4	46	Modify
5	55	Recheck
6	20	Remove

Estimated automatically generated link accuracy:
Correct links / total links = 388 / 459 => **85%**
where correct links are Null, Good, and Recheck in the REV_FLAG field, i.e. (267+66+55)

Estimated corrected link rate:
Incorrect links / total links = 71 / 459 => **15%**
where incorrect links are Modify, Remove, and Added in the REV_FLAG field, i.e. (46+20+5)

Estimated final link accuracy => 100%

Figure 13. Summary of link quality control results and accuracy estimates.

3.3. Results of Edgematch Adjustment

The adjustment is done by the EF tool. A few examples of the adjusted result using the MOVE_ENDPOINT method are given in *Figure 14*.

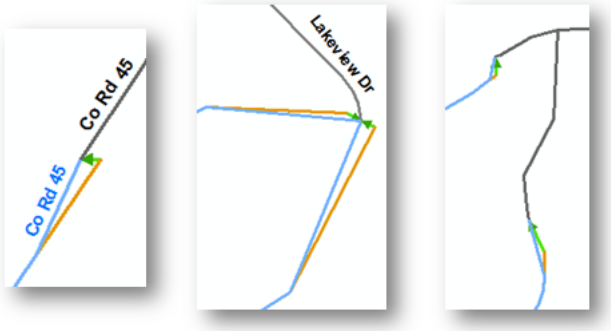

Figure 14. Adjusted features (blue lines) by MOVE_ENDPOINT method.

The total automated processing time to run the two models and the EF tool was 12.8 seconds; and the manual work took about 1.6 hours. The adjusted TIGER data connect with RMS data properly, therefore, can better support RMS's service needs that depend on routing.

4. Use Case 2: ELF Edgematching Project

According to the project website (ELF 2015), the European Location Framework (ELF) project, launched in March 2013, "will run for three years and deliver a pan European cloud platform and web services to build on the existing work of the INSPIRE Directive and enable access to harmonised data in cross border applications." Edgematching, as part of the ELF framework (Hopfstock et al. 2015), is technically essential to ensuring data harmonization crossing country borders in the project. With limited test data, we have been able to produce some initial edgematching results on hydrographic data and road data, as presented below.

4.1. Edgematching of ELF Hydrographic Data

The hydrographic datasets were obtained through the ELF project data sharing portal. They cover a small area on either side of the border between Norway and Sweden as shown in *Figure 15*. Using the model "GEL and Evaluation" as discussed above with a search distance of 100m, 26 edgematch links were generated along with their midpoints for visualization purposes, see *Figure 15*. Only one point (the red X in *Figure 15*) was produced at the location where two links touch. The longer one of the two touching links was determined to be incorrect and therefore deleted.

The Sweden data contains more details than Norway data because of different source capture scales. Although many features are near the border area, few correspondent features are found on the Norway side. If

a larger search distance was used, more false matches might be made. The initial test result further proves that the matching accuracy can reach above 95%.

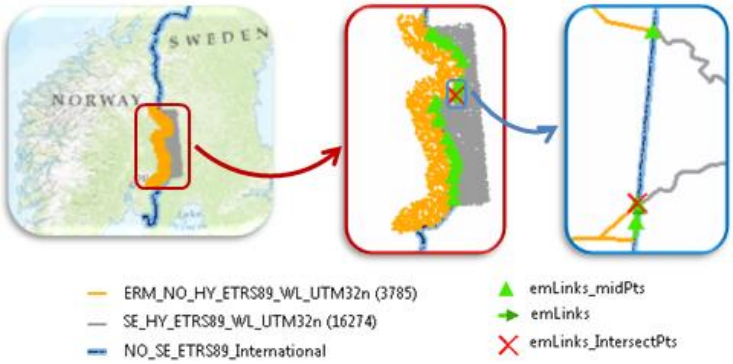

Figure 15. GEL and Evaluation results on ELF hydrographic data.

To perform the adjustment using these links, the EF tool was used on both hydrographic inputs from Norway and Sweden and the available border between them. The tool connects both inputs at border locations following the method described in *Session 2.2*. The two examples in *Figure 16* show in detail the input features, links, and the adjusted features which are now connected at border locations marked by the black circles.

Figure 16. Adjusted features are connected on the border.

To allow continued maintenance of edgematch status over time, the connecting points may have to be pre-defined and agreed by the neighboring countries; then the points can be supplied as connecting features (CFs) for the adjustment process, as described in the ELF project document (Brühl 2015). It should be easy to incorporate CFs in the feature matching tools and workflows. In the meantime, there seems no one definitive way of getting CFs. In some cases they will come from previous manual edgematch, but as soon as changes happen they may become outdated. The automatically deduced new connecting locations (marked by the black circles in *Figure 16*) on the border could be excellent candidates for CF points. These locations can be easily extracted by a geoprocessing tool.

4.2. Edgematching of ELF Road Data and Additional Thoughts

The road data near the border between Poland and Czech Republic were obtained through ELF project data sharing. Features within 700m to the border were selected to participate in the edgematching process, as shown in *Figure 17*. Again, the model "GEL and Evaluation" was used with a search distance of 100m; the results include 165 edgematch links and their midpoints, 10 points at intersecting locations, and 18 points at dangle ends of features near the border, as shown in *Figure 17*.

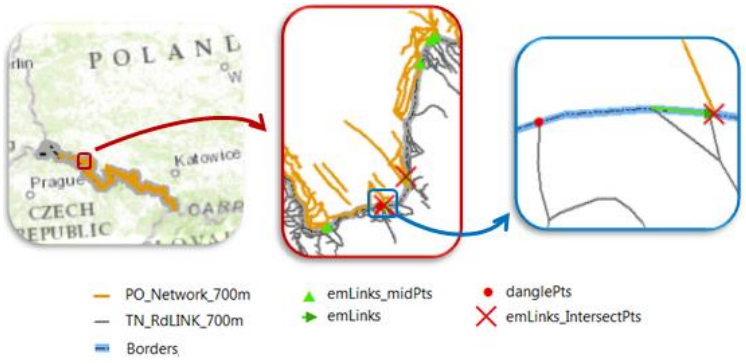

Figure 17. GEL and Evaluation results on ELF road data.

The quality improvement and feature adjustment processes are similar to what has been presented above. Due to the length limitation of the paper, the results and description of these processes are omitted. However during the review process the following areas were noticed, along with some thoughts on future enhancement and development.

- A link crossing road features as show in *Figure 18*. This happens where the road conditions are complex, for example the loops and road intersections occur near the border and the seemingly matching features don't have dangle ends. Additional analysis would be necessary to recognize the situation and identity the correct matches.
- Conflicting overlaps as shown in *Figure 18*. This issue cannot be resolved using a simple clipping of the features by the border features, especially where the overlap features are supposed to be coincident with the borders. It requires an analysis to find such situations and some spatial adjustment to align them properly. An initial prototype has been made to align features with a reference feature. The examples in *Figure 19* show two input boundaries (red and black lines) that are supposed to be coincident but apart. The feature alignment process brings the conflicting portion of the red line towards the black line. As shown in the enlarged area, the resulting blue line in the conflicting area is now per-

fectly aligned with the black line and remains continuous at the turning point. Feature alignment can be needed between linear features, polygon features, or combination of both.

Figure 18. Issues with complex road data near borders.

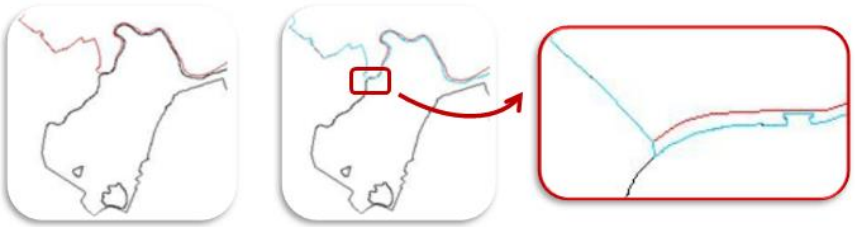

Figure 19. Prototype result of aligning one feature with another feature.

5. Conclusion

Streamlining cross-border data is critical to organizations that perform spatial analysis and map the information beyond the boundaries they maintain. The edgematching tools and the workflows presented above combine highly automated and manageable interactive processes and play a key role in improving cross-border data consistency and reliability. Input data quality and complexity certainly affect matching accuracy. Hydrographic data tend to be less congested than road data; less ambiguity results in higher matching accuracy and more correct links. The test cases indicate that a matching accuracy of 85-95% or better is achievable using the GEL tool; and the final results are further improved by the suggested quality control processes. Future research and development will focus on the following aspects, but not limited to them:

- Improvement in the matching analysis to further reduce incorrect edgematch links, including detection of links crossing other features and possible use of contextual features in match decision making.

- Refinement of the EM_CONF values so they better reflect the level of ambiguity or confidence.
- Detection of cross-border features in conflict (overlaps, braided common features, and so on) and spatial adjustment for feature alignment.

It is also important to continue enhancing the workflow for better support to long term maintenance – keeping the edgematched base data up to date. Changes happen frequently at local and global scales; timely and highly accurate detection of the differences between update and the base data is necessary. The Conflation tools in ArcGIS use feature matching techniques to detect feature changes and have great potential to facilitate data updating, attribute transfer, and multi-scale database linkages (Baella et al. 2014). Our solutions for data integration, conflation, and quality improvement will continue meeting the challenges and making data collaboration work more efficient.

References

Baella B, Lee D, Lleopart A, Pla M (2014) ICGC MRDB for topographic data: first steps in the implementation, The 17th ICA Generalization Workshop, 2014, Vienna, Austria

Brühl M (2015) D2.3.3 Edge Matching. ELF project documentation. http://elfproject.eu/sites/default/files/D2.3.3_EdgeMatching.pdf. Accessed 10 April 2015

ELF (2015) About ELF. http://www.elfproject.eu/. Accessed 10 April 2015

Hopfstock A, Brühl M, Laurent D, Dorie O, Grémeaux N, Delattre N, Werhahn S, Koistinen K, Latvala P, van Oosterom P (2015) D2.3 Data Maintenance and Processing Specification. ELF project documentation. http://elfproject.eu/sites/default/files/D2.3%20Data%20Maintenance%20and%20Processing%20specification.pdf. Accessed 10 April 2015

Lee D, Yang W, Ahmed N (2014) Conflation in Geoprocessing Framework - Case Studies, GEOProcessing, 2014, Barcelona, Spain