We formulate it as an optimization problem for choosing a best match plan that minimizes the total discrepancies between counterpart features. In particular, we employ the classic network flow models in operation research to describe the bi-directional matching problem for conflation as shown in **Figure 2**.

An Optimization-based Matching Method and its Application in Merging Administrative Boundary Data

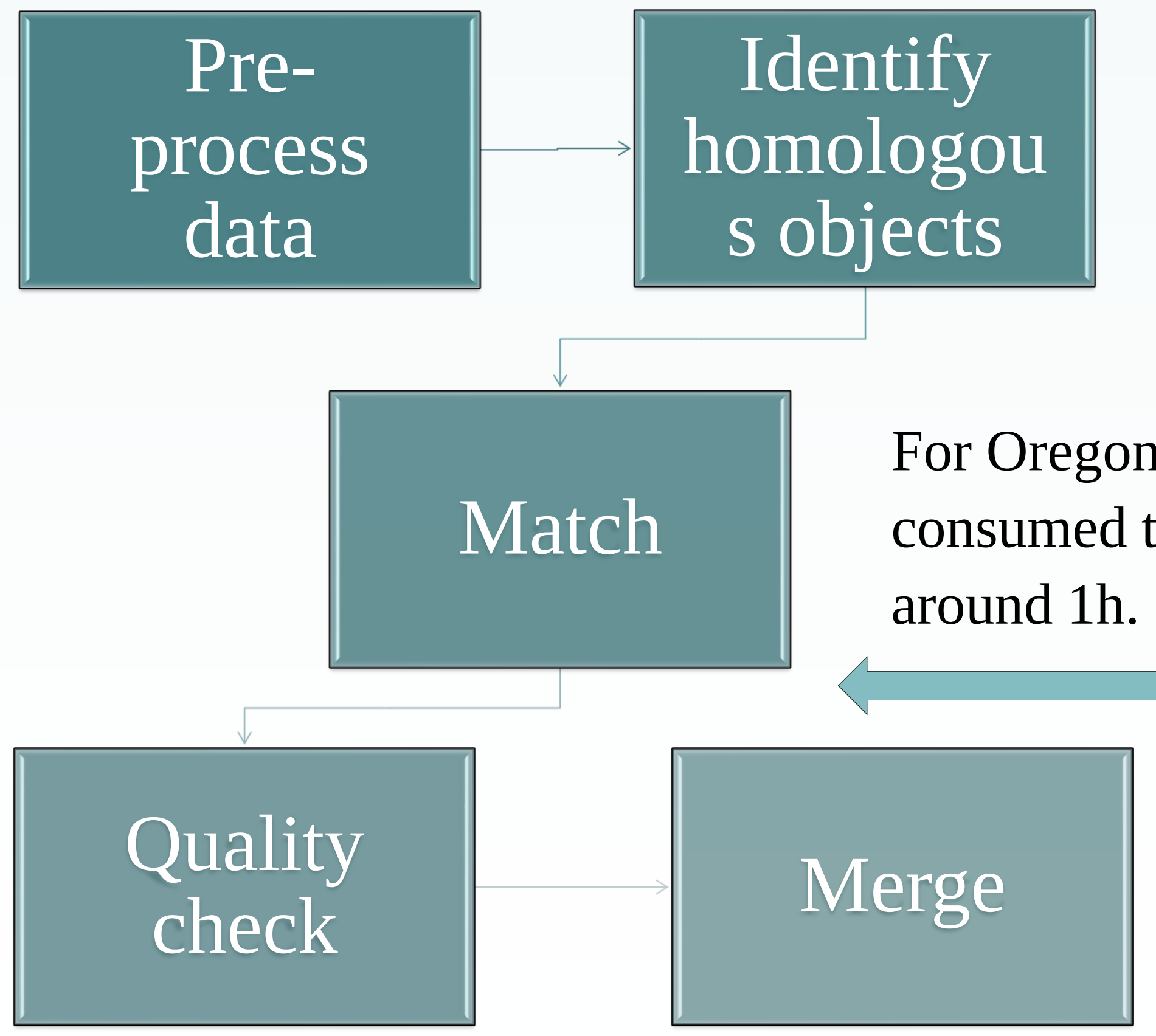
Ting Lei, Wenjun Yang
Geography and Atmosphere Science Department, University of Kansas

Abstract

As a critical data management task, conflation in GIS aims to determine the corresponding features from different datasets that in reality represent the same entities. This is called feature matching, which is used as a guidance to merge attributes of corresponding features between datasets. Based on the classification of features, there are point, polyline, and polygon matching methods. This study focuses on matching polygons and explores optimization–based matching methods for conflating two datasets.

Workflow

For completeness, we will explore both matching and merging in our experiments. The workflow can be broken down into five steps, the first two steps are data preparation. The third and fourth steps are matching normalized datasets and quality check. The last step is to merge attributes according to matching result.


Pre-process data

Identify homologous objects

Match

Quality check

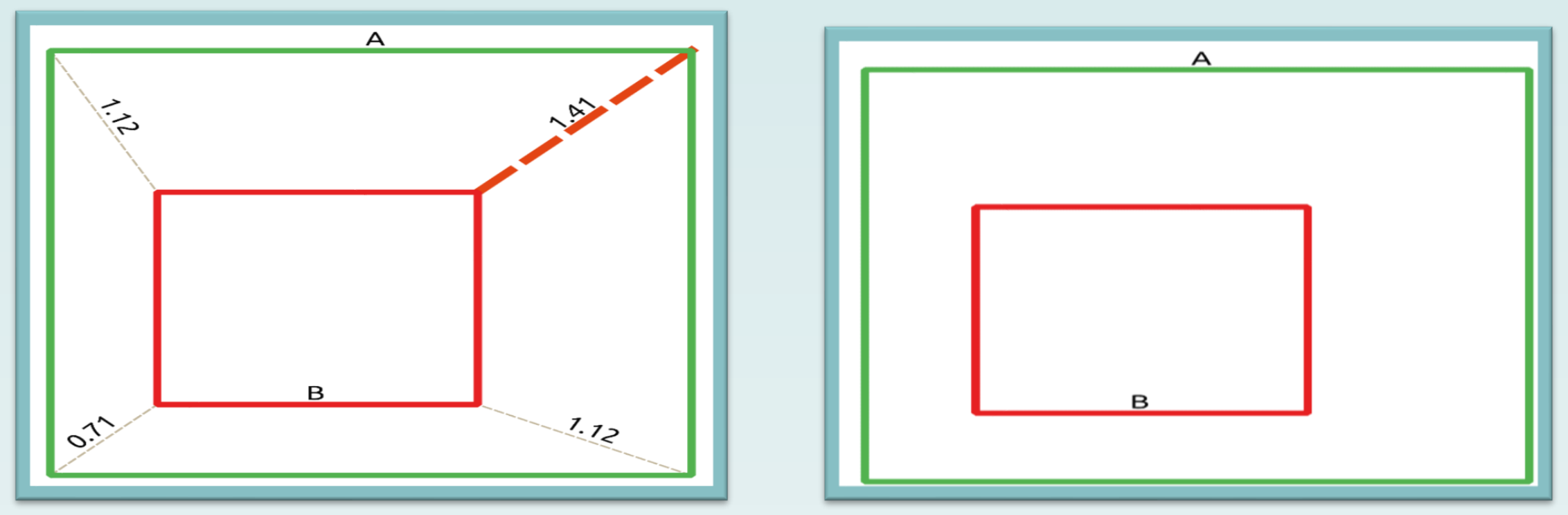
Merge

For Oregon data, time consumed till now is around 1h.

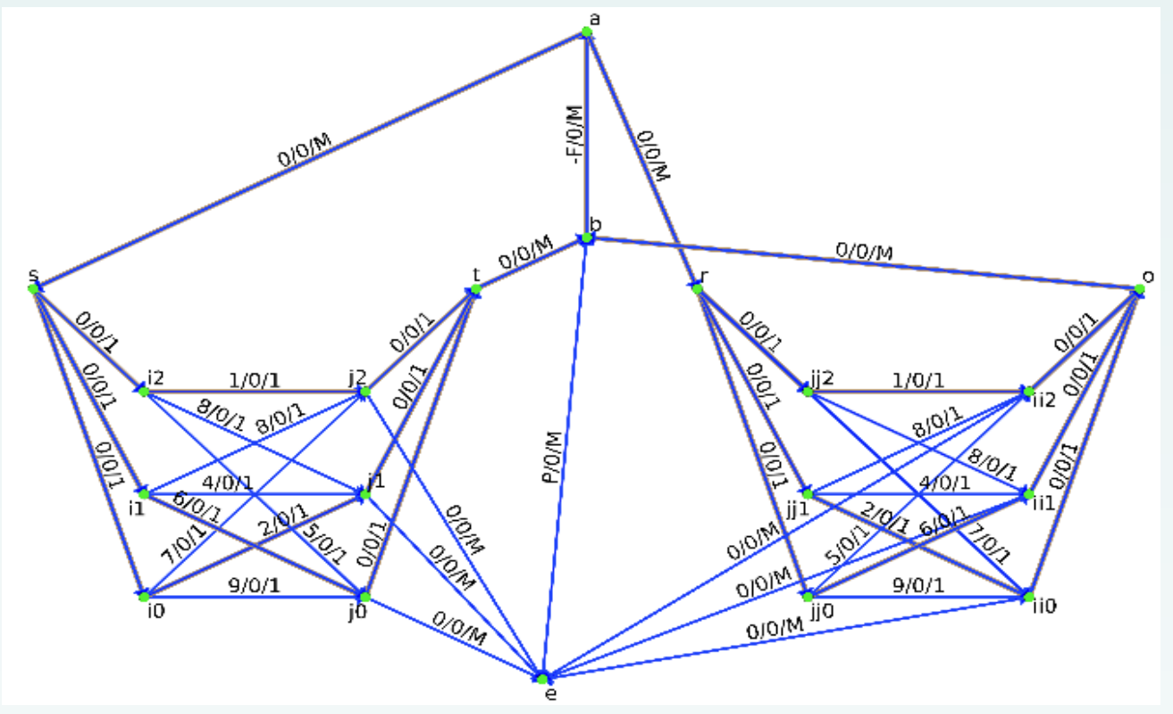
Methodology

We used a directed Hausdorff distance that can describe the part-whole relationship between polygons. If a polygon A in one dataset is spatially contained by a polygon B in the other dataset, the directed distance from A to B is zero. Otherwise, a non-zero value characterizes the deviation of A from B.

$$H_d(A,B) = \max_{p \in A} d(p,B)$$



a. directed distance from A to B   b. Directed distance from B to A



Network flow (fc-bimatching) model
*labels for each edge represent the cost/lower bound/capacity of flow.

We formulate it as an optimization problem for choosing a best match plan that minimizes the total discrepancies between counterpart features.

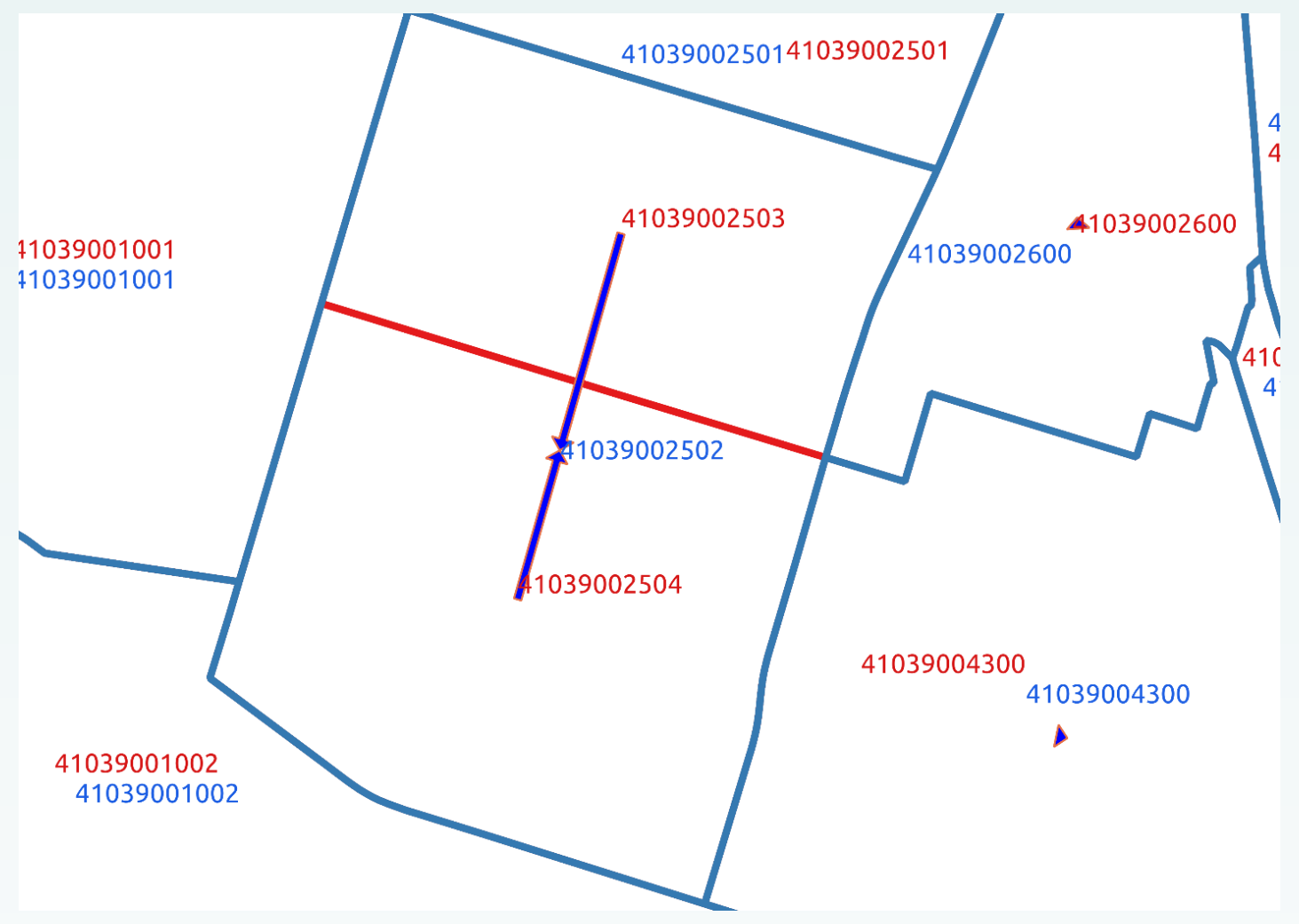
Matching type	Explanation
One-to-one	Boundaries have no change or minor adjustments can be safely ignored.
One-to-many	One unit splits into two or more units.
Many-to-one	A reverse relation of one-to-many. Two or more units consolidate into one unit.
Many-to-many	Complicated*

*Many-to-many type is complicated that we treat two or more units in one GDB as a holistic bigger unit and simplify the type into one-to-many or many-to-one.

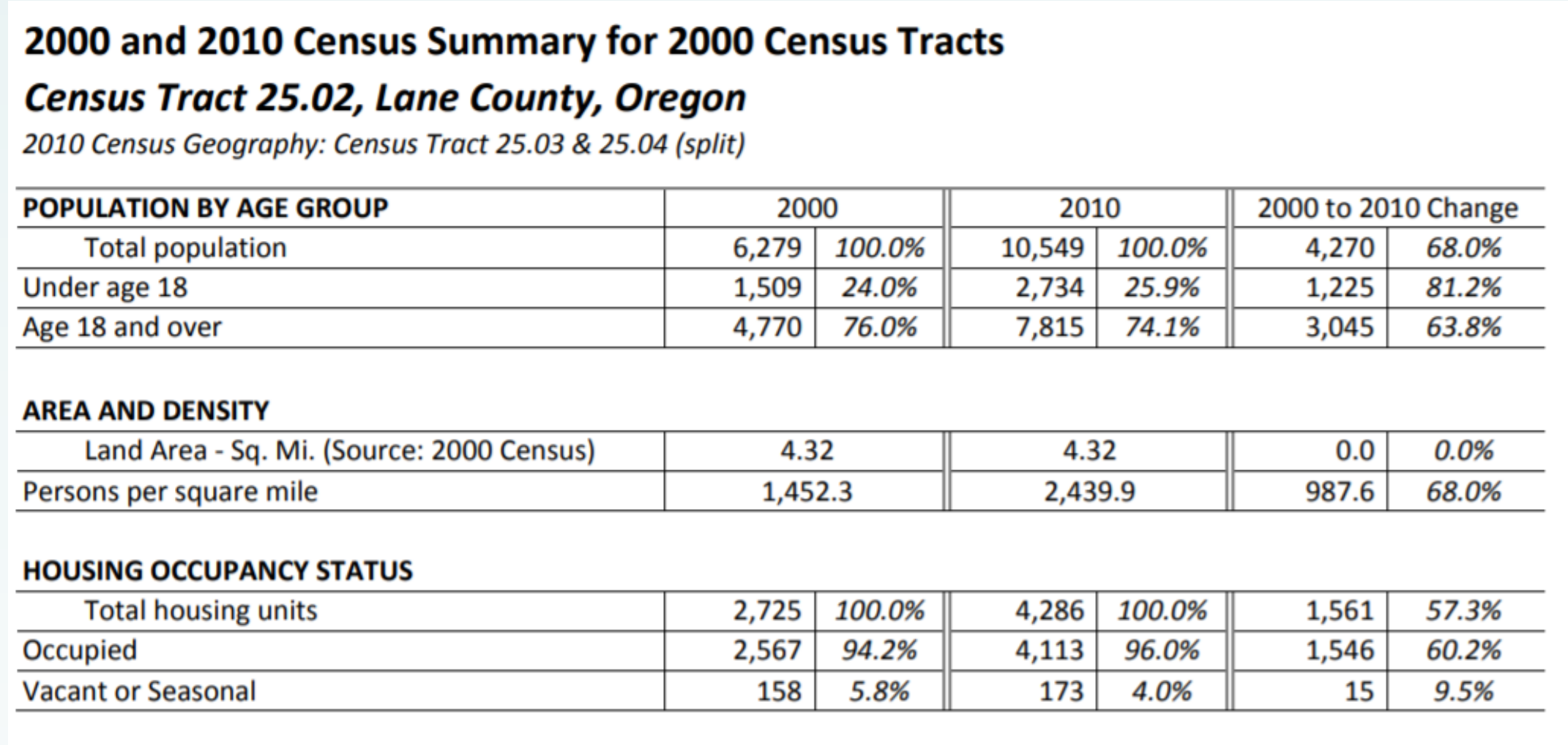
Experiment

We apply the method to match the tract-level maps of Oregon state from two decennial census--2000 and 2010. The experiment result demonstrates that the method can identify the split and consolidation of polygons between two datasets due to boundary changes. The process costs a relatively short time with little human interference.

To verify the accuracy of the matching result, the study referred to “2000 and 2010 comparison profiles” as the ground truth by Population Research Center of Portland State University. The profiles literally demonstrate block-level changes in census tracts. The following figures shows two matching examples and its ground truth:



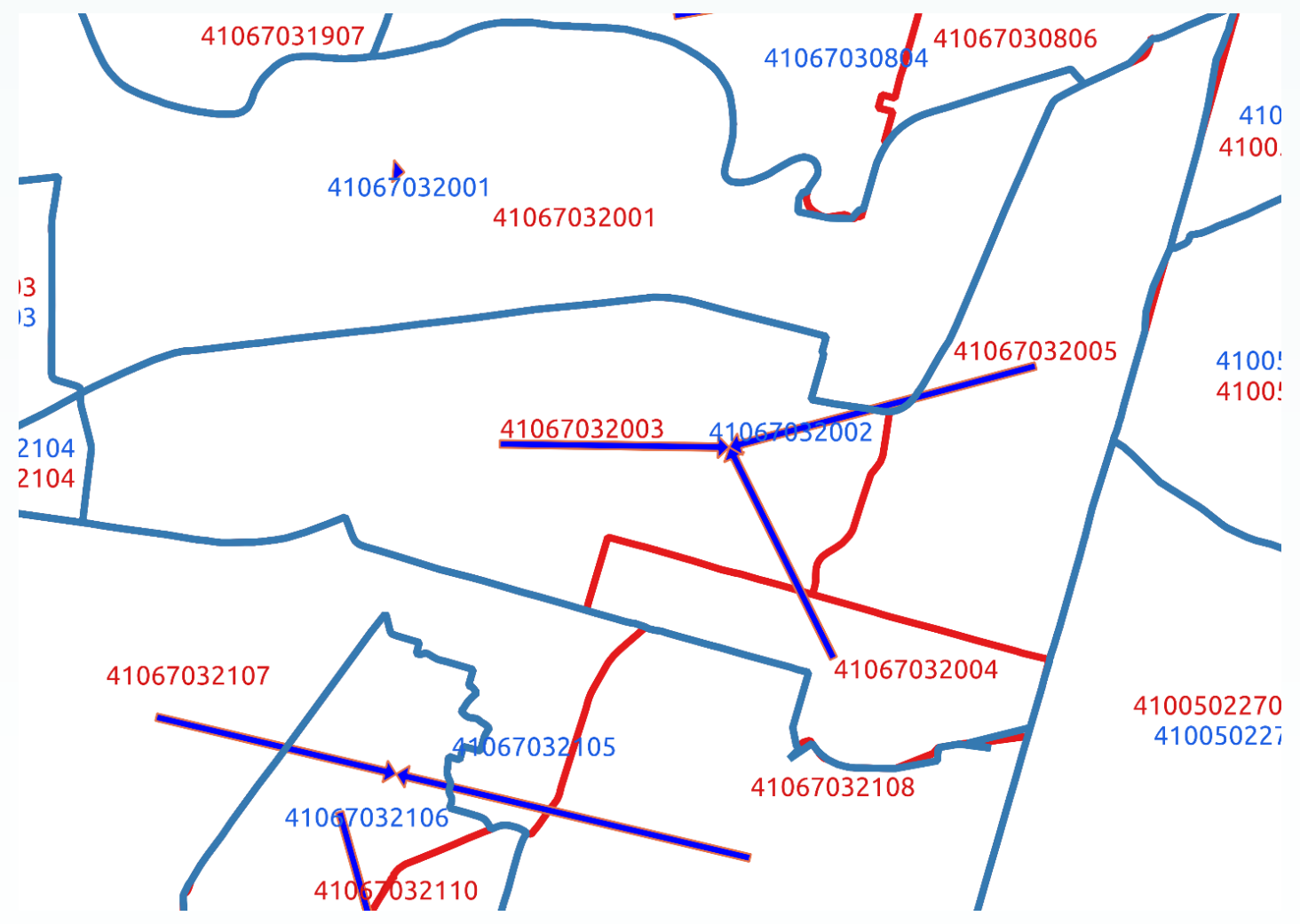
a. one-to-two



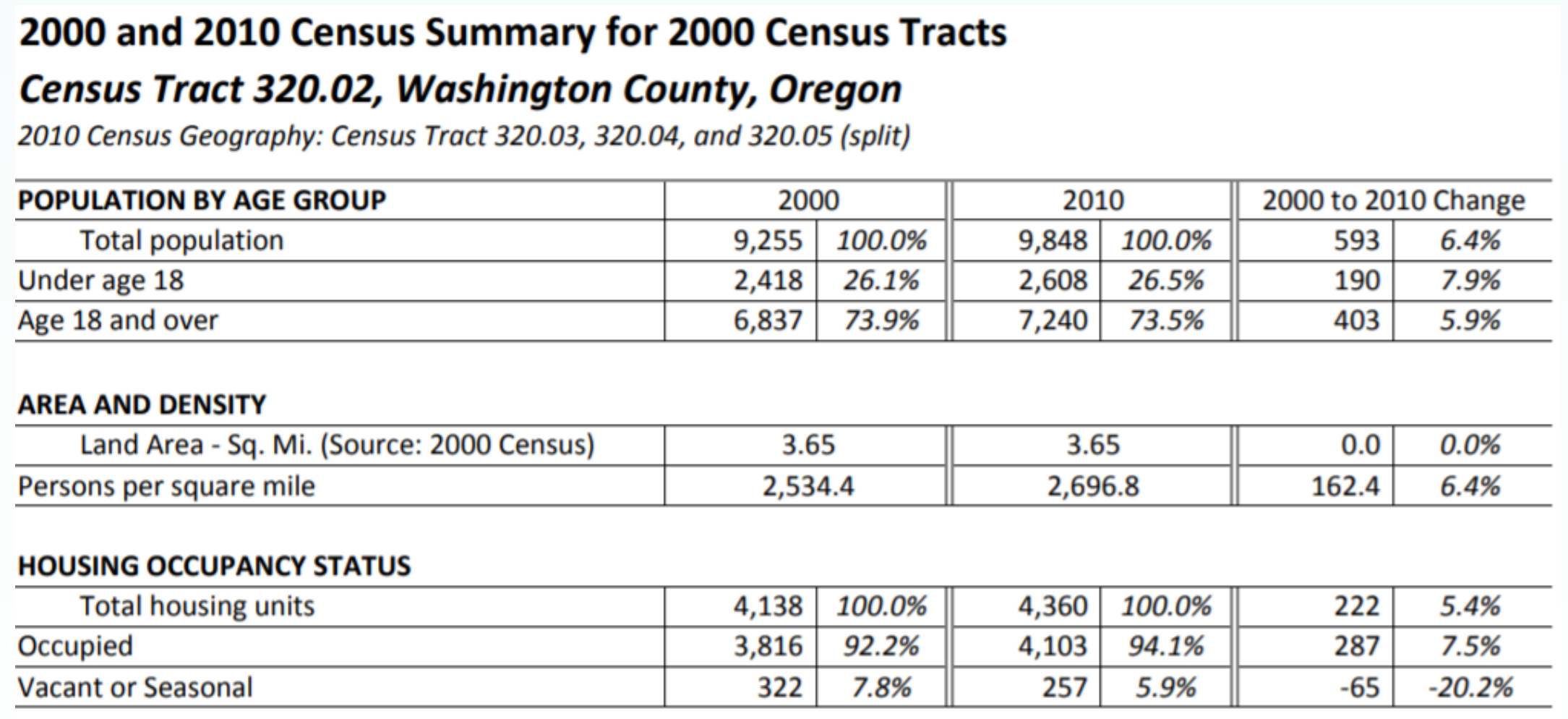
2000 and 2010 Census Summary for 2000 Census Tracts
Census Tract 25.02, Lane County, Oregon
2010 Census Geography: Census Tract 25.03 & 25.04 (split)

POPULATION BY AGE GROUP	2000		2010		2000 to 2010 Change	
Total population	6,279	*100.0%*	10,549	*100.0%*	4,270	*68.0%*
Under age 18	1,509	*24.0%*	2,734	*25.9%*	1,225	*81.2%*
Age 18 and over	4,770	*76.0%*	7,815	*74.1%*	3,045	*63.8%*
AREA AND DENSITY						
Land Area - Sq. Mi. (Source: 2000 Census)	4.32		4.32		0.0	*0.0%*
Persons per square mile	1,452.3		2,439.9		987.6	*68.0%*
HOUSING OCCUPANCY STATUS						
Total housing units	2,725	*100.0%*	4,286	*100.0%*	1,561	*57.3%*
Occupied	2,567	*94.2%*	4,113	*96.0%*	1,546	*60.2%*
Vacant or Seasonal	158	*5.8%*	173	*4.0%*	15	*9.5%*

b. the ground truth



a. one-to-two and one-to-three



2000 and 2010 Census Summary for 2000 Census Tracts
Census Tract 320.02, Washington County, Oregon
2010 Census Geography: Census Tract 320.03, 320.04, and 320.05 (split)

POPULATION BY AGE GROUP	2000		2010		2000 to 2010 Change	
Total population	9,255	*100.0%*	9,848	*100.0%*	593	*6.4%*
Under age 18	2,418	*26.1%*	2,608	*26.5%*	190	*7.9%*
Age 18 and over	6,837	*73.9%*	7,240	*73.5%*	403	*5.9%*
AREA AND DENSITY						
Land Area - Sq. Mi. (Source: 2000 Census)	3.65		3.65		0.0	*0.0%*
Persons per square mile	2,534.4		2,696.8		162.4	*6.4%*
HOUSING OCCUPANCY STATUS						
Total housing units	4,138	*100.0%*	4,360	*100.0%*	222	*5.4%*
Occupied	3,816	*92.2%*	4,103	*94.1%*	287	*7.5%*
Vacant or Seasonal	322	*7.8%*	257	*5.9%*	-65	*-20.2%*

b. the ground truth

🟦 2000 census tract boundaries 🟥 2010 census tract boundaries
*the blue triangles represent one-to-one match that the polygon has no change between decennial censuses.

Result

Oregon has 755 tracts in 2000 census and 827 in 2010. Consequently, the method automatically completed matching all 2000 tracts with 2010 tracts, which are 831 matching pairs in total. Compared with the ground truth, the rate of accuracy in matching one-to-many case is 98.0%. The rates of accuracy for many-to-one and no change categories are both 100%.

Compared with the previous matching method, the current one simplifies the merging process by focusing on changed features, reasonably ignoring the boundary changes caused by different cartographical techniques.

Acknowledgements

- The census data is derived from Census Bureau of the United States and National Historical GIS (NHGIS).
- Thanks to Ting Lei and the Department of Geography and Atmosphere Science, KU.
- Contact