OAK RIDGE National Laboratory

Enhancing and validating GeoNames with Digital Nautical Charts: A case study in the mapping of freeform map labels

Dakotah D. Maguire, Jason C. Kaufman, Alexandre Sorokine, Robert Stewart

Geospatial Science and Human Security Division

ORNL is managed by UT-Battelle LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

Background

- Temporal and spatial variation in locations vary among multiple sources
  - Cartographic offset vs. precise point locations
  - Name variations and aliases
- Do multiple sources validate each other and add value when combined?

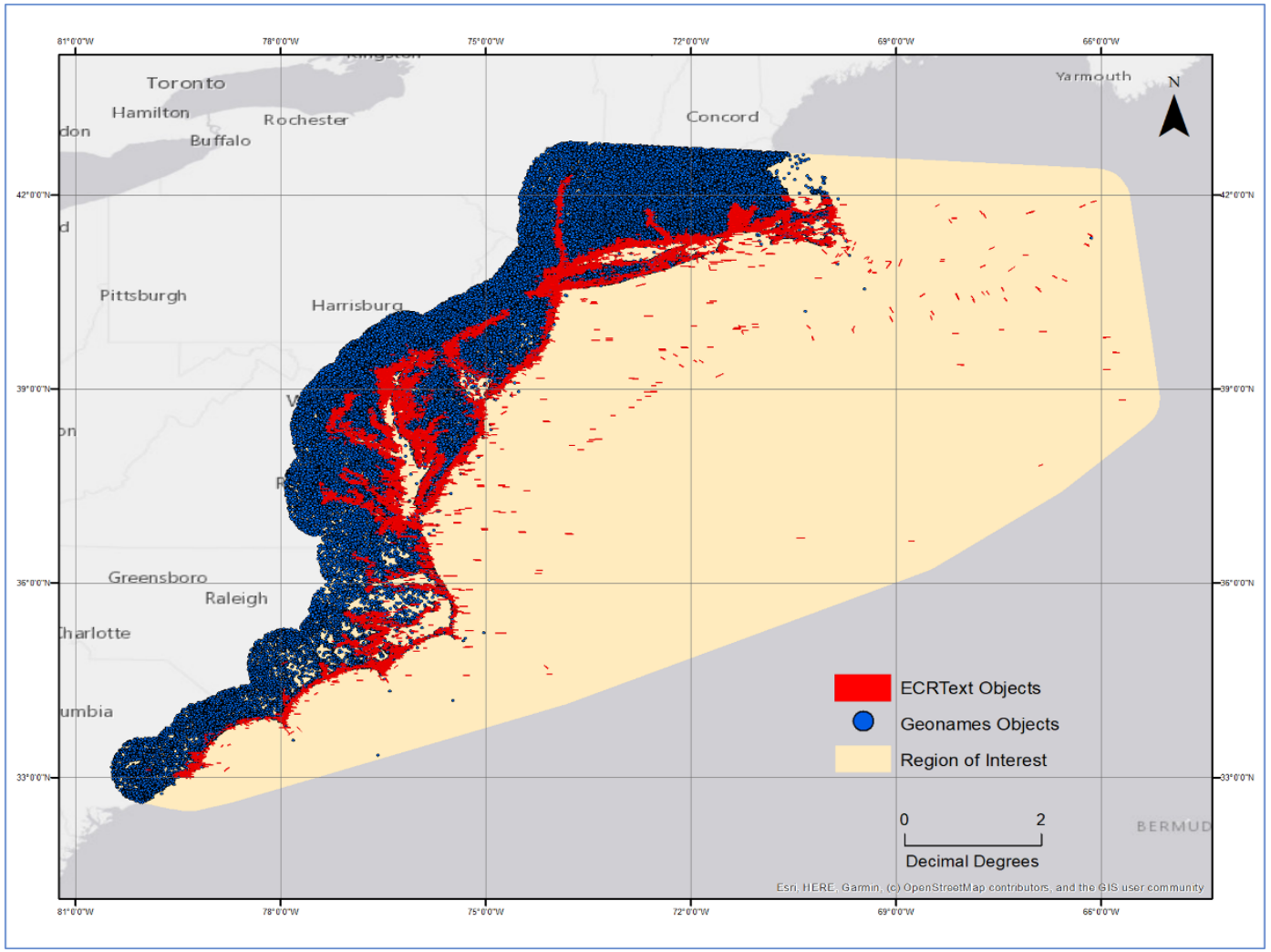
Study question

- Can Digital Nautical Charts (DNC) data enhance GeoNames data by supporting validation, expanding aliases, and filling in missing data?
  - Use textual similarity and spatial proximity
  - Detect instances of ECRText within GeoNames

Data

- Digital Nautical Charts is a substantial, long-lived, worldwide vector chart database for ship navigation
  - Developed and maintained by the National Geospatial-Intelligence Agency and the National Oceanic and Atmospheric Administration (NOAA, 2022)
- GeoNames is an open-source gazetteer with over 27 million geographic names

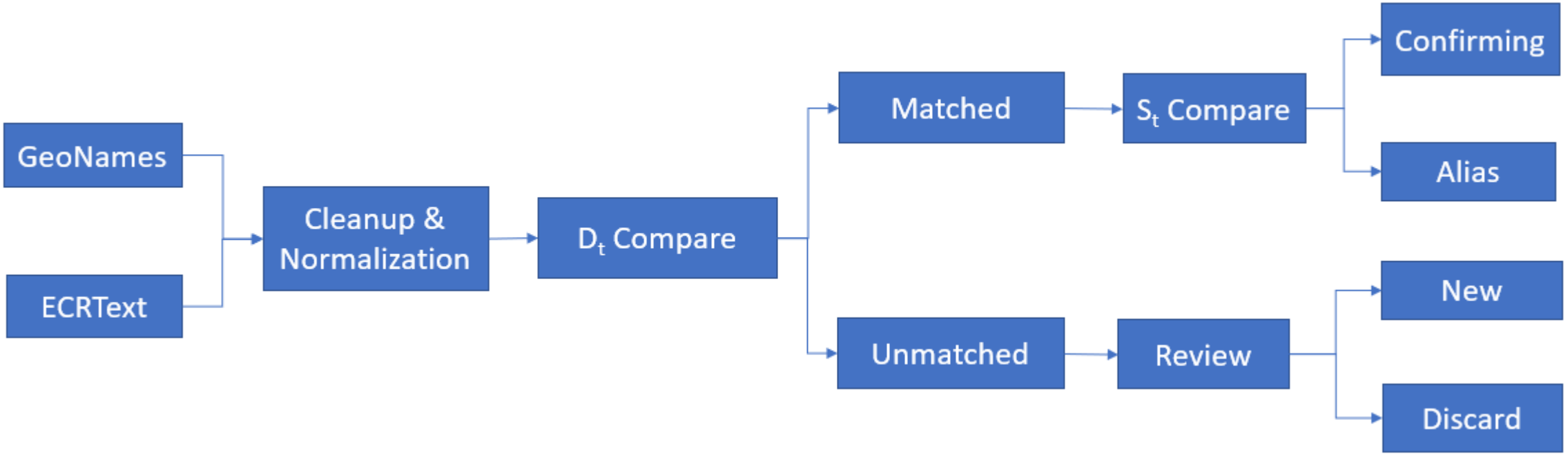
Study area and data



Our region of interest is DNC Region 17 (DNC17), containing 14,224 ECRText objects and 222,772 GeoNames locations along US East Coast from 42° N to 33° N latitude

OAK RIDGE National Laboratory

Methods



Methods: Pairs and similarity

- We can compute the distance between any ECRText–GeoNames data pair as the closest distance between the GeoNames point and the ECRText bounding polygon
  - We define a pair as "sufficiently close" if their distance is less than a distance threshold *dt*
- We use a computationally effective trigrams method to calculate the similarity of the names; results in similarity values in the range of 0–1 (Dunn, 2020)
  - We define a pair as "sufficiently similar" if the similarity value is greater than a threshold *st*

Methods: Pairs and similarity (continued)

- Semiautomated workflow developed to classify ECRText
- ECRText first cleaned and normalized using US Chart No. 1 (NGA, 2019)
- ECRText and GeoNames within *dt* distance (a buffered concave hull around ECRText features) are considered matched pairs and unmatched otherwise
  - Distance threshold value of 0.5° was found to provide reasonable classification outcomes

Methods: Pairs and similarity (continued)

- Matches then compared for text similarities s and classified as confirming or alias
  - Similarity threshold value = 0.8
  - Unmatched pairs manually reviewed to discard results not obviously geographic names; remaining are candidates for new locations

Potential outcomes

- **Confirming:** Here, an ECRText is part of a sufficiently close pair (distance ≤ *dt*) with an exact text–name match (similarity = 1). The interpretation here is that ECRText very close to a GeoNames location with exactly the same name provides some confirmation that the GeoNames is likely accurate.

- **Alias:** Here, an ECRText is part of a sufficiently close pair (distance ≤ *dt*) with a sufficiently similar name (*st* ≤ similarity < 1). The interpretation here is that ECRText found very close to a GeoNames location with nearly the same name may well be a new alias (i.e., alternate name) for the already existing GeoNames location.

- **New:** Here, an ECRText is either too distant from any GeoNames (distance > *dt*) or is part of a pair with significantly different text–name matches (similarity < *st*). The interpretation here is that the ECRText object is too distant or too different to be reasonably associated with an existing GeoNames location. These may well be candidates for new locations to enhance the GeoNames dataset.

- **Discard:** Here, an ECRText is too distant from a GeoNames location with a label that is not associated with a geographic name (e.g., Unexploded Ordinance). The interpretation here is these place names are highly unlikely to have any clear benefit to GeoNames.

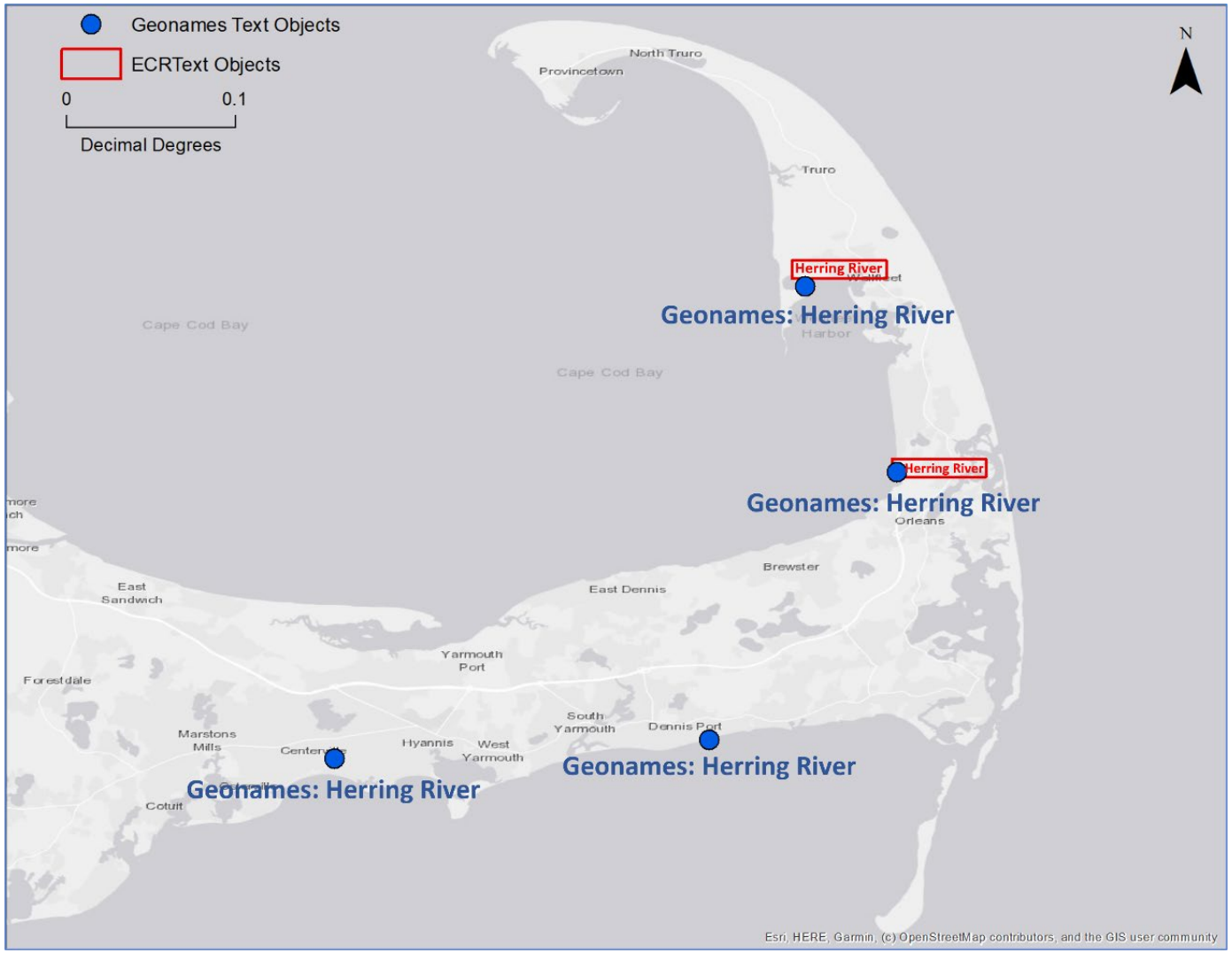
Initial results

- Unique text objects in the study area:
  - 14,224 ECRText objects
  - 208,203 GeoNames
- Most significant benefit is validation with 87% of ECRText data classified as confirming
- ECRText in the Alias (2%) and New (8%) categories has the potential to enhance and extend existing locations

Classification	Total ECRText Objects
Confirming	12,439 (87%)
Alias	275 (2%)
New	1,103 (8%)
Discard	407 (3%)

Challenges

- Handling crowded places where names are very similar or identical
  - Near things are more related than distant things



Ongoing work

- Conflationary step
  - Duplication across libraries and across scales
- Additional text similarity methods
  - Levenshtein vs. Trigrams performance
- OpenStreetMap as a third data source
  - Potential for more frequent name changes (i.e., added aliases)
  - Duplicate entries, or closely related objects with nearly identical names

Copyright and disclaimer