Evaluating Trustworthiness of Multiple Overlapping Data Sources

Yeasmin Ara Akter[1], Aberto Abelló[2], Petar Jovanovic[3], Tomer Sagi[4], Katja Hose[5]

[1,2,3] Universitat Politècnica de Catalunya, Barcelona, Spain
[4,5] Aalborg University, Denmark

[1] yeasmin.ara.akter@upc.edu, [2] aabello@essi.upc.edu, [3] petar@essi.upc.edu, [4] tsagi@cs.aau.dk, [5] khose@cs.aau.dk

04/07/2022

INTRODUCTION

Information Fusion

- Integrating multiple data sources to produce more consistent, accurate, and useful information than that provided by any individual data source.
- Fusing data from multiple overlapping data sources
- Discovering the true value from noisy information

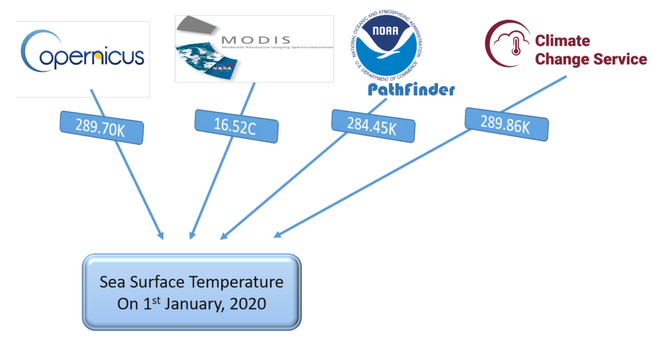

Figure 1 – Sea surface temperature from different sources

Related Work

Table 1: Related Work

Systems	Type	Uncertainty Handling	Truth Discovery Method			Evaluation Metric
			Considered Source Dependency	Truth Computation	Ground Truth Evaluation	
Apollo-social [2]	Probabilistic Graphical Model	×	×	Maximum Likelihood	×	Precision, Recall
CATD [3]	Optimization	×	×	Weighted averaging	×	MAE, RMSE
RCHDTD [4]	Optimization	×	×	Weighted Voting Weighted Median	✓	Mean Normalized Absolute Distance (MNAD)
SmartMTD [6]	Probabilistic Graphical Model	×	✓	Majority Voting	✓	Precision, Recall, F1-Score, Execution Time
EPTD [5]	Iterative	×	×	Majority Voting	✓	MAE, RMSE
SRTD [7]	Iterative	✓	×	Majority Voting	✓	Specificity (SPC), Matthews Correlation Coefficient (MCC), Cohen's Kappa (Kappa)
RPPTD [8]	Optimization	×	×	Majority Voting	✓	Execution Time
RTD [9]	Iterative	×	×	Mean Shift Clustering	✓	MAE, MSE, R-Squared

Limitations:

- Uncertainty is ignored in most of the trustworthiness evaluation system
- Different data type must be treated differently
- Use of gold standard data
- Error is not traced throughout the workflow
- No specific evaluation metric to provide overall degree of trustworthiness

Objectives

- Determining a representation method for both uncertain and missing data
- Determining an efficient attribute conflict resolution method that supports aligning data from multiple sources
- Developing an efficient tracing method of data transformations with the help of data provenance techniques to represent the propagation of trust
- Determining a metric to estimate the degree of trustworthiness of sources given multiple overlapping data sources

Proposed Architecture

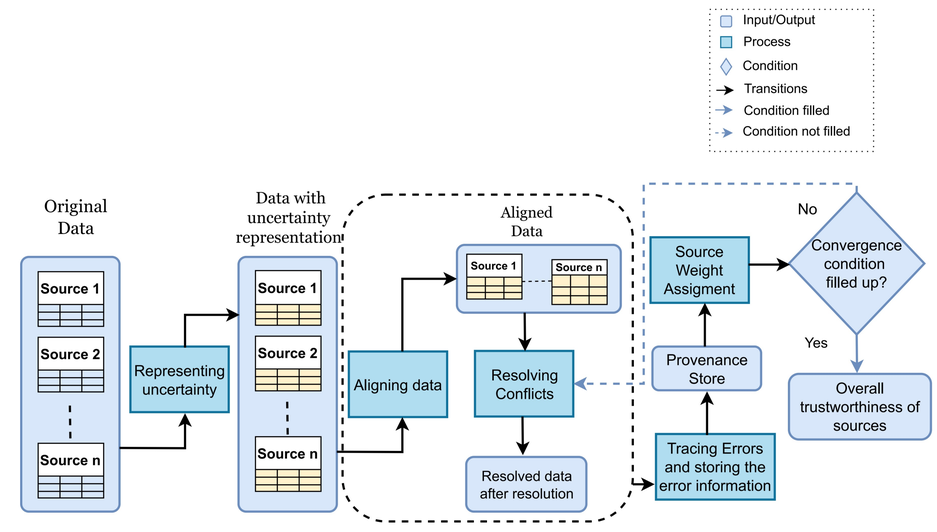

Figure 2 – Trustworthiness Development Process

Challenges

- Differentiating Contradictory and missing data
- Source Dependency
- Proper Domain Subdivision
- Tracing the errors throughout the data transformation workflow
- Identifying best provenance technique
- Defining a metric to decide the overall degree of trustworthiness
- Having master data is difficult to evaluate the system

Use Case

Environmental data source has been used as the use case where four different sources provide sea surface temperatures. Table 2 shows how much data are provided by each individual source. Figure 3 & Figure 4 illustrate the data statistics (Min, Max, Average, Median) of the sources.

Table 2: Data Source Desciptions

Source Name	Values	Resolution	Number of Non-Null Data (%)	Number of Null
Copernicus	Temperature (Daily)	0.05	6.3	93.7
Climate Data	Temperature (day and Night)	0.05	0.74	99.25
Modis-Aqua	Temperature (Daily)	0.04	21.10	78.8
Pathfinder	Temperature (Day and Night)	0.04	96.5	3.4

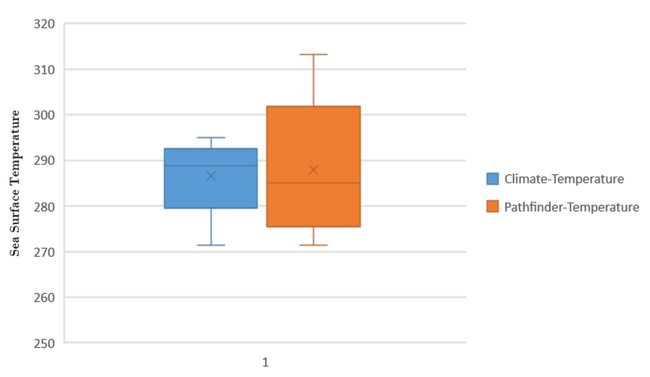

Figure 3 – Data Statistics of Climate and PathFinder Source.

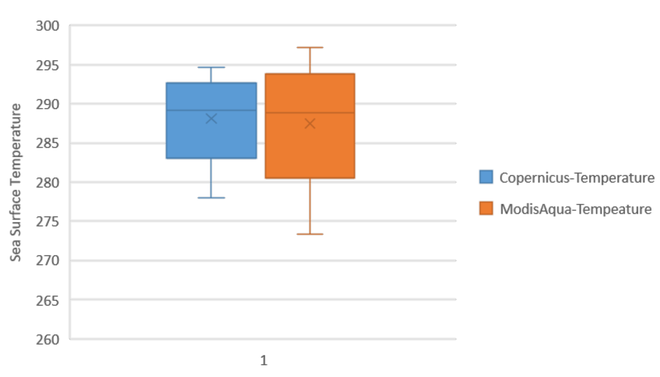

Figure 4 – Data Statistics of Modis and Copernicus Source.

Results

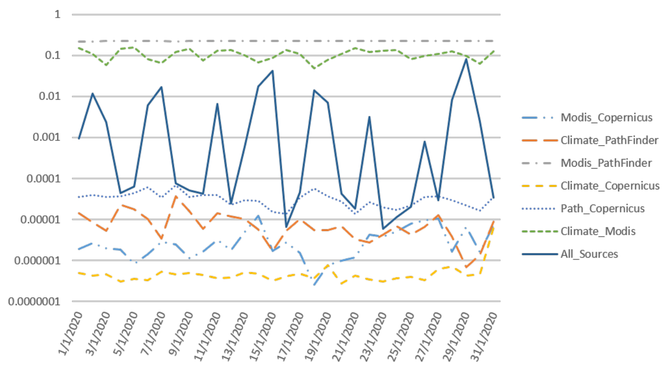

Figure 5 – Discordance among multiple sources.

Conclusion

- Discordance among the sources varies according to different cost function.
- If data sources provide non null values in same latitude and longitude, they have less errors.
- Absolute representation of uncertain data has high impact on the discordance among the sources.