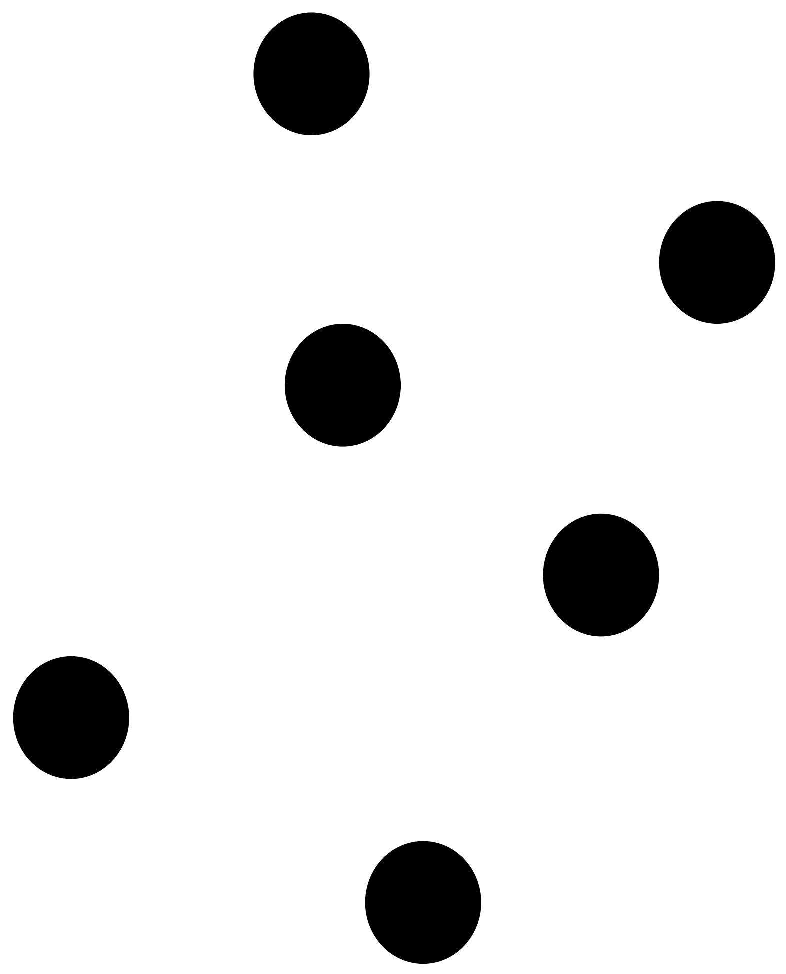
Self-consistent Deep Geometric Learning for Heterogeneous Multi-source Spatial Point Data Prediction

Dazhou Yu

2024

Background: spatial data

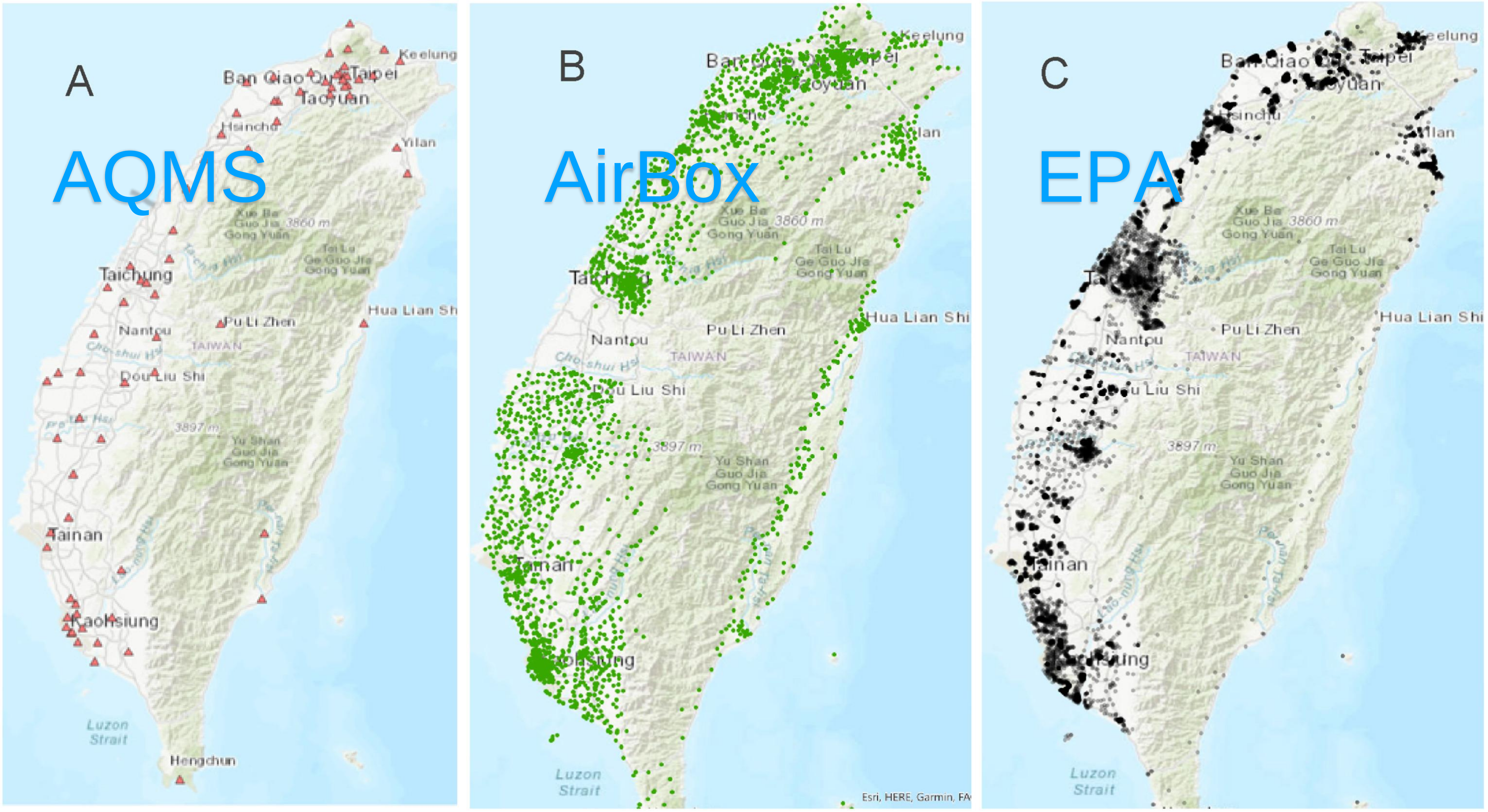
- Coordinate & attributes
- Vector data: point
- Raster data: height*width*channel



(1) Vector data

(2) Raster data

Multi-source Spatial Point Prediction

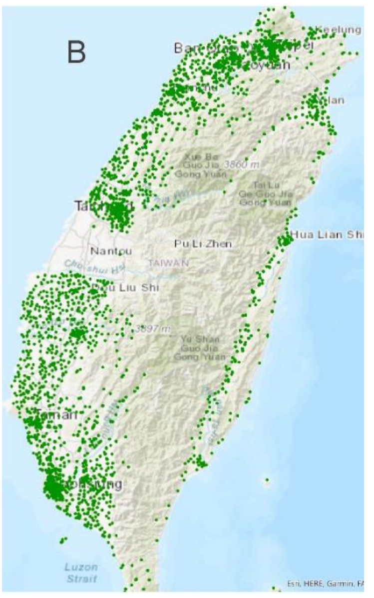


An example of multi-source spatial point prediction problem: Varying distribution of three data sources including (A) 74 AQMSs, (B) 3704 LASS AirBox sensors, and (C) 9701 EPA MicroStations.

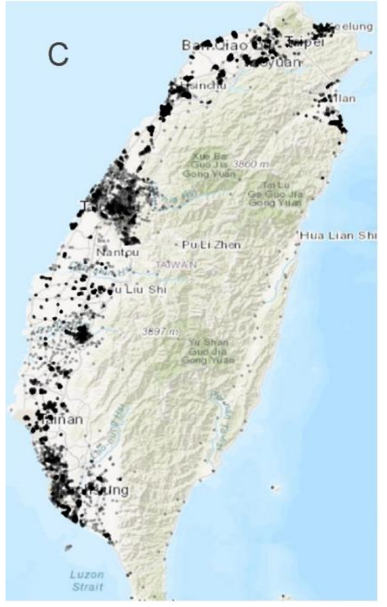
Problem of Multi-source Spatial Point Prediction

Location s
(coordinates & attributes)

$D^{(1)}$



$D^{(2)}$



$D^{(3)}$

Prediction y

Challenges & Contributions

- Different data sources without ground truth
  - Self-supervised Deep Multi-source Spatial Prediction (DMSP)
- Different qualities
  - Fidelity score, a learnable parameter
- Different spatial locations
  - Geographically aware multi-source graph neural network

Adaptive Self-supervised Fusion of Multiple Data Sources

$$MI(\tilde{Y}, \hat{Y}) = \sum_{\hat{y} \in \hat{Y}} \sum_{\tilde{y} \in \tilde{Y}} w(\tilde{y}, \hat{y}) p(\tilde{y}, \hat{y}) \log \frac{p(\tilde{y}, \hat{y})}{p(\tilde{y}) p(\hat{y})}$$

$$= \sum_{i=1}^{N} \sum_{\hat{y} \in \hat{Y}} \sum_{y^{(i)} \in Y^{(i)}} \mathbb{1}(y^{(i)} = \tilde{y}) p(y^{(i)}, \hat{y}) \log \frac{p(y^{(i)}, \hat{y})}{p(y^{(i)}) p(\hat{y})}$$

$$\max \sum_{i=1}^{N} C_i MI(Y^{(i)}, \hat{Y}) \quad \Longleftrightarrow \quad \min \sum\nolimits_{i=1}^{N} C_i \mathcal{L}_i(Y^{(i)}, \hat{Y})$$

Deep Multi-source Spatial Prediction (DMSP) Framework

Deep Multi-source Spatial Prediction (DMSP) Framework

Deep Multi-source Spatial Prediction (DMSP) Framework

Deep Multi-source Spatial Prediction (DMSP) Framework

Deep Multi-source Spatial Prediction (DMSP) Framework

Datasets

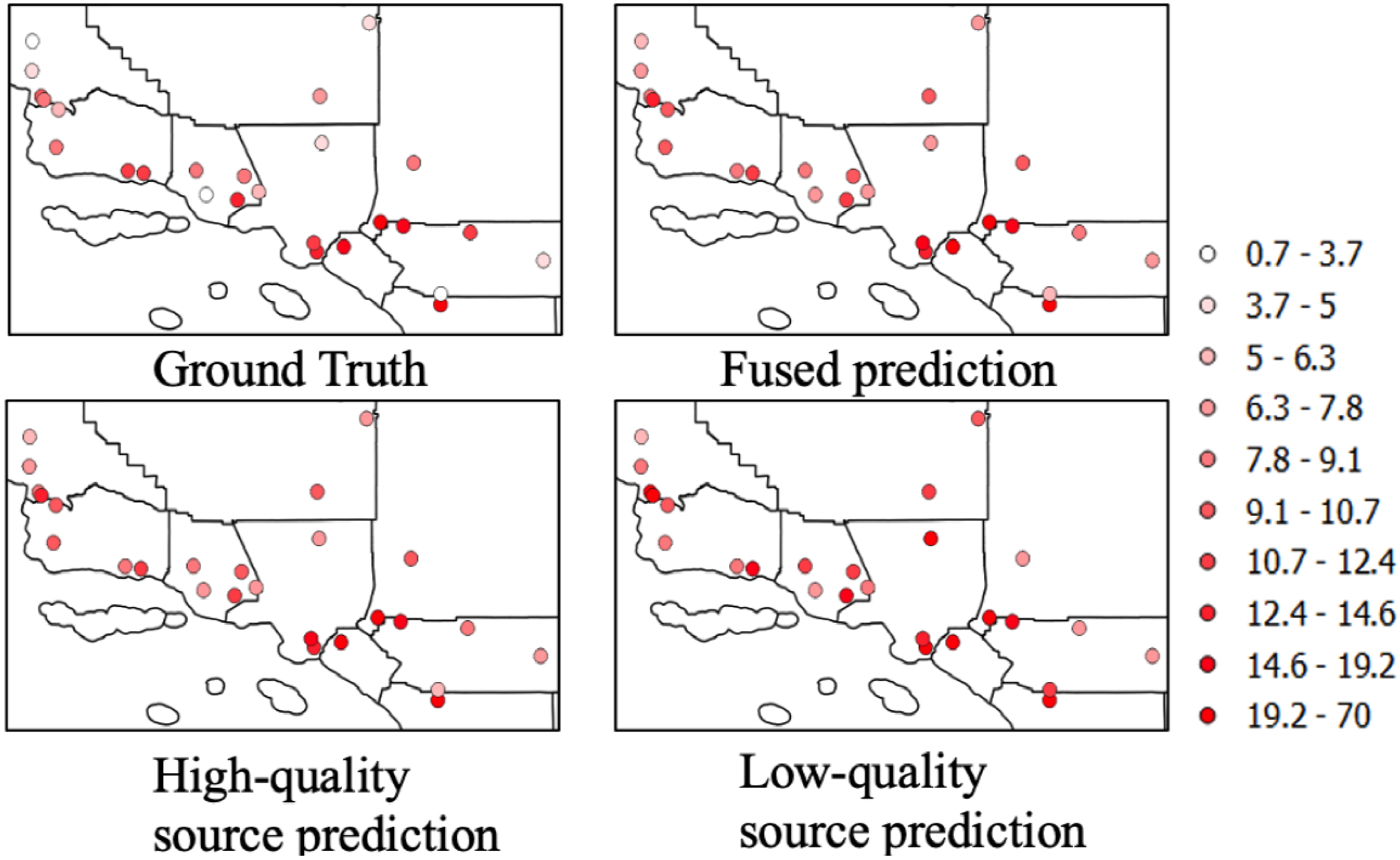
Name	Number of Locations in source 1	Number of Locations in source 2	Time period	Task
SouthCalAir Dataset	26	512	2019/01/01 to 2019/12/31	PM2.5 prediction
NorthCalAir Dataset	63	1110	2019/01/01 to 2019/12/31	PM2.5 prediction
Flu Dataset	48	156	2010-2015	Flu count prediction

Effectiveness Analysis

Dataset	Method	MAE	RMSE	EVS	CoD	Pearson
SouthCalAir	SRA-MLP	3.211±0.059	5.305±0.091	0.416±0.017	0.411±0.015	0.686±0.012
	RR-XGBoost	5.811±0.047	8.450±0.107	-0.258±0.050	-0.2644±0.056	0.351±0.010
	NARGP	4.476±0.853	7.000±1.484	0.084±0.334	0.076±0.331	0.487±0.138
	DMSP	**3.112±0.059**	**4.878±0.234**	**0.542±0.026**	**0.504±0.036**	**0.737±0.021**
	GeoPrior	3.236±0.305	5.926±0.254	0.411±0.107	0.411±0.109	0.670±0.041
	Space2Vec	3.135±0.303	5.996±0.283	0.401±0.122	0.395±0.107	0.672±0.056
	DMSP-H	4.091±0.486	6.106±0.344	0.277±0.119	0.263±0.139	0.555±0.059
	DMSP-F	14.835±2.850	22.445±3.768	-6.972±2.271	-9.361±2.379	0.055±0.074
NorthCalAir	SRA-MLP	3.000±0.059	5.374±0.378	0.404±0.026	0.390±0.024	0.636±0.021
	RR-XGBoost	3.705±0.030	6.177±0.334	0.121±0.066	0.119±0.068	0.557±0.014
	NARGP	3.317±0.035	6.26±0.580	0.186±0.158	0.185±0.158	0.579±0.052
	DMSP	**2.423±0.083**	**4.474±0.489**	**0.590±0.052**	**0.586±0.046**	**0.768±0.034**
	GeoPrior	2.845±0.055	4.920±0.154	0.481±0.027	0.480±0.029	0.690±0.021
	Space2Vec	2.641±0.103	4.509±0.283	0.585±0.022	0.584±0.027	0.762±0.056
	DMSP-H	3.768±0.823	6.048±1.394	0.236±0.274	0.221±0.286	0.506±0.205
	DMSP-F	17.110±3.289	42.426±20.589	-32.214±33.699	-37.378±35.118	0.003±0.013
SCR	SRA-MLP	0.782±0.031	0.969±0.029	0.007±0.008	-0.016±0.014	0.092±0.080
	RR-XGBoost	0.939±0.014	1.208±0.032	-0.573±0.177	-0.646±0.236	0.031±0.040
	NARGP	0.616±0.045	0.773±0.059	0.457±0.024	0.446±0.025	0.698±0.012
	DMSP	**0.478±0.032**	**0.574±0.025**	**0.606±0.046**	**0.605±0.045**	**0.780±0.027**
	GeoPrior	0.505±0.035	0.600±0.054	0.553±0.037	0.553±0.039	0.731±0.021
	Space2Vec	0.498±0.015	0.615±0.063	0.584±0.031	0.580±0.029	0.751±0.046
	DMSP-H	0.516±0.026	0.631±0.030	0.507±0.135	0.489±0.158	0.726±0.071
	DMSP-F	0.484±0.018	0.609±0.014	0.596±0.048	0.588±0.047	0.776±0.026

- DMSP achieves the highest score
- The performance degrades if only learns from a single source or treats all sources equally

Prediction visualization



- Fused prediction more closely matches the ground truth

Summary

- Self-supervised Deep Multi-source Spatial Prediction (DMSP)
- Fidelity score, a learnable parameter
- Geographically aware multi-source graph neural network
- DMSP framework not only outperforms existing state-of-the-art methods but also provides meaningful insights into the data.

Paper

Code