ST-MVL: Filling Missing Values in Geo-sensory Time-series Data


Xiuwen Yi[1,2], Yu Zheng[2,1]

Southwest Jiaotong University, Chengdu, China

Microsoft Research, Beijing, China

Released Codes & Data

Motivation

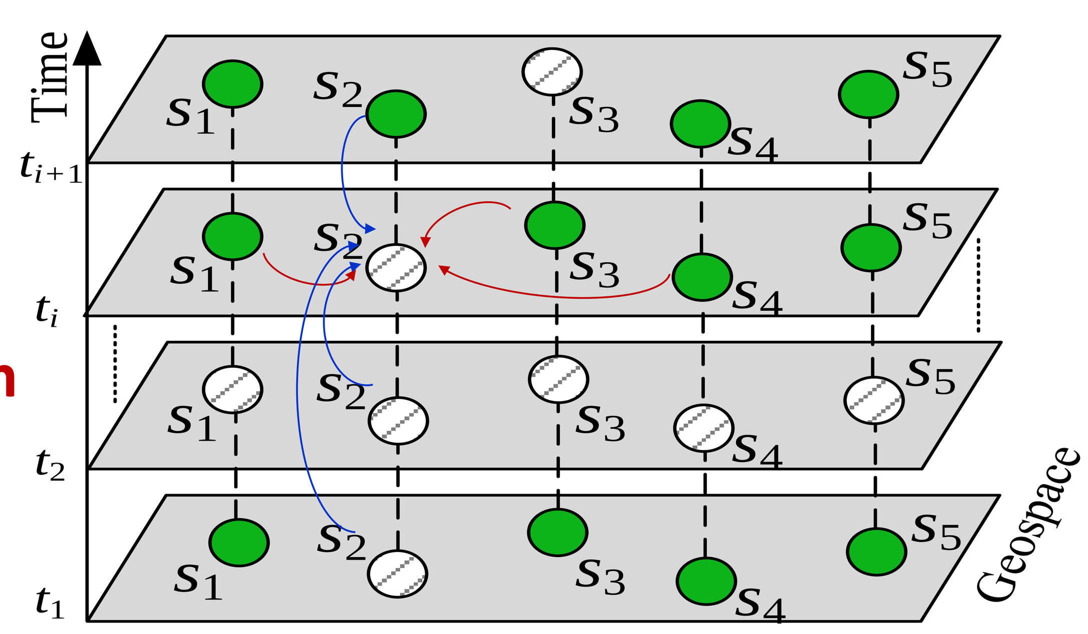
Data missing is a very common phenomenon in LOT data

- Due to communication or device errors
- Affect real-time monitoring and further data analytics

Goal

Filling the missing values in a collection of geo-sensory time series data using collective information:

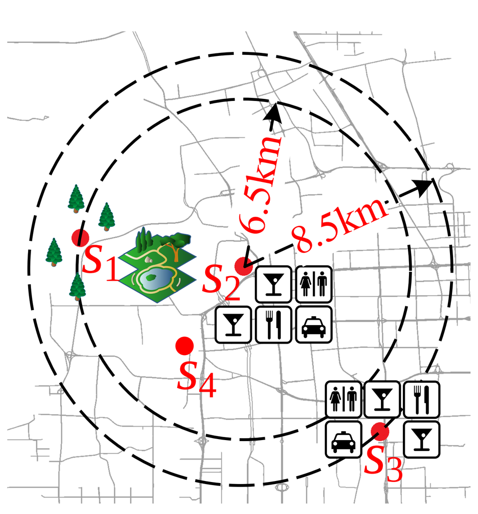
- Data of a sensor
- its neighborhoods



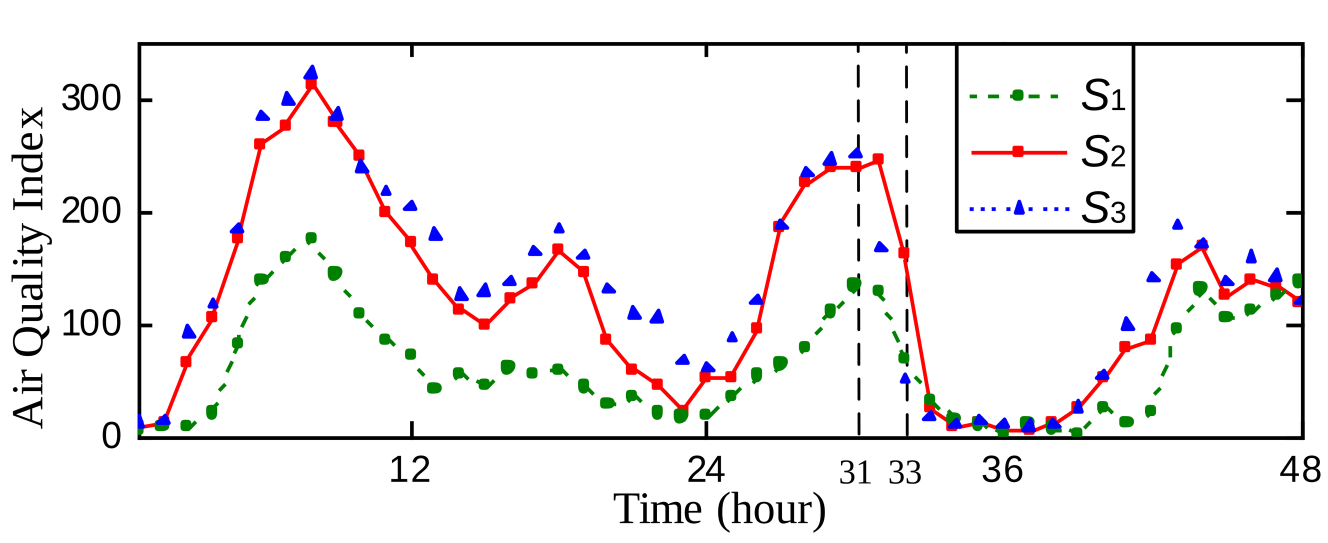
A fundamental problem

Challenges

- Random missing and block missing
  - Lose readings of multiple sensors simultaneously
  - Lose readings of a sensor at consecutive timestamps
  - Hard to find stable inputs for a model
- Readings changing over time and location non-linearly
  - Not handled by simple interpolations

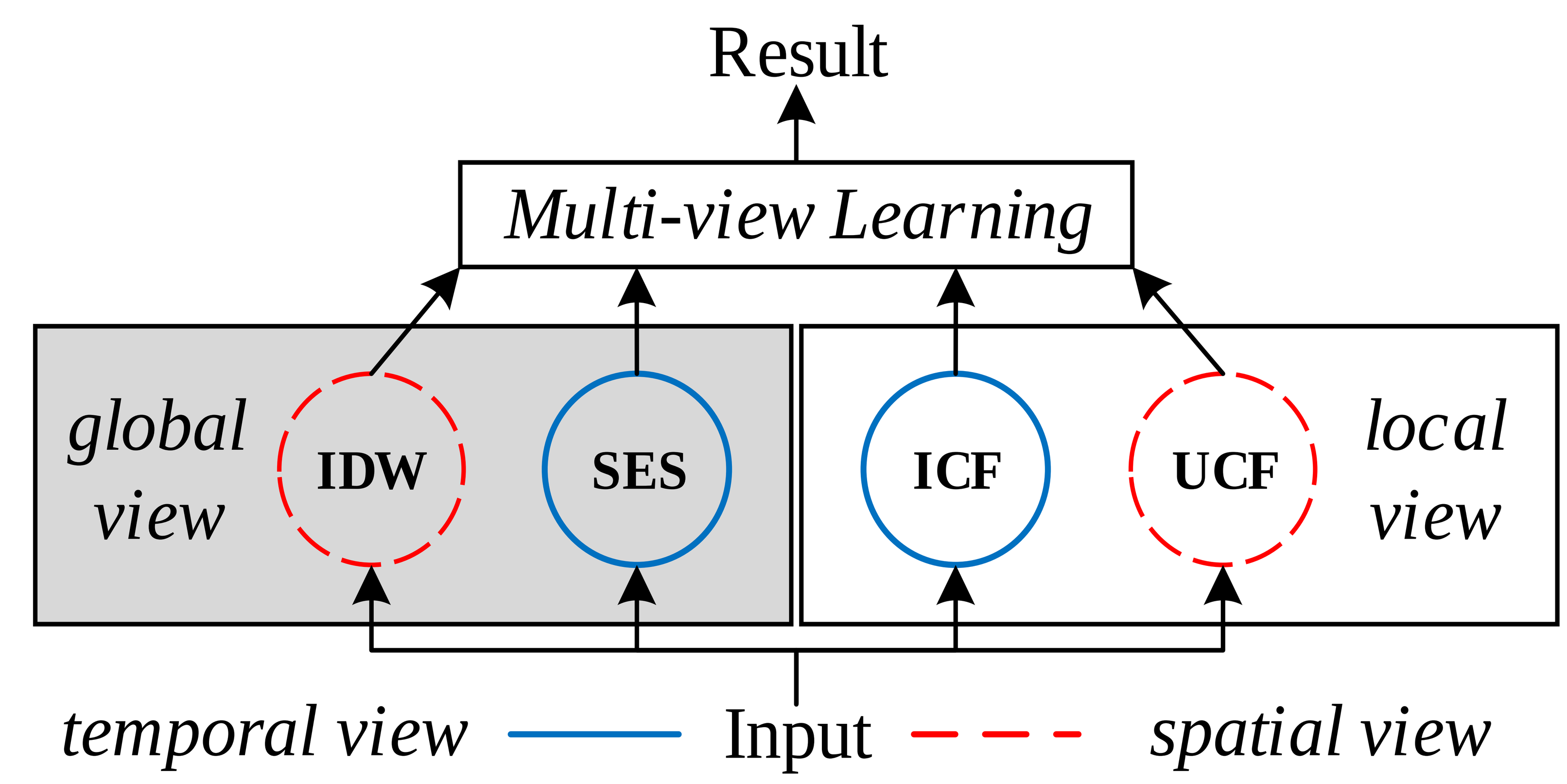


A) Geo-location of sensors



B) Air quality index over time

Methodology

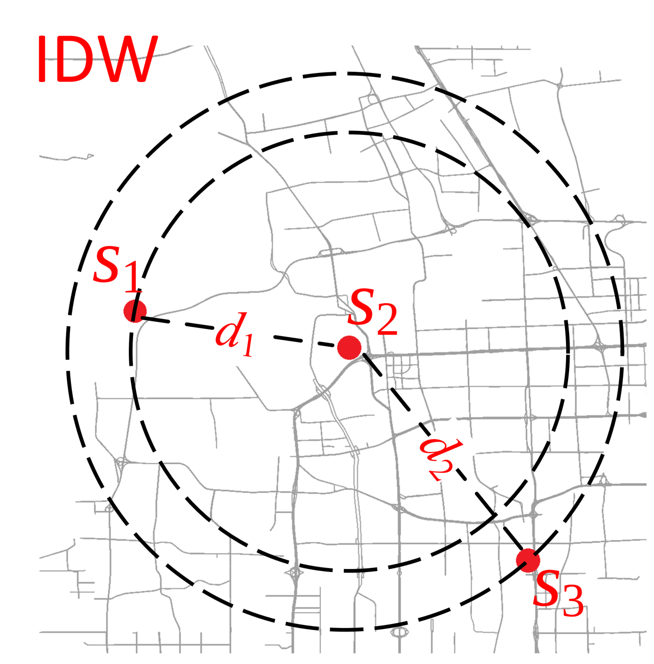


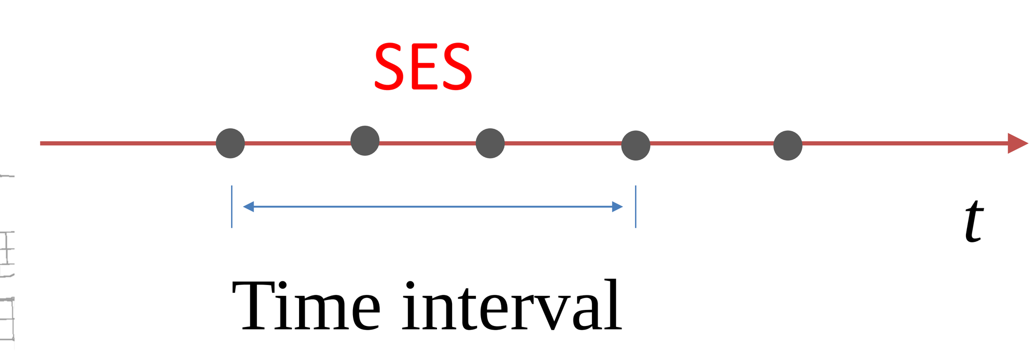
Global spatial view: Inverse Distance Weighting (IDW)

Global temporal view: Simple Exponential Smoothing (SES)

Local spatial view: User-based Collaborative filtering (UCF)

Local temporal view: Item-based Collaborative filtering (ICF)





UCF: Sensors → users

ICF: Time intervals → items

Multi-view learning

- Four views combination by linear least square

$$\hat{v}_{mvl} = w_1 * \hat{v}_{gs} + w_2 * \hat{v}_{gt} + w_3 * \hat{v}_{ls} + w_4 * \hat{v}_{lt} + b$$

Overview

Integrating four perspectives

- Spatial and Temporal perspectives
  - Spatial neighbors
  - Temporally adjacent time intervals
- Global and local perspectives
  - Global: Long-term patterns
  - Local: Recent context

	t_1	t_2	……	t_{j-2}	t_{j-1}	t_j	t_{j+1}	t_{j+2}	……	t_{n-1}	t_n
s_1	230	230	….	205	164	185		188	….	223	249
s_2	200	188	….	173	136	X	146	185	….	199	255
s_3	118	93	….	72	56	59	44	78	….	99	111
⋮			⋮						⋮		
s_m	121	102	….	60	30	40	33	56	….	88	106

Temporal

Spatial

Local

Global

Evaluation

Datasets

		PM2.5	NO_2	Humidity	Wind Speed
Block missing	Spatial	2.2%	3.9%	9.8%	11.8%
	Temporal	3.5%	6.5%	9.6%	19.5%
General missing		8.2%	6.8%	4.6%	4.0%
Overall		13.3%	16.0%	21.5%	30.3%

Baselines

Method	Spatial	Temporal	Spatial + Temporal
Global	IDW	SES	IDW+SES
Lobal	UCF	ICF, ARMA	CF, NMF, stKNN
Global+Local	Kriging	SARIMA	AKE, DESM, NMF-MVL

Comparison among different methods (based on PM2.5)

Method	General Missing		Spatial Block Missing		Temporal Block Missing		Sudden Change		Overall	
	MAE	MRE	MAE	MRE	MAE	MRE	MAE	MRE	MAE	MRE
ARMA	22.61	0.331	29.26	0.369	\	\	51.11	0.567	27.47	0.394
Kriging	15.53	0.221	\	\	15.62	0.222	42.32	0.407	16.59	0.234
SARIMA	14.69	0.220	23.92	0.319	31.20	0.561	52.80	0.586	18.76	0.278
stKNN	12.84	0.188	19.91	0.235	12.72	0.226	35.13	0.390	14.00	0.201
DESM	13.65	0.191	19.24	0.233	12.66	0.224	42.87	0.425	15.59	0.228
AKE	13.34	0.195	19.08	0.229	12.14	0.22	41.54	0.403	14.27	0.211
IDW+SES	11.64	0.171	18.25	0.215	11.95	0.213	34.33	0.381	12.70	0.183
CF	12.20	0.178	19.27	0.234	12.25	0.218	34.91	0.388	13.40	0.193
NMF	11.21	0.163	18.98	0.239	12.73	0.217	34.37	0.381	13.08	0.188
NMF-MVL	11.16	0.162	18.97	0.238	12.66	0.217	34.33	0.380	13.06	0.187
ST-MVL	**10.81**	**0.158**	**17.85**	**0.217**	**11.71**	**0.208**	**33.15**	**0.368**	**12.12**	**0.174**