Spatial Transfer Learning for Estimating $PM_{2.5}$ in Data-poor Regions

Shrey Gupta[1], **Yongbee Park**[4], Jianzhao Bi[2], Suyash Gupta[3], Andreas Züfle[1], Avani Wildani[1,5], Yang Liu[1]
[1] Emory University, USA, [2] University of Washington, USA, [3] University of California, Berkeley, USA, [4] Ingkle, South Korea, [5] Cloudflare, USA.

EMORY UNIVERSITY

INTRODUCTION

Particulate Matter 2.5 (PM 2.5):
- Aerosol with size < 2.5 μm.
- Causes respiratory and cardiovascular illnesses.
- Caused due to vehicles, wildfires, etc.

Limitations in PM 2.5 data collection:
- **Remote-sensing Data:** inaccurate due to weather-related factors.
- **Ground-sensor Data:** equipment is costly to install, maintain and scale.

Transfer Learning (TL) to the Rescue!! But …
- Previous TL models are forecasting models.
- TL models don't account for spatial dependencies.
- Do not account for semantic dependencies.
- Poor performance on unknown test locations.

GOAL

Design a solution to achieve spatial transfer learning such that it accounts for spatial and semantic dependencies, can predict on unknown test locations as well as perform nowcasting.

DATASETS AND MODELS

Target Dataset Regions
a) California-Nevada (US) [128 sensors] [2]
b) Lima (Peru) [10 sensors][3]

Source Dataset Regions
a) Eastern US
b) North-eastern US

Transfer Learning Scenarios
- **a) Simulated Transfer:**
  - **a) Target:** California-Nevada; **Source:** Eastern US
  - **b) Target:** California-Nevada; **Source:** North-eastern US
- **b) Real-world Transfer:**
  - **a) Target:** Lima; **Source:** Eastern US

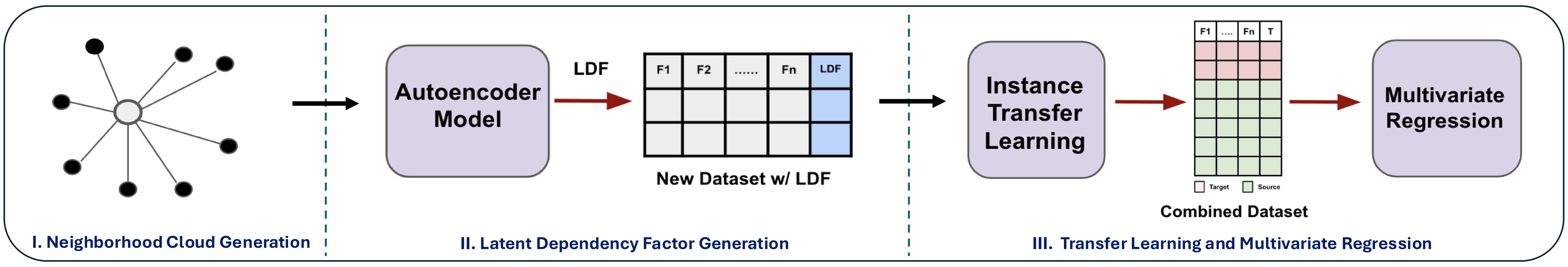
Experimental Setup
Simulated Transfer:
 # Target sensors: [5,7,9, 11]
Real-world Transfer:
 # Target sensors: 10

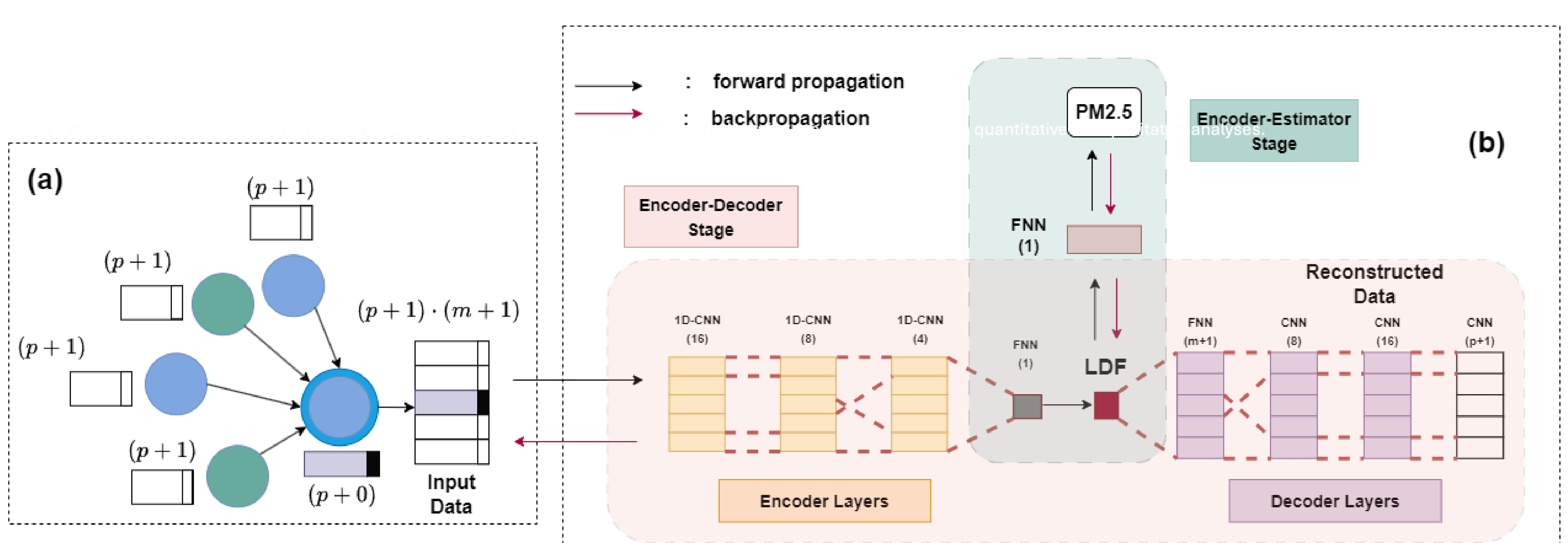
Regression Models:
a) Random Forest (RF)
b) Gradient Boosting (GBR)

Transfer Learning Models:
a) Nearest Neighbor Weighing (NNW)
b) Kullback-Leibler Importance Estimation Procedure (KLIEP)
c) Kernel Mean Matching (KMM)
d) Fully-connected NN (FNN)

NNW, KLIEP and KMM are Instance Transfer models and FNN is Parameter Transfer model [1].

METHODOLOGY





Neighborhood Cloud

Two-stage Autoencoder for LDF Generation

Neighborhood Cloud Dataset: 12 neighbors with (p+1) features selected based on similarity to objective location. Clustered dataset has size (p+1) x (12 + 1).

Two-Stage Autoencoder

STAGE I: Encoder-Decoder
- Encoder summarizes input data to generate latent value. Decoder employs backpropagation.
- Both have 3 1-D CNN layers w/ varying filter size.

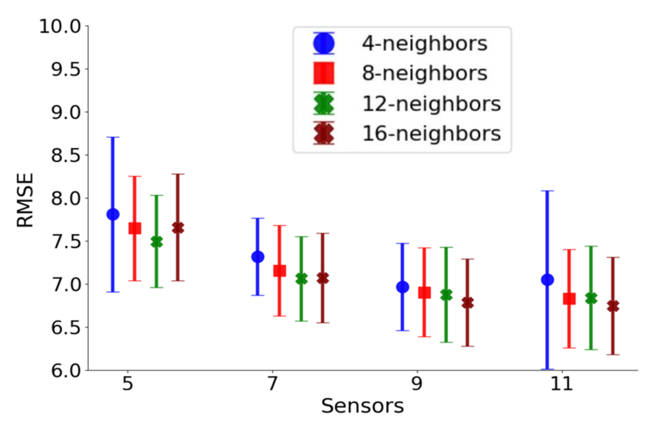
STAGE II: Encoder-Estimator
- Increases attention on $PM_{2.5}$ labels
- Has 1 FNN layer w/ 1 weigh + bias
- Utilizes $PM_{2.5}$ value of objective location.

The two stages alternate training over epochs.

LDF-A: Uses $PM_{2.5}$ and Aerosol Optical Depth (AOD) in the Encode-Estimator stage.

Optimal 'k' Neighbors

- 12 neighbors selected from set {4, 8, 12, 16}.
- NNW is used for experiments.
- Similar results with other transfer models.



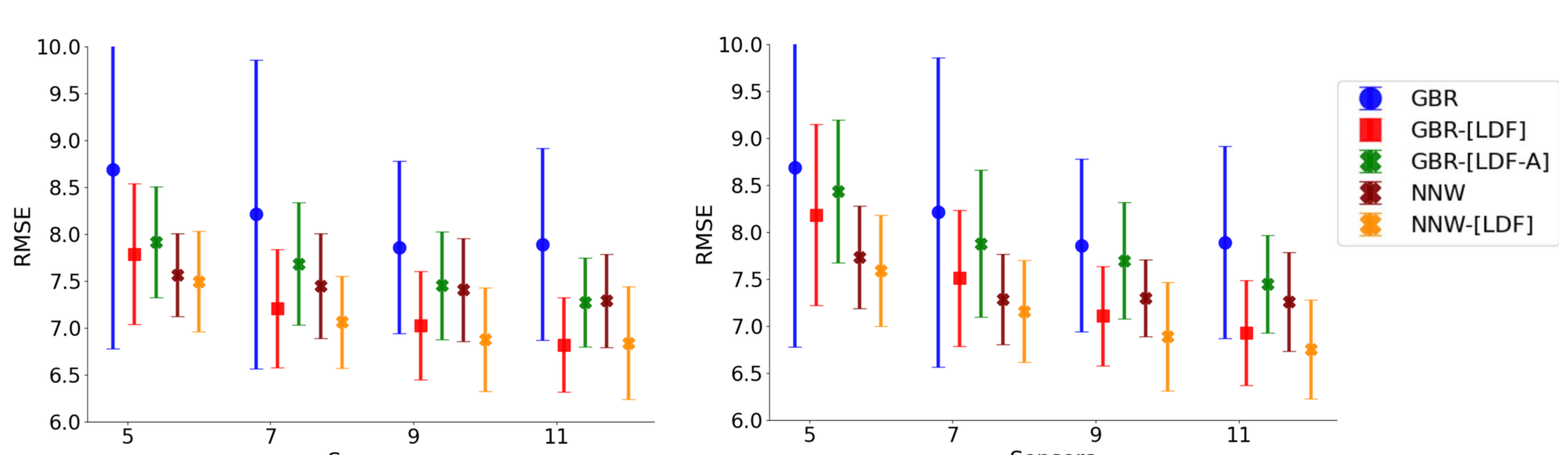
QUANTITATIVE RESULTS

COMPARISON OF BEST PERFORMING BASELINES W/ & W/O LDF FEATURES								
Sensors →	5		7		9		11	
Models	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE
Gradient Boost	-0.061	8.684	0.064	8.210	0.177	7.857	0.157	7.891
SOURCE: EASTERN US								
NNW	0.236	7.563	0.263	7.447	0.280	7.406	0.296	7.288
NNW [LDF]	**0.247**	**7.494**	**0.336**	**7.061**	**0.378**	**6.874**	**0.378**	**6.838**
NNW [LDF-A]	0.225	7.596	0.298	7.230	0.359	6.973	0.359	6.924
SOURCE: NORTH-EASTERN US								
NNW	0.199	7.732	0.294	7.286	0.301	7.297	0.298	7.257
NNW [LDF]	0.225	7.592	**0.317**	**7.157**	**0.376**	**6.886**	0.392	6.751
NNW [LDF-A]	**0.201**	**7.702**	0.320	7.122	0.378	6.873	**0.374**	**6.847**

- For the ML models (GBR, RF), only target data samples are utilized.
- GBR performs the best among ML models.
- NNW performs the best among all transfer models w/ and w/o LDF feature.

Ablation Study

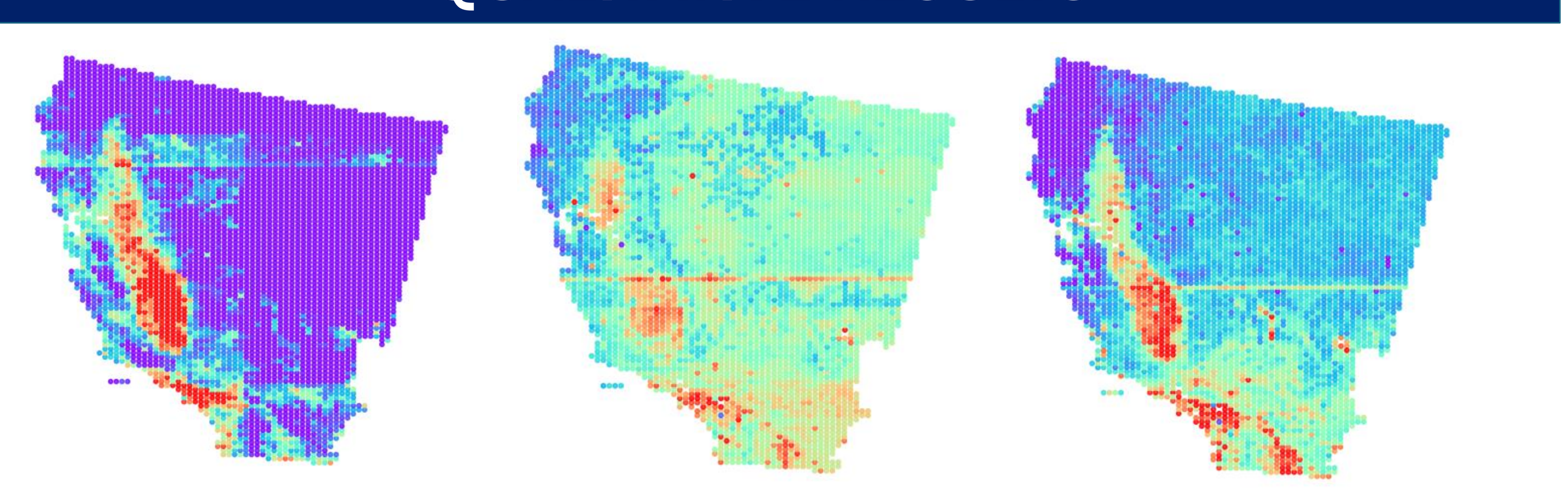
GBR [LDF] show performance improvement, however, NNW [LDF] still outperforms it indicating that LDF is suited for instance transfer models.



Source: Eastern US

Source: North-Eastern US

QUALITATIVE RESULTS



GBR (Ground Truth)

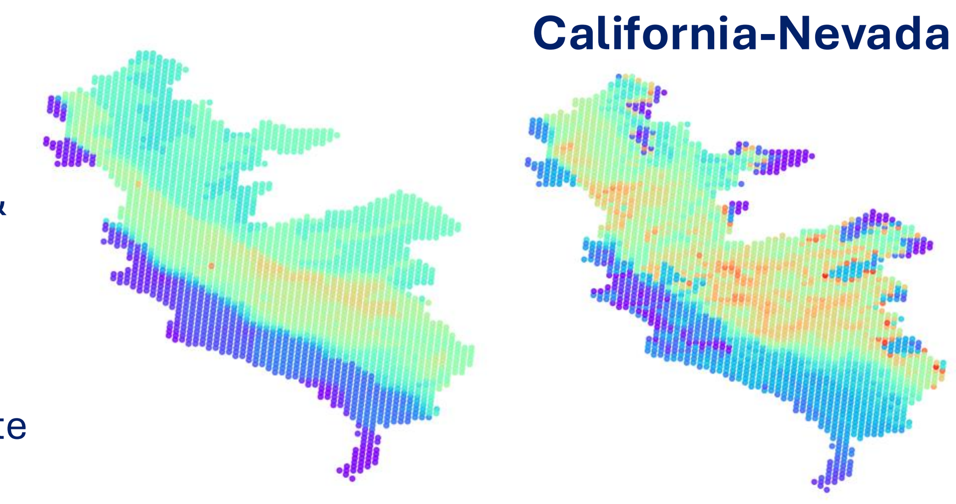
NNW

NNW [LDF]

California-Nevada

California-Nevada
- NNW [LDF] has the most optimal $PM_{2.5}$ estimation.
- Other models have obscure & patchy estimation.

Lima
NNW [LDF] has the most accurate concentration gradient of $PM_{2.5}$ estimation in inland and the Andes mountain region.



GBR

NNW [LDF]

Lima

CONCLUSION

The Latent Dependency Factor (LDF) feature improves the prediction accuracy for transfer learning models by 19.34% over the baseline models.

References:
[1] Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345-1359.
[2] Park, Y., Kwon, B., Heo, J., Hu, X., Liu, Y., Moon, T.: Estimating pm2. 5 concentration of the conterminous united states via interpretable convolutional neural networks. Environmental Pollution 256, 113395 (2020)
[3] Vu, B.N., Sánchez, O., Bi, J., Xiao, Q., Hansel, N.N., Checkley, W., Gonzales, G.F., Steenland, K., Liu, Y.: Developing an advanced pm2. 5 exposure model in lima, peru. Remote sensing 11(6), 641 (2019)

Checkout the Github repository for the paper.

E: yongbee.park@ingkle.com
E: shrey.gupta@emory.edu
Website : Scan here