Multimodal Geospatial Representation Learning

Xinran Liu [1]

Abstract

The goal of this project is to learn geospatial embeddings from multimodal regional features and evaluate their effectiveness in downstream spatial interpolation tasks. We construct a synthetic dataset by simulating wind fields and air pollution concentration over the contiguous United States. A graph transformer–based self-supervised model is used to learn region-level embeddings from these multimodal inputs. The learned embeddings are then used to predict two downstream targets: environmental burden and respiratory disease risk. Experimental results show that the learned embeddings capture meaningful spatial and multimodal structure, leading to improved prediction performance compared to baseline methods.

1. Results and Discussions

Table 1 shows the regression performance at both county and postal code scales. The proposed method consistently outperforms the baseline methods (kNN and IDW) across all targets and metrics. This indicates that the learned embeddings capture meaningful regional structure beyond simple geographic proximity. An average performance over 5 runs is reported. Evaluation metrics with specific seeds are documented in Evaluation/results.csv.

The current spatial network can be extended to a spatio-temporal setting by introducing separate spatial and temporal attention modules. At each time step t, a *graph spatial attention module* extracts spatial features from the graph, and a *graph temporal attention module* extracts temporal features at the same time step via temporal self-attention across the time dimension to model long-range dependencies at each spatial location. *Optimal Transport* is then used to align the spatial and temporal feature distributions, and the aligned features (optionally concatenated with the original spatial and temporal features) are used to produce the final node embeddings.

[1]Department of Computer Science, Vanderbilt University contact: xinran.liu@vanderbilt.edu.

Table 1. Regression performance at postal code and county levels (mean$^{\pm \text{std}}$ over 5 seeds: [0, 1, 2, 3, 4]).

Scale	Target	Method	MAE ↓	RMSE ↓	R^2 ↑
County	Resp.	kNN	$0.1628^{\pm 0.0061}$	$0.2222^{\pm 0.0097}$	$0.7705^{\pm 0.0163}$
		IDW	$0.1601^{\pm 0.0058}$	$0.2171^{\pm 0.0090}$	$0.7793^{\pm 0.0142}$
		Ours	$\mathbf{0.0915}^{\pm 0.0053}$	$\mathbf{0.1179}^{\pm 0.0067}$	$\mathbf{0.9353}^{\pm 0.0081}$
	Env.	kNN	$0.1712^{\pm 0.0085}$	$0.2321^{\pm 0.0128}$	$0.7831^{\pm 0.0181}$
		IDW	$0.1681^{\pm 0.0078}$	$0.2266^{\pm 0.0120}$	$0.7947^{\pm 0.0144}$
		Ours	$\mathbf{0.1136}^{\pm 0.0074}$	$\mathbf{0.1452}^{\pm 0.0094}$	$\mathbf{0.9152}^{\pm 0.0104}$
Postal	Resp.	kNN	$0.0609^{\pm 0.0009}$	$0.0897^{\pm 0.0017}$	$0.9598^{\pm 0.0013}$
		IDW	$0.0617^{\pm 0.0007}$	$0.0900^{\pm 0.0015}$	$0.9595^{\pm 0.0008}$
		Ours	$\mathbf{0.0345}^{\pm 0.0015}$	$\mathbf{0.0455}^{\pm 0.0017}$	$\mathbf{0.9896}^{\pm 0.0007}$
	Env.	kNN	$0.0637^{\pm 0.0009}$	$0.0920^{\pm 0.0018}$	$0.9624^{\pm 0.0013}$
		IDW	$0.0648^{\pm 0.0009}$	$0.0929^{\pm 0.0019}$	$0.9616^{\pm 0.0012}$
		Ours	$\mathbf{0.0386}^{\pm 0.0013}$	$\mathbf{0.0508}^{\pm 0.0014}$	$\mathbf{0.9885}^{\pm 0.0006}$

To incorporate modality awareness, the same encoder can operate on multimodal inputs, producing modality-aware spatial and temporal feature distributions by splitting dimensions. Optimal Transport alignment and concatenation are then applied in the same manner, resulting in embeddings that jointly preserve spatial, temporal, and modality-specific structure.

2. Synthetic Data Generation

2.1. Regions

We construct regional graphs where vertices correspond to geographic locations specified by longitude and latitude coordinates, and edges connect each vertex to its 6 nearest neighbors based on the Haversine distance, i.e., the great-circle distance that respects the spherical geometry of the Earth. Additional state-id information is incorporated thanks to the GeoPandas library (Jordahl et al., 2020). The vertex locations are 1) **randomly sampled** from a uniform distribution over the longitude range [-125,-66] and latitude range [25,49], and subsequently 2) **filtered** to retain only those within the contiguous United States (CONUS) boundary. For the sample sizes, we consider two scales (separately):

- At county level. Since CONUS contains approximately 3,000–3,100 counties and county equivalents, we initially draw 5,500 random location samples using seed 0. After filtering by the CONUS boundary, this results in 3,213 vertices.

- At postal code level. We similarly sample 50,000 locations, resulting in 28,922 vertices after filtering by the CONUS boundary. Figure 1 illustrates the resulting region graph.

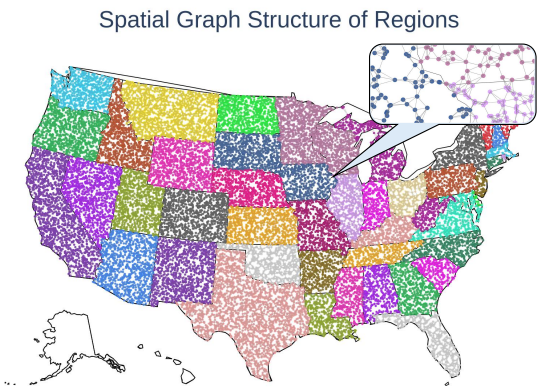

Figure 1. Randomly sampled region graph mimicking postal code–level spatial resolution in the contiguous United States (CONUS). The graph contains 28,922 vertices and 86,345 edges constructed using 6-nearest neighbor connectivity.

2.2. Multi-modal Features

Features are generated on a regular longitude–latitude grid covering the longitude range [-125,-66] and latitude range [25,49]. Each location is associated with a 4-dimensional feature vector, consisting of a 3-dimensional block representing the wind field and one dimension representing pollution concentration.

- **Wind Field.** The wind field (Figure 2) is generated using a synthetic stream function constructed from random spherical harmonic expansions, producing smooth, spatially coherent flow patterns. Large-scale circulation, mid-latitude turbulence, and latitude-dependent zonal winds are incorporated to mimic realistic atmospheric dynamics, with additional perturbations introduced near coastlines. Wind velocity vectors are obtained from the gradient of the stream function, with a small divergent component and reduced magnitude over land to simulate surface friction. The resulting wind field is represented using three features: the zonal velocity component u, the meridional velocity component v, and the corresponding wind speed magnitude $\sqrt{u^2 + v^2}$.

- **Pollution Concentration.** The pollution level (Figure 3) is generated by randomly placing 2,000 emission sources across the CONUS region and modeling their contributions using Gaussian kernels based on great-circle distance, producing smooth concentration patterns with localized peaks. To account for atmospheric transport, pollution levels are attenuated in regions with stronger wind speeds using an exponential wind-dependent penalty. The resulting concentration is then normalized to $[0, 1]$.

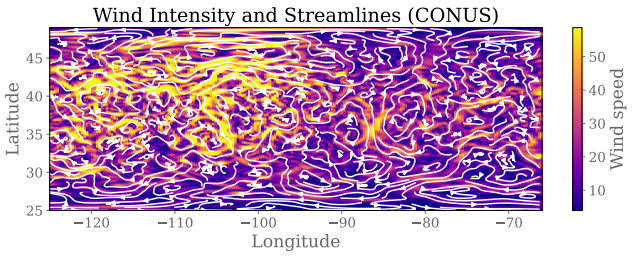

Figure 2. Wind features over CONUS. Colors indicate wind speed and streamlines show flow direction.

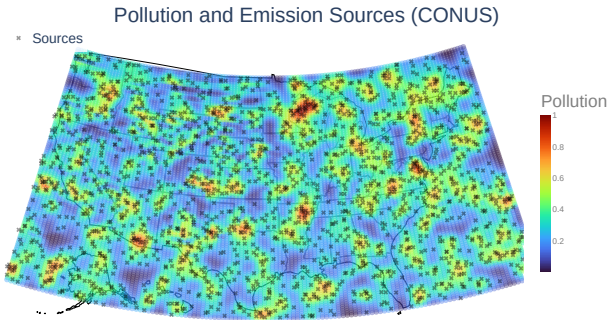

Figure 3. Pollution feature over CONUS. Colors show concentration and crosses indicate emission sources.

After generating features on the regular longitude–latitude grid, we use linear interpolation to evaluate the feature fields at the regional locations defined by the previously constructed graph.

2.3. Regression Targets

Two regression targets are designed to reflect environmental exposure and health risk processes. Both targets are constructed using pollution concentration, wind speed, a smooth latent spatial field and noise.

Let $a_i \in [0, 1]$ denote the normalized pollution concentration, $w_i \in [0, 1]$ the normalized wind speed magnitude, and $s_i \in [0, 1]$ the smooth spatial field at region i.

Environmental Exposure Burden primarily reflects the physical transport and accumulation of pollutants. Its formulation combines local pollution concentration, wind-driven transport, and geographic baseline variability, requiring models to jointly reason about pollution and wind features as well as their interaction. The interaction term $a_i w_i$ ensures that exposure depends not only on local emissions but also on atmospheric transport, which introduces nontrivial

cross-modal dependence. Therefore $y^{(1)}_{\text{region}}$ is defined as

$$y^{(1)}_{\text{region},i} = 2.5\, a_i + 2.5\, w_i + 1.0\, a_i w_i + 0.8\, s_i + 0.2\, \epsilon^{(1)}_i, \quad (1)$$

where $\epsilon^{(1)}_i \sim \text{Gamma}(2.0,\, 0.05)$ represents environmental variability.

Respiratory Health Risk represents a downstream impact that depends on both cumulative exposure and nonlinear physiological response. This target incorporates nonlinear transformations, threshold effects, and dependence on the first target, mimicking real-world dose–response relationships and cascading environmental effects. As a result, accurate prediction requires capturing higher-order structure and shared latent spatial factors. And $y^{(2)}_{\text{region}}$ is defined as

$$y^{(2)}_{\text{region},i} = 0.5\, y^{(1)}_{\text{region},i} + 1.2\, a_i^2 + 1.0\, \sqrt{w_i} + 0.1 \quad (2)$$

$$s_i + \max(a_i - \tau,\, 0) + 0.1\, \epsilon^{(2)}_i, \quad (3)$$

where $\epsilon^{(2)}_i \sim \text{Gamma}(2.0,\, 0.05)$ and τ is the 70th percentile of pollution concentration.

The final regression vector for each region is

$$\mathbf{y}_{\text{region},i} = \left(y^{(1)}_{\text{region},i},\; y^{(2)}_{\text{region},i}\right). \quad (4)$$

Note that wind and pollution are intentionally given larger weights than the spatial field because pollution concentration and wind-driven transport are the primary physical mechanisms governing environmental exposure, whereas geographic location mainly contributes indirect baseline effects.

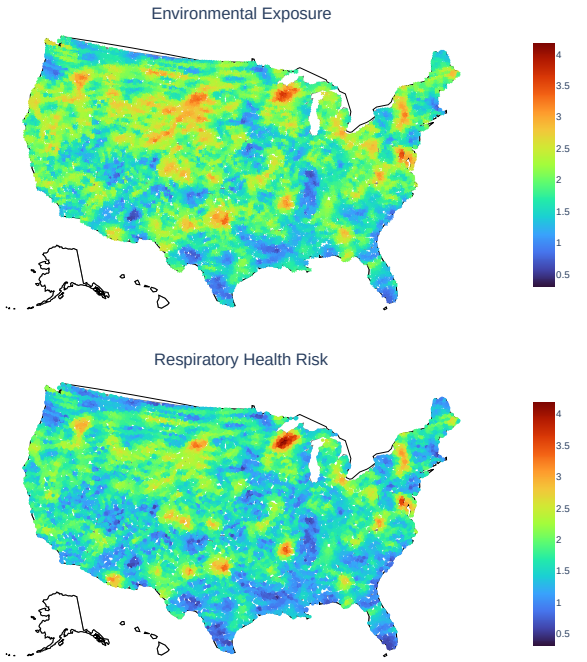

Figure 4. Regression targets. Both targets share common spatial patterns because respiratory risk is partly driven by environmental exposure and pollution transport. However, they are not identical: respiratory risk additionally incorporates nonlinear dose–response effects, threshold behavior at high pollution levels, and additional spatial variability, resulting in distinct local differences. This creates a realistic setting where targets are correlated through shared environmental mechanisms while retaining task-specific structure.

3. Self-Supervised Model Pretraining

Model Architecture and Modality Awareness. We adopt a masked graph autoencoder with a shared Graph Transformer (Shi et al., 2021) encoder to learn spatial representations in a self-supervised manner. The model operates on the regional graph, where each node is associated with multimodal environmental features, geographic coordinates, and a state identifier. The state identifier is mapped to a learnable embedding to provide regional context. The encoder consists of 6 Graph Transformer layers with residual connections, layer normalization, and dropout. The encoder produces a shared embedding for each node, which is explicitly partitioned into a 128-dimensional shared embedding, consisting of a 64-dimensional wind embedding and a 64-dimensional air quality embedding. Separate modality-specific decoders are used to reconstruct features from their respective embedding components.

Training with Corruption Mechanism. During training, features at 30% of locations are randomly masked or corrupted, and the model is trained to reconstruct the masked values. We use Huber loss as the reconstruction objective.

4. Downstream Evaluations

Data are randomly split into 70% training, 10% validation, and 20% test sets. To evaluate embedding quality, we train an MLP with three hidden layers on top of the learned embeddings for regression. Input normalization is performed separately for each split to avoid leakage. We compare against k-nearest neighbors (kNN) and inverse distance weighting (IDW) baselines, both using the Haversine distance computed from geographic coordinates. Performance is evaluated using mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2).

5. AI Assistance and Disclosure

ChatGPT was used to assist with drafting code comments, documentation, and preliminary test files. All such content was thoroughly reviewed, corrected, and validated by the author. The research ideas, methodology, implementation, experiments, and conclusions were conceived, executed, and verified by the author.

References

Jordahl, K., den Bossche, J. V., Fleischmann, M., Wasserman, J., McBride, J., Gerard, J., Tratner, J., Perry, M., Badaracco, A. G., Farmer, C., Hjelle, G. A., Snow, A. D., Cochran, M., Gillies, S., Culbertson, L., Bartos, M., Eubank, N., maxalbert, Bilogur, A., Rey, S., Ren, C., Arribas-Bel, D., Wasser, L., Wolf, L. J., Journois, M., Wilson, J., Greenhall, A., Holdgraf, C., Filipe, and Leblanc, F. geopandas/geopandas: v0.8.1, July 2020. URL https://doi.org/10.5281/zenodo.3946761.

Shi, Y., Huang, Z., Feng, S., Zhong, H., Wang, W., and Sun, Y. Masked label prediction: Unified message passing model for semi-supervised classification. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 1548–1554. International Joint Conferences on Artificial Intelligence Organization, 2021.

A. Computational Resources

All experiments were conducted on a single NVIDIA RTX A6000 GPU. The available resources were sufficient to train the proposed models without requiring distributed training or specialized hardware.

While the hardware did not impose strict limitations, it informed several design choices aimed at improving efficiency and reproducibility. In particular, we used moderately sized embedding dimensions, which reduced training time and memory usage while preserving model performance. No large-scale hyperparameter sweeps or extensive multi-seed experiments requiring cluster-scale resources were performed.