



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

异构多源空间点数据预测的自洽深度几何学习



2024/9/29



I Introduction

引言部分为读者提供了研究问题的背景、当今面临的挑战、以及文章提出的解决方案



I Introduction

研究背景

- 在环境预测、自然资源管理、交通规划等各个领域，空间预测至关重要。在空间预测中，需要收集不同的属性，并利用这些属性来拟合目标变量的预测模型。
- 在一个经典的PM2.5预测中，可以采用空气质量监测站(AQMS)提供观测数据。虽然AQMS提供高质量数据，但其部署和维护成本高，导致在广泛地区覆盖不足。为了弥补这些覆盖空白，常见的做法是部署大量低成本微型传感器。
- 这些不同来源收集的属性各不相同，某些污染物，如非甲烷碳氢化合物，由空气质量监测系统专门监测。为了充分利用每个源的独特属性，需要一种专门的多源空间数据预测方法。

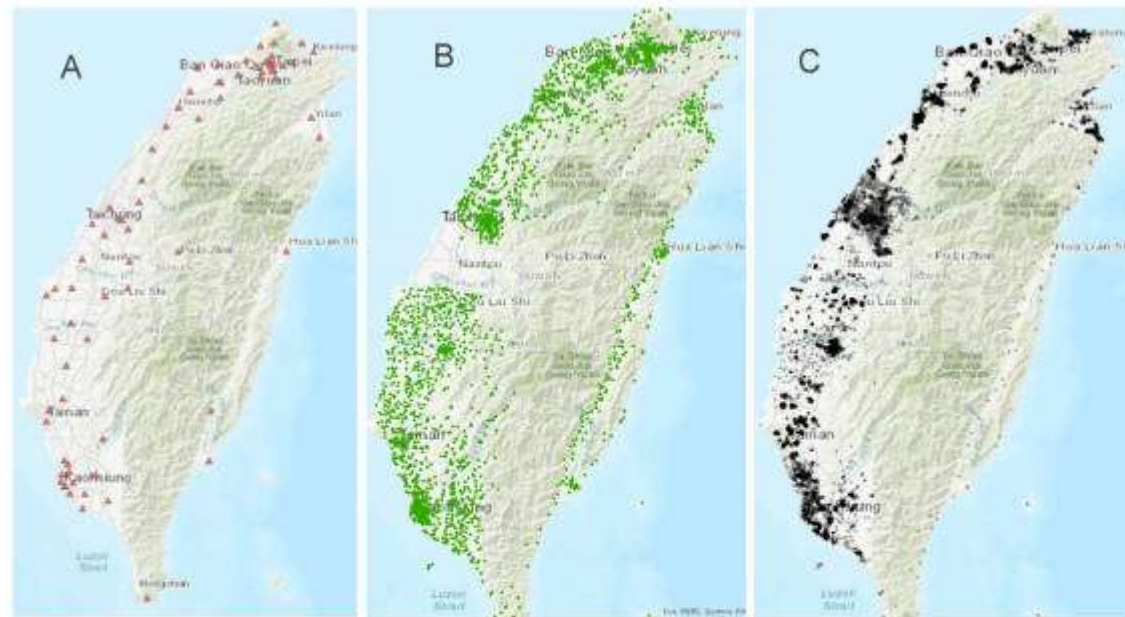


Figure 1: An example of multi-source spatial point prediction problem: Varying distribution of three data sources including (A) 74 AQMSs, (B) 3704 LASS AirBox sensors, and (C) 9701 EPA MicroStations.

aqms、中央研究院维护的AirBox传感器，和台湾环保局的空气质量微站数据



I Introduction

当前面临的挑战

- 挑战一：在没有地面真实数据的情况下，如何有效地跨数据源对齐信息？
- 挑战二：如何在不同数据源质量差异的情况下对齐信息？
- 挑战三：如何在不同空间位置对齐信息？

解决方法

- 提出一种自监督策略，通过最大化模型估计和每个数据源的目标变量之间的互信息来对齐数据,这种方法旨在以自我监督的方式调整来自所有来源的数据，利用它们的集体优势来提高预测准确性
- 引入“保真度分数”概念，这是一个可学习的参数，用于量化每个数据源的质量。该参数不受约束，设计为可学习的，允许模型使用梯度下降算法动态调整和量化每个数据源的质量。
- 开发一个地理感知的多源图神经网络，专门处理不同数据源之间的空间关系和特征异构性的复杂性。



II Related work

在相关工作部分，文章回顾了多源空间点数据预测领域的三种不同模型



II- Related work

■ 传统模型

基于领域知识的理论模型，通常应用于空气污染物建模等领域。虽然这些模型具有可解释性，但其计算需求较高，且对不同领域的适应性较差。

例如，使用气象原理模拟污染物的化学和物理过程。一些著作[1]利用气象原理和数学方法模拟污染物的化学和物理过程，便于预测。虽然这些模型有坚实的理论基础支撑，但它们对某些领域的特殊性对不同情景的适应性造成了重大限制

■ 基于高斯过程(GP)的模型

这些模型通过邻近样本的加权平均来估计未知数据点。尽管GP方法在数据源之间存在强相关性时表现良好，但其假设过于简化，难以捕捉复杂的非线性交互。最近的进展，如NARGP[28]，将机器学习与GP相结合，以增强灵活性和拟合性，代表了多源预测问题中GP方法的最新进展。



■ 机器学习模型

相比传统和GP模型，机器学习方法更灵活且计算效率更高。随着越来越复杂的架构，机器学习模型已经变得善于捕捉复杂的数据相关性。随机森林等方法常用于多源空间问题，虽然有效，但在显式学习跨数据源空间点关系时表现较弱。



III Methodology

在方法论部分主要介绍了：

- ✓ 自我监督方式细节
 - ✓ 保真度评分具体内容
 - ✓ 深度多源预测框架 (DMSP) 复杂细节
- 
- 

III-A 问题表述

1. 地理空间表示

文中将二维地理空间中的点表示为 $s \in \mathbb{R}^2$ 。作者假设有 N 个不同的数据集 $\{D^{(i)}\}_{i=1}^N$ 每个数据集对应于不同的数据源（例如：地面站或传感器）

2. 地理位置与数据集

每个数据集 $D^{(i)}$ 由一组唯一的地理位置 $s^{(i)} = \{s_j^{(i)}\}_{j=1}^{n_i}$ 构成，其中 n_i 是数据集中位置的数量。这种设置允许数据集从不同地区捕捉多样的环境信息。

3. 目标变量

数据集 $D^{(i)}$ 中的每个地理位置 $s_j^{(i)}$ 者 $(x_j^{(i)}, y_j^{(i)})$ 和 p_i 关，其中 \in 表示辅 p_i 属性（例如温度、风速），表 $y_j^{(i)}$ 这些属性的数量。 $Y^{(i)}$ 观测值 被视为随机变量 \tilde{y} 样本，该变量是实际环境变量 的近似值。

4. 预测问题

作者将多源空间点预测问题定义为：给定多个数据集 $\{D^{(i)}\}_{i=1}^N$ ，目标是预测二维地理空间中的 $\tilde{y} \sim \tilde{Y}$ 量。也就是说，模型必须为该地理空间中的任意位置预测环境状况。

5. 目标

主要目标是最大化预测值 Y 与实际值 \tilde{y} 之间的互信息（mutual information）。通过此互信息，模型能够确保预测值与真实值的高度一致，从而实现准确的预测。

1. 地理空间表示:

文中将二维地理空间中的点表示为 $s \in \mathbb{R}^2$ 。作者假设有 N 个不同的数据集 $\{D^{(i)}\}_{i=1}^N$ ，每个数据集对应于不同的数据源（例如，地面站或传感器）。

2. 地理位置与数据集:

每个数据集 $D^{(i)}$ 由一组唯一的地理位置 $S^{(i)} = \{s_j^{(i)}\}_{j=1}^{n_i}$ 构成，其中 n_i 是数据集中位置的数量。这种设置允许数据集从不同地区捕捉多样的环境信息。

3. 目标变量:

数据集 $D^{(i)}$ 中的每个地理位置 $s_j^{(i)}$ 都与样本 $(x_j^{(i)}, y_j^{(i)})$ 相关联，其中 $x_j^{(i)} \in \mathbb{R}^{p_i}$ 表示辅助属性（例如温度、风速）， p_i 表示这些属性的数量。目标观测值 $y_j^{(i)}$ 被视为随机变量 $Y^{(i)}$ 的样本，该变量是实际环境变量 \tilde{y} 的近似值。

4. 预测问题:

作者将多源空间点预测问题定义为：

给定多个数据集 $\{D^{(i)}\}_{i=1}^N$ ，目标是预测二维地理空间中的目标变量 $y \sim \tilde{y}$ 。也就是说，模型必须为该地理空间中的任意位置预测环境状况。

5. 目标:

主要目标是最大化预测值 Y 与实际值 \tilde{y} 之间的互信息（mutual information）。通过此互信息，模型能够确保预测值与真实值的高度一致，从而实现准确的预测。

III-B 多数据源自适应自监督融合

- 目标：目标是最大化互信息实现数据自适应性融合
- 方法：在缺乏真实数据的情况，通过将某些传感器的高质量观测数据视为真实标签，模型通过如下互信息公式进行优化

- 互信息公式：

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

权重 w 是广义互信息的权值，其是基于每个传感器的准确观测概率来计算的

- 经过推导，互信息最大化本质上转化为对数条件似然期望的加权和最大化：

$$\begin{aligned} \max \sum_{i=1}^N C_i MI(Y^{(i)}, \hat{Y}) &= \sum_{i=1}^N C_i \left(H(Y^{(i)}) - H(Y^{(i)} | \hat{Y}) \right) \\ \Leftrightarrow \max - \sum_{i=1}^N C_i H(Y^{(i)} | \hat{Y}) &= \sum_{i=1}^N C_i \mathbb{E}_{p(y^{(i)}, \hat{y})} \log p(y^{(i)} | \hat{y}), \end{aligned}$$

III-B 多数据源自适应自监督融合

■ p分布的学习方式：

方程中的实际条件分布 p 未知，我们可以让预测模型学习一个近似真实条件分布 p 的分布 q 。根据KL散度的非负性，我们可以证明当估计的条件分布与实际分布一致时，条件熵最小。

Theorem 4.1. *For two random variables $Y^{(i)}, \hat{Y} \in \mathbb{R}$, any variational approximation of the conditional distribution $p_{Y^{(i)}|\hat{Y}}$ will increase the conditional entropy $H(Y^{(i)}|\hat{Y})$.*

■ 训练流程：

使用自监督策略，逐步优化每个数据源的可信度分数，提升预测性能。其中优化目标为：

$$\min \sum_{i=1}^N C_i \mathcal{L}_i(Y^{(i)}, \hat{Y}),$$

其中 $\mathcal{L}_i(Y^{(i)}, \hat{Y})$ 为两个随机变量的定制损失函数。此公式可根据问题的具体要求，灵活地假设任何适当的噪声分布。

III-C 保真度分数

自洽训练过程的核心是将互信息目标转化为适合自监督训练的可计算函数。尽管没有直接的真值 (ground truth) 数据，但可以使用所有数据源的目标变量观测值来训练模型。每个数据源都有一个与之相关的质量权重，称为保真度分数 (fidelity score)，这个分数是作为可学习的参数。

训练步骤：

- ① 对于数据集中的每个样本 $(x_j^{(i)}, y_j^{(i)} | s_j^{(i)})$ 其中是 $x_j^{(i)}$ 特征, $y_j^{(i)}$ 是目标变量。
- ② 将目标变量掩蔽，得到部分掩蔽的数据集 $D(i)$ 。
- ③ 使用提出的 DMSP 框架，根据数据集的集合 预测掩蔽的目标变量。
- ④ 根据预测计算损失 L 。
- ⑤ 更新保真度分数和 DMSP 的参数。

III-C 保真度分数

自洽训练过程细节：

Algorithm 1: Single epoch training procedure

Input: $\{D^{(i)}\}_{i=1}^N$, initialized $\{C_i\}_{i=1}^N$, initialized Φ_W ,
Learning rate α .
Output: $\{C_i\}_{i=1}^N$, Φ_W .



```
1 for  $i \in \{1, \dots, N\}$  do
2   for  $j \in \{1, \dots, n_i\}$  do
3     Get  $(x_j^{(i)}, y_j^{(i)} | s_j^{(i)})$ ;
4      $D'_i \leftarrow D^{(i)}$  with  $y_j^{(i)}$  masked;
5      $\hat{y} \leftarrow \Phi_W(D^{(1)}, \dots, D'^{(i)}, \dots, D^{(N)})$ ;
6      $\mathcal{L} \leftarrow C_i \mathcal{L}_i(\hat{y}, y_j^{(i)})$ ;
7     for  $C \in \{C_1, \dots, C_N\}$  do
8        $C \leftarrow C - \alpha \frac{\partial \mathcal{L}}{\partial C}$ ;
9     end
10     $W \leftarrow W - \alpha \frac{\partial \mathcal{L}}{\partial W}$ ;
11  end
12 end
13 return  $\{C_i\}_{i=1}^N$ ,  $\Phi_W$ 
```

- 第1行遍历所有数据集。
- 第2行遍历当前数据集的所有样本。
- 第3行获取 j -th示例。
- 第4行屏蔽了示例中的目标值，以获得一个时态数据集。
- 第5行使用提议的DMSP框架预测被屏蔽的目标值。
- 第6行计算训练损失。
- 第7行到第10行更新DMSP的保真度分数和参数



IV Experiment

在实验部分主要介绍了：

- ✓ 数据集来源，对比方法，实验指标
 - ✓ 实验结果
 - ✓ 消融实验
- 
- 

IV Experiment

数据集:

- SouthCalAir:来自美国环境保护署(EPA)空气质量系统(AQS)的26个传感器和515个PurpleAir 传感器在2019年1月1日至2019年12月31日在南加州的细颗粒物(PM 2.5)数据。
- NorthCalAir: 该数据来自美国环境保护局(EPA)空气质量系统(AQS)的63个传感器和北加州1110个PurpleAir传感器。
- Spatially correlated regression (SCR):模拟空间相关数据生成的合成数据集, 目标变量中加入高斯正态噪声来模拟低质量的数据源。特征被不可逆地转换, 以避免退化为简单的回归任务。
- Flu: 此数据集集整合了来自两个不同来源的数据:2010年至2015年疾病控制与预防中心(CDC)和谷歌流感趋势计划, 数据可靠但是范围有限。

对比方法:

- SRA-MLP: 逐步回归分析与多层感知器神经网络相结合的模式。
- NARGP: 基于高斯过程回归和非线性自回归方案的概率框架
- GeoPrior: 在以地理位置为条件的概率估计方法。
- Space2Vec: 一种对位置绝对位置和空间关系进行编码的表示学习模型。

评价指标:

平均绝对误差 (MAE)、均方根误差 (RMSE)、解释方差分数 (EVS)、决定系数 (CoD) 以及皮尔逊相关系数 (Pearson)

IV Experiment

实验结果:

Table 1: The performance of the proposed model (including ablation variants) and the comparison methods.

Dataset	Method	MAE	RMSE	EVS	CoD	Pearson
SouthCalAir	SRA-MLP	<u>3.211±0.059</u>	<u>5.305±0.091</u>	<u>0.416±0.017</u>	<u>0.411±0.015</u>	<u>0.686±0.012</u>
	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	<u>3.000±0.059</u>	<u>5.374±0.378</u>	<u>0.404±0.026</u>	<u>0.390±0.024</u>	<u>0.636±0.021</u>
	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	<u>0.616±0.045</u>	<u>0.773±0.059</u>	<u>0.457±0.024</u>	<u>0.446±0.025</u>	<u>0.698±0.012</u>
	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

IV Experiment

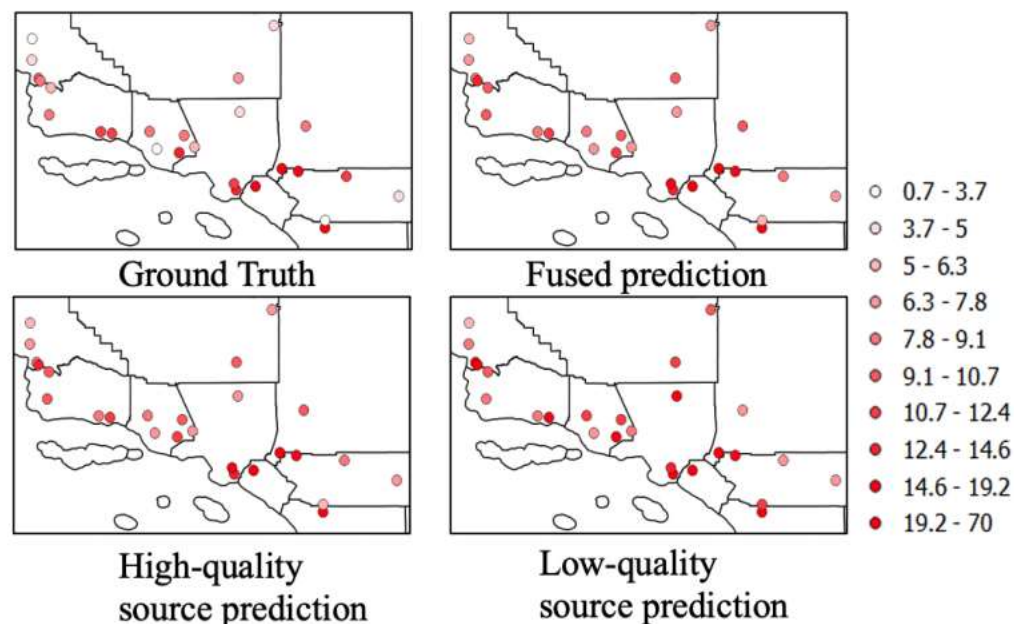
消融实验:

- **单一数据源的贡献**: 研究了仅使用单一数据源时的表现, 结果表明单一数据源不能提供足够的信息来达到最优性能。
- **多个数据源的平等处理**: 去除忠实度分数, 将所有数据源平等对待, 结果表明多个数据源的融合效果显著, 尤其在低质量数据源较多的空气污染数据集中表现突出。
- **不同的空间嵌入方法**: 比较了使用其他两种空间嵌入方法 (GeoPrior和Space2Vec) 的表现, 结果显示DMSP的空间感知GNN能更好地处理空间点数据预测。

可扩展性分析:

测试了DMSP在处理不同样本规模时的运行时间, 结果显示模型在样本数量增加时, 运行时间呈线性增长, 表明该方法具有良好的可扩展性。

SouthCal和SCR数据集高质量数据源的对比



V Conclusion



V Conclusion

- 本研究提出了一种新的多源空间点数据预测框架。该框架的核心是一个自我监督的训练目标，该目标不仅可以对齐来自各种来源的数据，还可以在没有任何真实数据的情况下利用它们的集体力量。保真度评分定量地评估每个数据源的质量。
- 此外，地理位置感知图神经网络还能熟练地处理不同位置之间复杂的空间关系。
- 此方法的有效性已经在合成和真实世界的多源数据集上进行了严格的测试和验证。结果表明，我们的框架不仅优于现有的最先进的方法，而且还提供了对数据有意义的见解。

谢谢！

