【Café速递】韦晶：中国高分辨率高精度近地表细颗粒物遥感反演研究

一、 核心提示

遥感技术为大范围常规 PMx 监测提供了有效手段，由于影响气溶胶反演的因素比较多，导致当前广泛使用的气溶胶产品存在很大误差；同时从气溶胶大气柱的总含量转换到地面细颗粒物的估算误差来源很多，由此估算得到的 PMx 精度通常较低。而且当前已存在的 PMx 卫星数据集空间分辨率较为粗糙，很难反映中小尺度尤其是城市地区空气污染变化情况，因此也不能满足污染监控和研究的需要。本期报告将介绍如何利用卫星遥感技术估算得到中国长时间序列高空间分辨率高质量 PMx 数据集(ChinaHighPMx)，这对于我国在不同时空尺度上监控空气污染、理解其形成变化规律等都具有重要意义。

主持：修田雨　录像：王克险　文字：韩佳明 审核：修田雨

二、 人物名片

韦晶，北京师范大学和美国马里兰大学联合培养博士，师从李占清教授。截至目前，共发表学术论文 70 余篇，包括以第一或通讯作者在 RSE、ES&T、JGR、ACP 和 TGRS 等国际顶级期刊发表 SCI 论文 20 余篇，总被引 700 余次，H-index 为 16，3 篇入选 ESI 全球热点(TOP<0.1%)论文，4 篇入选 ESI 全球高被引(TOP<1%)论文，1 篇入选 JGR 亮点论文。担任 RSE、ISPRS、JGR、TGRS、EP 和 AE 等 10 余个国际知名 SCI 期刊审稿人；荣获李小文遥感科学青年奖。

三、 报告现场

2020 年 7 月 10 日晚上 9 点，韦晶嘉宾做客 GeoScience Café第 263 期线上分享活动。韦晶嘉宾从大气污染的研究背景出发，介绍了近地表细颗粒物遥感估算的典型算法、中国大气细颗粒物污染时空覆盖变化规律，以及基于多源遥感技术和人工智能方法构建的中国高分辨率质量大气细颗粒物数据集 (ChinaHighPMx)，并对观众的提问进行了详细的解答。

01 研究背景介绍

大气气溶胶来源复杂，人类污染排放、生物质燃烧和野火燃烧等都会向大气排放大量的气溶胶。一般在晴天时，气溶胶浓度较低；然而当大气污染较为严重时，对大气环境及人类健康都会产生非常重要的影响。在气溶胶中，大气细颗粒物，即空气动力学当量直径小于 2.5（$PM_{2.5}$）或者 1 微米（PM_1）的颗粒物，能够到达人的肺部，会对人体健康产生严重影响，已成为当前大气研究中一个热点问题。

针对大气污染，中国政府也颁布了一系列的大气污染防治措施。如 2013 年颁布的大气污染防治行动计划，主要目的是整体改善全国的空气质量，特别是中国京津冀、长三角和珠三角三个典型地区；接着在 2018 年启动了蓝天保卫战计划，目的是大幅度减少大气污染物排放总量，协同减少温室气体的排放，进一步明显降低 $PM_{2.5}$ 的浓度，明显减少重污染天数，从而大大改善环境空气质量，增加人民的蓝天幸福感。

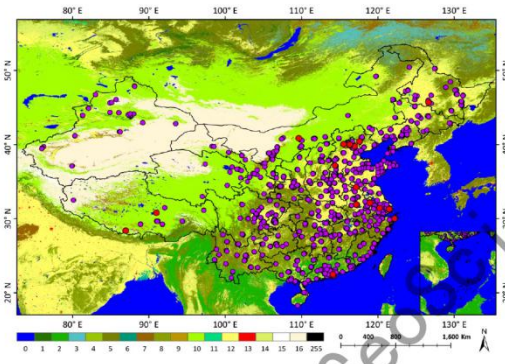
02 近地表细颗粒物遥感估算研究

基于遥感技术的近地表细颗粒物估算研究具有十分重要的意义。由于人为气溶胶的不断

增加，空气污染日益严重，近地表细颗粒物（特别是 $PM_{2.5}$、PM_1）已成为影响城市环境的首要污染物，受到了公众的广泛关注。传统的地基观测由于站点分布稀疏，很难从广尺度实现对大气污染的监测。卫星遥感技术的发展特别是气溶胶遥感为大范围常规 PM 监测提供了有效手段。但由于影响气溶胶反演的因素比较多，同时从气溶胶大气柱的总含量转换到 PM 的估算误差来源复杂，导致了传统方法估算得到的 PM 精度通常较低。已有的 PM 卫星数据集的空间分辨率较为粗糙，很难反映中小尺度区域内尤其是城市地区内空气污染变化情况，因此也不能满足污染监控和研究的需要。因此，获得高分辨率和高精度的 PM 数据，对于我国在不同时空尺度上监控空气污染、理解其形成变化规律等都具有重要意义。

首先韦晶介绍了**时空-随机森林模型**来估算中国 $PM_{2.5}$ 浓度。数据源为中国环保部发布的逐小时 $PM_{2.5}$ 地基观测数据，以及其他遥感和再分析辅助数据，共选取包括 AOD、地表覆盖高程夜间灯光和气象条件在内的 13 个独立变量。

中国环保部逐小时$PM_{2.5}$地基观测数据



到2016年，总共有1480个监测站点在中国东部均匀分布；相反在中国西部分布相对稀疏

其他遥感和再分析等辅助数据

Product	Content	Unit	Spatial Resolution	Temporal Resolution
$PM_{2.5}$	$PM_{2.5}$	ug/m^3	-	Hourly
MCD19A2	MAIAC AOD at 550 nm	-	1 km ×1 km	Daily
MOD13A3	NDVI	-	1 km ×1 km	Monthly
MCD12Q1	Land use cover	-	500 m × 500 m	Annually
VIIRS	Night light	-	500 m × 500 m	Annually
SRTM	Elevation	m	90 m × 90 m	-
ERA-Interim	2m air temperature	K	0.125°×0.125°	6-hour
	Surface pressure	hPa	0.125°×0.125°	6-hour
	10m U wind component	m/s	0.125°×0.125°	6-hour
	10m V wind component	m/s	0.125°×0.125°	6-hour
	Boundary layer height	m	0.125°×0.125°	3-hour
	Total precipitation	mm	0.125°×0.125°	3-hour
	Evaporation	mm	0.125°×0.125°	3-hour
	Relative humidity	%	0.125°×0.125°	3-hour

共选取包括AOD、地表覆盖、高程、夜间灯光和气象条件在内13个独立变量

图一：数据源介绍

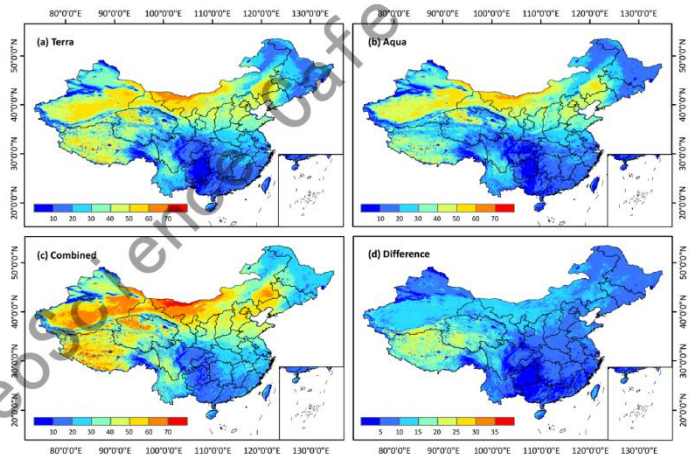
数据预处理方面采用了线性回归平均法对 Terra 和 Aqua 卫星的 AOD 产品进行融合，从而使空间覆盖率提高 15%-20%，有效样本数量提高 25%-32%。接着利用双线性内插法将所有数据的空间分辨率提升到 1 千米。

数据集成

- 利用线性回归平均法融合Terra和Aqua卫星AOD产品
- 空间覆盖率提高15%-20%
- 有效样本数量提高25%-32%

数据重采样

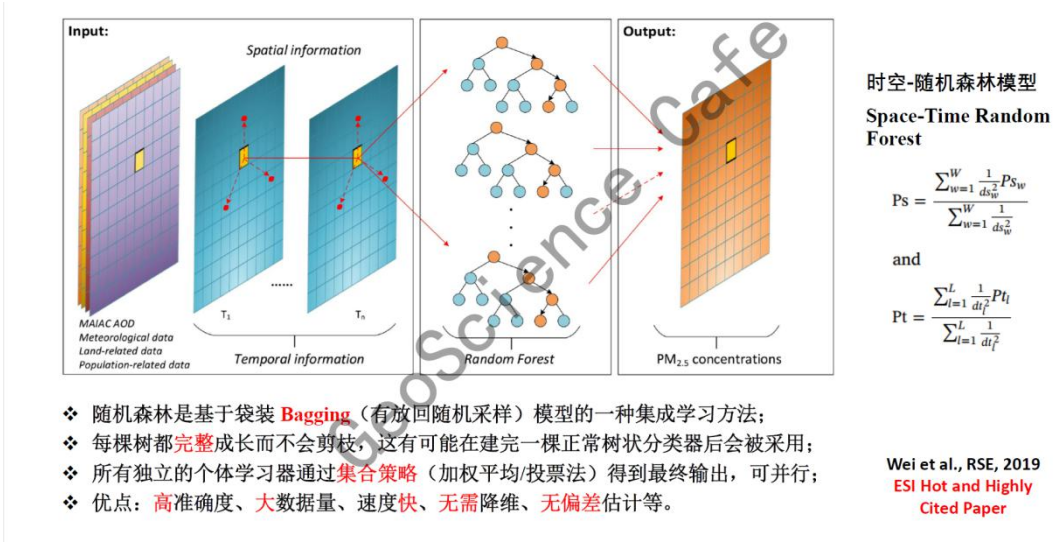
- 选择双向线性内插法将所有粗糙或细致空间分辨率的数据重采样到1千米 (≈ 0.01°×0.01°)



中国Terra、Aqua和合成后的AOD产品空间覆盖率

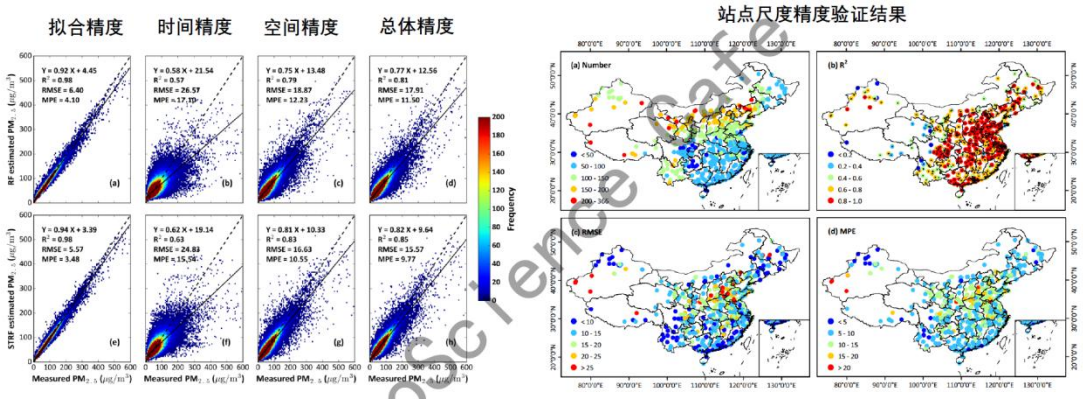
图二：数据预处理

接下来，韦晶介绍了时空-随机森林模型。随机森林模型是目前比较热门的一种方法，被广泛应用于多个领域。随机森林是基于袋装 Bagging 模型的一种集成学习方法，属于机器学习里的一种方法。每棵树都完整成长而不会剪枝；所有独立的个体学习器可通过集合策略与并行处理得到最终输出。其优点在于准确度高、数据量大、速度快、无需降维、无偏差估计等。然而传统的方法并没有考虑大气污染明显的时空变化，因此时空-随机森林模型采用了将时空信息融合在一起的方法，即参考地理时空加权模型，针对空间中的一个像素，同时考虑与它距离最近的几个点及时间序列对它产生的影响。

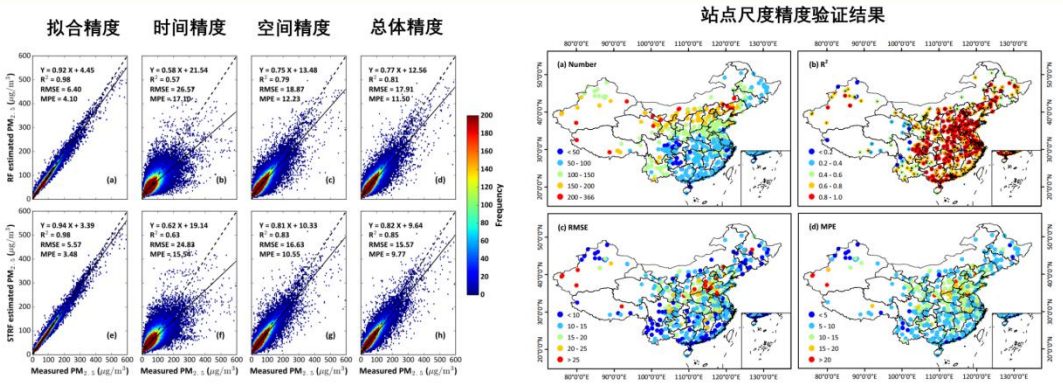


图三：时空-随机森林模型

在随后的精度验证及与传统方法的对比上，证明了时空-随机森林模型的优异性能；同时将空间分辨率提高了 3-10 倍，生产得到了中国 1 公里 $PM_{2.5}$ 数据，。



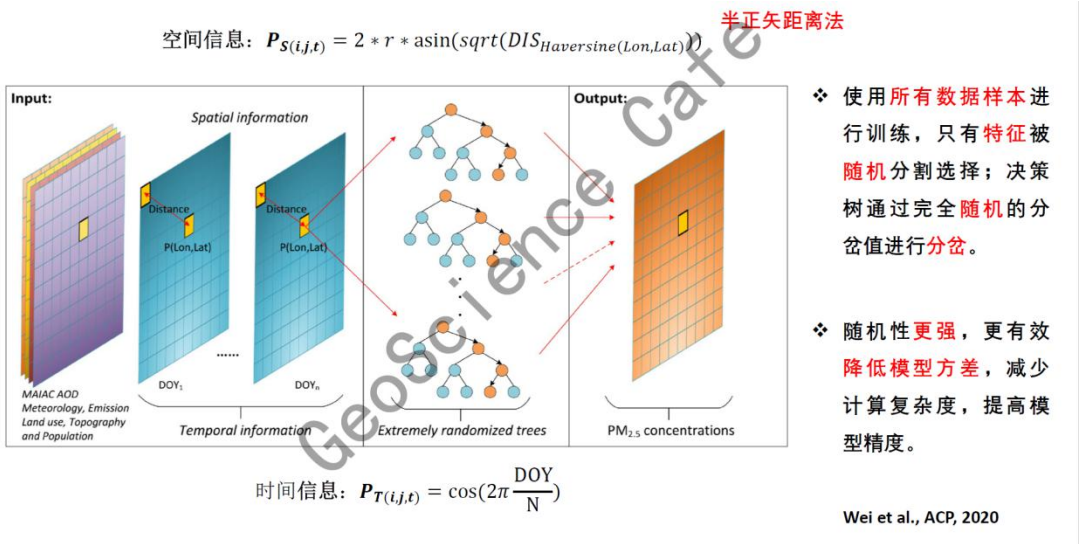
精度验证(十折交叉法)



- STRF模型拟合精度达到0.98，模型不会过拟合；整体优于传统随机森林模型；
- STRF模型具有很好的时空预测能力，整体精度达到0.85；
- ~90%: CV-R^2 > 0.7; ~88%: RMSE < 20 μg/m³; ~93%: MAE < 15 μg/m³

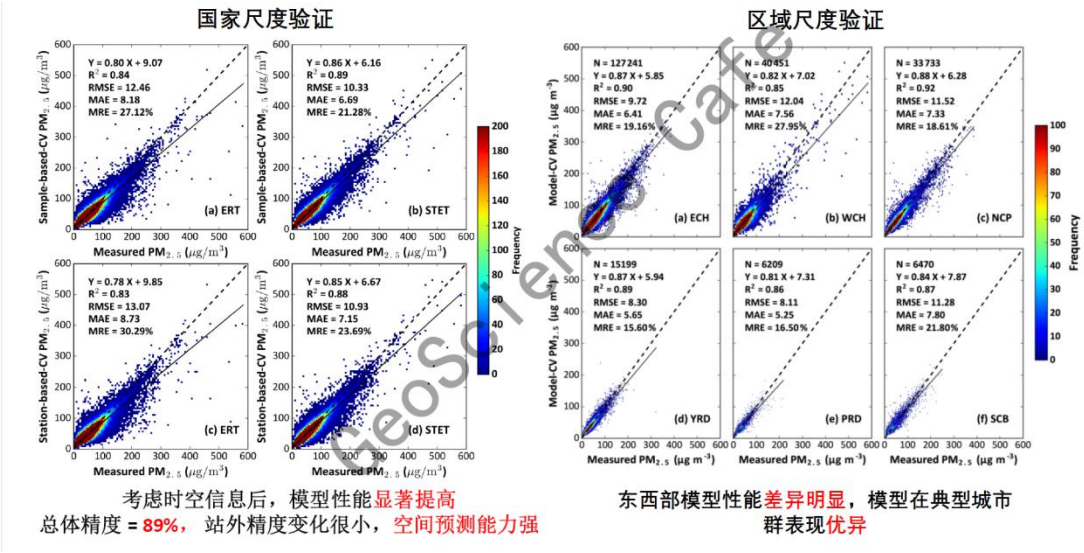
图四：时空-随机森林模型精度验证（十折交叉法）

尽管时空-随机森林模型的初步尝试获得了较高质量的 $PM_{2.5}$ 估算结果，但仍存在一些问题，如时空确定因素还有待提高。因此韦晶接着介绍了发展的新**时空-极端随机树模型**来提高算中国 $PM_{2.5}$ 浓度的估算精度。时空-极端随机树模型相比于时空-随机森林，具有更多的优势。如该模型使用所有数据样本进行训练，只有特征被随机分割选择；决策树通过完全随机的分岔值进行分岔。更强的随机性能更有效的降低模型方差，减少计算复杂度，提高模型精度。



图五：时空-极端随机树模型

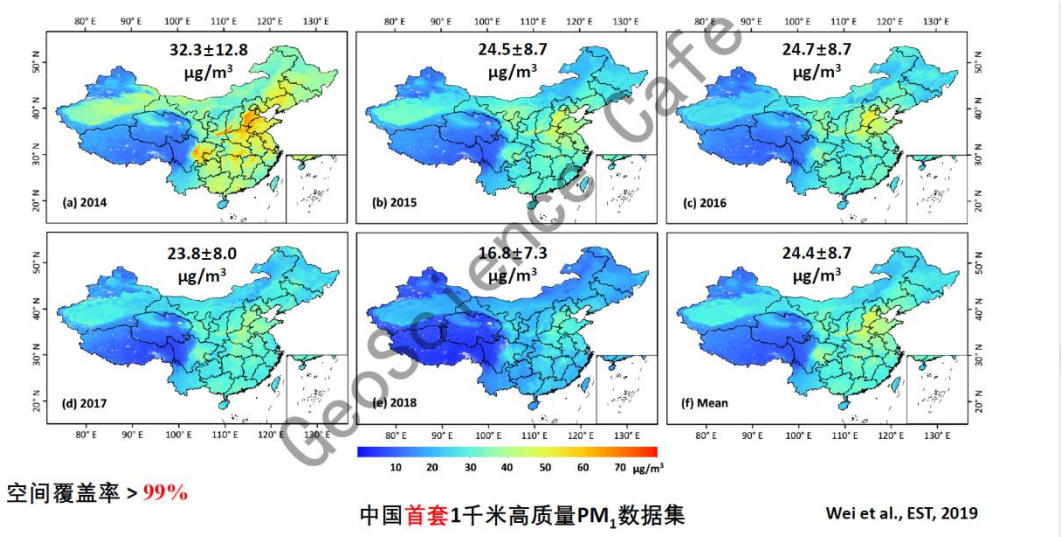
随后韦晶对数据源和模型特征选择进行了详细的介绍，结果表明，时空-极端随机森林模型在考虑时空信息后，模型性能显著提高，总体精度达到 89%，站外精度变化小，空间预测能力强；模型在中国东西部的表现性能存在差异，在典型城市群表现优异。



图六：结果分析与验证

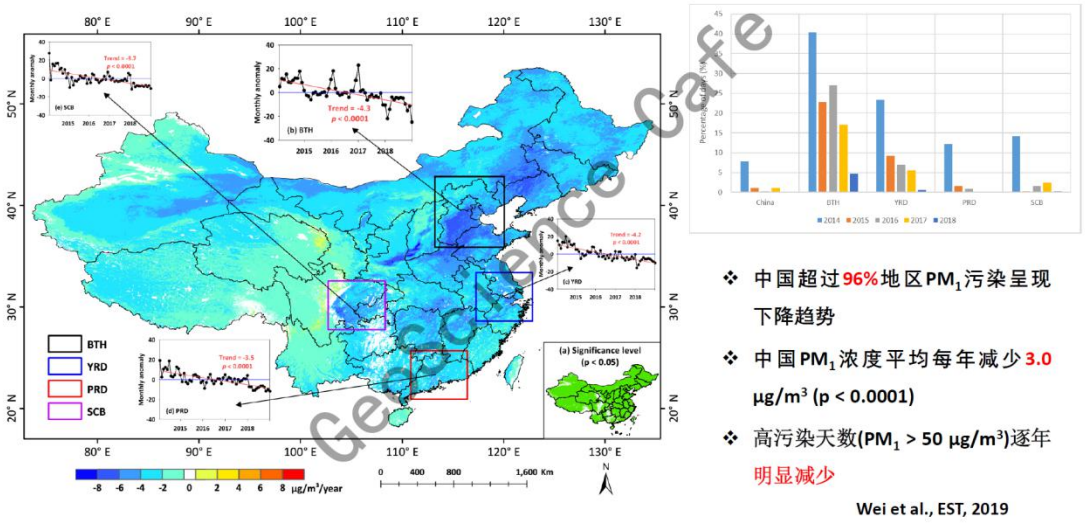
03 中国大气细颗粒物污染时空覆盖变化

而后韦晶介绍了中国 2014 年至 2018 年 PM_1 污染的空间分布与变化。相比于 $PM_{2.5}$, PM_1 的研究较少，但是它对人体的危害非常巨大。基于时空-极端随机树模型，首次生产得到了中国长时间序列 1 千米高质量 PM_1 数据集。从图七可以看到，PM_1 浓度在 2014 年相对较高，而到 2018 年时有明显的降低，这对国家的战略制度进行了一定的验证。



图七：中国 PM1 污染空间分布

时空分析结果发现，近年来，中国超过 96%地区 PM_1 污染呈现下降趋势，中国 PM_1 浓度平均每年减少 3.0ug/m^3，高污染天数逐年明显减少。



图八：中国 PM1 污染变化趋势

04 中国高分辨率质量大气细颗粒物数据集

最后韦晶介绍了团队开发和维护的中国高分辨率高质量大气细颗粒物数据集（ChinaHighPM$_x$），并欢迎相关研究人员免费下载使用。

ChinaHighPM$_x$数据集

中国高分辨率高质量大气细颗粒物数据集(ChinaHighPM$_x$)是一套基于多源卫星遥感技术(MODIS、VIIRS和Himawari-8等)，利用人工智能方法，综合考虑气象条件、地表覆盖和变化、人类分布与污染排放等因素，解决大气污染时空异质特性，生产得到的中国不同粒径近地表细颗粒物遥感数据集，由韦晶博士和李占清教授团队开发与维护。该数据集主要包括PM$_1$、PM$_{2.5}$和PM$_{10}$三种粒径颗粒物数据，空间覆盖整个中国大陆，目前时间覆盖2000–2019年，不断更新；空间分辨率最高可达1千米，时间分辨率为小时/日/月/年。该数据集免费开放，欢迎使用从事相关科学研究。

数据链接：https://weijing-rs.github.io/product.html

图九：ChinaHighPMx 数据集

四、 交流互动

提示：由于篇幅有限、提问内容较多，本节仅选取部分提问内容。

问题一：冬季为什么很多地方没有数据?

韦晶：冬天数据缺失主要分布在中国高纬度或高海拔地区，这些地区冬天主要有长期的冰雪覆盖，在光学遥感中，传统算法没有设计在冰雪表面进行气溶胶反演，导致 AOD 没有数据，因此 PM$_{2.5}$ 也无法进行反演。

问题二：时空随机森林方法对站点密度是否有要求，在站点密度低的地区如中国西部，精度是否较低？
韦晶：是的，中国东西部的精度存在差异，这一方面也是因为不同的站点数目，站点分布越稀疏，训练样本量就越少，同时描述空间信息的能力也越弱。

问题三：多源遥感数据是怎么联合使用，需要进行经过质量一致性处理吗？
韦晶：一般多元遥感数据联合使用的主要问题是时间跟空间上的匹配问题。因此，我们需要把所有数据规整化到相同的分辨率，同时也要把它归一到相同的时间尺度。质量一致性方面，不同团队（如 NASA）在发布遥感产品时，本身就已经做过一些质量控制了。

问题四：在进行不同精度不同分辨率的气溶胶数据的融合的过程中，需要注意哪些方面呢？有什么窍门呢？
韦晶：目前我们这个方法是对 Terra 和 Aqua 两颗卫星的 AOD 产品进行融合。在进行融合的时候，主要考虑的问题是它们的成像时间，一个是上午、一个是下午。所以我们就利用线性回归的方法，对 Terra 和 Aqua 卫星的数据进行融合，这样会降低这方面的误差。对于不同精度不同分辨率的，我暂时还没有做过这方面的实验。目前张良培老师团队有很多学生和老师从事数据融合这方面的研究，你可以看一下他们的相关文章。

问题五：我国北部冬季数据缺失的地区，年均浓度计算时怎么处理呢？
韦晶：我们算年均浓度的时候一般会有一个标准，比如说我们一般都是基于每天算每月的，都是保证在某个像元上有大约超过 10%或者 20%的有效值，我们再进行平均。比如说 MODIS 团队，他们在做月产品的时候，基本也会采用这种方式。

问题六：看到您在建立模型时使用的气象要素没有能见度是吗？因为能见度和颗粒物的相关性比较高，请问这里为什么没有采用能见度建模呢？
韦晶：采用能见度建模也是可以的，这个确实是很好的参考数据，但是目前比较好的能见度遥感产品还是没有的。现在利用能见度数据的研究是挺多的，因为能见度观测的数据可以追溯到一九七几年，当然我们也有一篇文章是基于能见度进行的 $PM_{2.5}$ 反演，这个是可行的。但我在做遥感反演时，因为没有比较好的能见度数据，所以我还没有考虑。如果有相关的遥感产品的话，我们是可以考虑把能见度引入到模型中。

问题七：之前看论文里面涉及到对影像数据进行垂直校正，湿度校正。现在针对这些校正方法有没有什么创新建议？
韦晶：传统的物理方法就是从 AOD 转化到近地表细颗粒物，主要是进行了湿度和垂直高度校正。当然这个方法只是非常早期的。因为影响 $PM_{2.5}$ 因素很多，很难找到一种比较准确的物理模型去表达 $PM_{2.5}$-AOD 之间的关系，所以他们计算的结果也不是特别好。当然我们机器学习方法，通过考虑 BLH、DEM 和 RH，是可以利用数据挖掘这种形式来进行类似的垂直和湿度校正。。

问题八：模型好像融合了 AOD 数据，MODIS 和葵花 8 的 AOD 产品发布时间分别是 1 天和 1 小时，那对模型来说，是不是无法实时预测当天的 PM2.5 浓度呢？
韦晶：是的，MODIS 这种太阳同步轨道卫星的重访周期是一天或者 2 天一次，获得的数据可能就是某个时刻的。如 Terra 就是上午 10:30、Aqua 就是下午 1:30，然后两者融合得到的 PM 数据也处于这个时间段的；葵花八可以逐小时观测。但是因为光学卫星有效观测时间基

本是白天，如上午 8 点到下午 6 点，夜间 PM2.5 反演研究基本上很少。不过现在已经有一些初步研究利用 VIIRS 的夜间波段进行夜间 PM2.5 反演。我想如果利用多元遥感数据的话，是可以预测全天不同时刻的 PM2.5 的。

问题九：因为国控站点主要是在城市区域，其地表反射率特征和广大的农村地区是不同的，那么其训练集是不是主要为建成区反射率特征的数据集，请问本文模型的方法是如何克服因为数据集不充分问题的呢？

韦晶：我们产品主要是服务于城市地区的。因为大城市地区污染较为严重，也是大家较为关注的，而郊区的污染则相对较轻，所以城市不管是环境还是健康都是关注的焦点，我们首先要保证城市地区的模型精度要可靠。当然在农村地区，因为站点分布不均匀可能导致精度有所下降。不过目前中国东部的站点已经是非常密集的了，2013 年的站点主要是城市地区，到 2020 年，已有 1800 多个站点，覆盖了大部分的农村地区。你说的这个问题确实是个不可避免的问题，这对模型构建和反演都会造成一定的影响。但是以后站点肯定是越来越多，现在每年大约要增长 50~100 个观测点，中国对这方面的投入非常大，对我们以后研究会很有帮助。

参考文献：

1. Wei, J., Huang, W., Li, Z., Xue, W., Peng, Y., Sun, L., and Cribb, M. Estimating 1-km-resolution $PM_{2.5}$ concentrations across China using the space-time random forest approach, Remote Sensing of Environment, 2019a, 231, 111221. https://doi.org/10.1016/j.rse.2019.111221
2. Wei, J., Li, Z., Guo, J., Sun, L., Huang, W., Xue, W., Fan, T, and Cribb, M. Satellite-derived 1-km-resolution PM_1 concentrations from 2014 to 2018 across China, Environmental Science & Technology, 2019b, 53(22), 13265-13274. https://doi.org/10.1021/acs.est.9b03258
3. Wei, J., Li, Z., Cribb, M., Huang, W., Xue, W., Sun, L., Guo, J., Peng, Y., Li, J., Lyapustin, A., Liu, L., Wu, H., and Song, Y. Improved 1 km resolution $PM_{2.5}$ estimates across China using enhanced space-time extremely randomized trees, Atmospheric Chemistry and Physics, 2020, 20(6), 3273-3289. https://doi.org/10.5194/acp-20-3273-2020

本期海报回看链接：https://mp.weixin.qq.com/s/BT017i7t8DANucjY1jmqEw

问卷反馈链接：https://www.wjx.cn/jq/74279734.aspx

本期海报预览链接	问卷反馈链接

GeoScience Cafe 以“谈笑间成就梦想”为目标，于**每周五晚 7:00** 在实验室四楼休闲厅，邀请-位嘉宾，为大家带来学术报告或经验分享。报告内容包括摄影测量与遥感、地理信息系统、导航与定位服务、人文社科等研究方向，听众可在报告结束后向嘉宾提问、与嘉宾交流探讨，同时每学期还会举办期求职就业讲座和场导师信息分享会。每期报告会根据嘉宾意愿在站开设**直播**，使不能来到现场的听众同步参与。**报告 PPT** 和**视频回看**会在征得嘉宾同意的情况下在 **QQ 群**和 **B 站**上发布。

更多精彩内容（讲座预告、讲座回顾、报告、报告视频）敬请通过以下方式获取：

公众号：GeoScienceCafe
QQ 群：865459403
Bilibili：geocafe