一 实验背景：

自 2013 年“大气十条”实施以来，全国各地狠抓大气污染治理，我国大气污染防治成效显著。近年来，空气质量监测站收集到大量具有高维、时序特点的空气质量数据，如何利用此类数据，分析理解大气污染传输模式，并为决策者提供有效建议十分具有挑战性。

利用大数据分析技术和可视化方法，能够分析大气污染问题及成因、监测大气污染发展趋势、分析大气污染的地域相关性，快速感知大气污染的时变规律，辅助工作人员因地制宜地制定防治策略。大数据可视分析与可视化将数据智能处理、视觉表征和交互分析有机地结合，使机器智能和人类智慧深度融合、优势互补，为大气污染防治工作的分析、指挥和决策提供有效手段和决策依据。

二 数据选取：

数据利用竞赛提供的 2013－2018 年中国高分辨率大气污染再分析开放数据集（http://naq.cicidata.top:10443/chinavis/opendata），该数据是覆盖全国范围的基于地理空间网格的空气质量再分析数据和对应的气象数据，包括六项常规污染物、风速、温度、气压、相对湿度和经纬度在内的 13 个属性。

2.1 六种大气污染物每日数据

数据集由官方提供，包括 2013 年 1 月 1 日-2018 年 12 月 31 日全国各经纬度区域内的大气基本指标数据（表 1）。

表 1.数据介绍

字段	类型	数据描述	示例
PM2.5	float	细颗粒物，以微克为单位	20.21
PM10	float	可吸入颗粒物，以微克为单位	23.56
SO2	float	二氧化硫，以微克为单位	8.65
NO2	float	二氧化氮，以微克为单位	4.58
CO	float	一氧化碳，以微克为单位	0.29
O3	float	臭氧，以微克为单位	63.44

2.2 四项气象因素每日数据

此数据来源于官方所提供的数据集，包括了 2013 年 1 月 1 日-2018 年 12 月 31 日全国范围内各经纬度区域的四项气象因素数据（表 2）。

表 2.气象数据介绍

数据名称	数据类型	数据描述	示例
风速	float	单位为 m/s。包括具体风速大小和风向。	[-2.48，-2.31]
温度	float	表示各地区的每日温度，单位为 K。	291.54
湿度	float	表示各地区的每日相对湿度，指单位体积空气中，实际水蒸气的分压与相同温度和体积下水饱和蒸气压的百分比，用百分数表达。	68.14
大气压强	float	表示各地区间的每日压强，单位为 Pa。	100350.11

三 任务分析：

- 大气污染时空分布模式分析，大气污染时空演变态势
- 大气污染源分析，识别主要污染物
- 城市大气污染类型及分布
- 大气污染预测，预测大气污染发展趋势
- 大气污染物的健康效应

四 问题与挑战：

- 大气相关数据量大：竞赛提供的的数据每一个日数据 CSV 里有 4 万余条数据，共 2013 到 2018 六年时间；
- 时间跨度广： 时序数据，从 2013 到 2018 六年；
- 数据维度多 ： 13 列高维数据
- 专业要求较高：需要一定的大气污染物相关知识
- 系统的交互设计

五 数据处理：

使用高德地图开放 API 对数据集中的经纬度做逆地理编码，(总共 4 万多坐标点点，每天可以查 5000 条)，获得每对经纬度坐标对应的行政区划信息到二级地级行政区。

在此过程中，删除了部分不在高德地图查询范围内的坐标数据；对直辖市地区统 一使用直辖市的名称。 获取到行政区划信息后，将该信息添加到源数据集的最后作为新的列，得到经过逆地理编 码与数据清洗后的数据。

对处理后的数据，按照行政区划，分别统计某省份（省级行政区）和某城市（地级行政区）内所有经纬度点数据的平均值。

根据相关标准和论文，使用公式计算 AQI 和 IAQI。

$$IAQI_p = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{lo}}(C_p - BP_{Lo}) + IAQI_{Lo}$$

每种指标按照对应时间区间取值，在相同时间区间下的所有污染物的 IAQI 计算结果的最大值作为这个时间区间的 AQI 数值。同时这个最大值对应的污染物就是主要污染物。AQI 从 0 到 500 。AQI 值越高，空气污染程度越高，健康问题也越大。 例如，AQI 值 50 或以下代表空气质量良好，而 AQI 值超过 300 代表空气质量有害。AQI 分为六类。 每个类别对应不同程度的健康问题。每个类别也有特定的颜色。

空气质量指数	空气质量指数级别（状况）及表示颜色	对健康影响情况	建议采取的措施
0 - 50	一级（优）	空气质量令人满意，基本无空气污染	各类人群可正常活动
51 -100	二级（良）	空气质量可接受，但某些污染物可能对极少数异常敏感人群健康有较弱影响	极少数异常敏感人群应减少户外活动
101-150	三级（轻度污染）	易感人群症状有轻度加剧，健康人群出现刺激症状	儿童、老年人及心脏病、呼吸系统疾病患者应减少长时间、高强度的户外锻炼
151-200	四级（中度污染）	进一步加剧易感人群症状，可能对健康人群心脏、呼吸系统有影响	儿童、老年人及心脏病、呼吸系统疾病患者避免长时间、高强度的户外锻炼，一般人群适量减少户外运动
201-300	五级（重度污染）	心脏病和肺病患者症状显著加剧，运动耐受力降低，健康人群普遍出现症状	儿童、老年人及心脏病、肺病患者应停留在室内，停止户外运动，一般人群减少户外运动
300+	六级（严重污染）	健康人群运动耐受力降低，有明显强烈症状，提前出现某些疾病	儿童、老年人和病人应停留在室内，避免体力消耗，一般人群避免户外活动

六 污染物特征值与污染分类：

分析某地区的地表污染类型和起因是环境空气质量研究中至关重要的一环，其中利用污染物数据进行分析是不可或缺的。根据《环境空气质量标准》的规定，我国环境空气质量监测中的六项常规指标在浓度变化和不同污染物之间存在显著的数量级差异。若仅采用比值法对大气污染特征进行分析，则这些指标微小的特征变化将会因为这种差异而被掩盖，导致无法在时间序列或空间序列上区分污染特征的差异。

我们采用了段菁春等人在 2018 年提出的特征雷达图设计中的数据处理算法，对特定地区不同时间的污染物数据进行了处理。处理包括归一化成分谱计算、污染物特征值计算，以及污染物特征标准值和上下限的计算。随后，利用这些处理后的数据与该地区当月的特征标准值和上下限进行比较，以确定某一时间点该地区大气污染物的真实情况，并对当日的大气污染类型进行分类。

对污染物特征值的计算限定于对某一特定城市某日的污染物数值，与该地当月平均特征值的比较。其中，污染物特征值计算主要包括四部分处理。

根据特定地点特定日期每种污染物的特征值与其标准值、上下限的比较，对污染类型进行分类。对 PM2.5，PM10，SO2，NO2，CO 这五种污染物的主要来源进行分析。

根据污染物特征值的不同组合，将大气污染类型分为以下几类：

- 标准型：各污染物特征值均未超出污染物特征值上下限，污染特征未发生显著变化。
- 偏二次型：PM2.5 超出上限，表明污染特征受二次颗粒物生成影响显著。
- 偏沙尘型：PM10 超出上限。
- 偏机动车型：NO2 与 CO 特征值超出上限，表明污染物特征受机动车影响显著。
- 偏燃煤型：SO2 特征值明显超出上限，表明污染特征受燃煤排放影响显著。

- 偏烟花型：PM2.5 和 SO2 特征值明显超出上限，污染特征可能受烟花燃放过程影响。
- 偏钢铁型：SO2、NO2、与 CO 特征值超出上限，表明污染特征受工业排放过程影响。
- 其他型：没有明显的污染物特征。

七 算法预测：

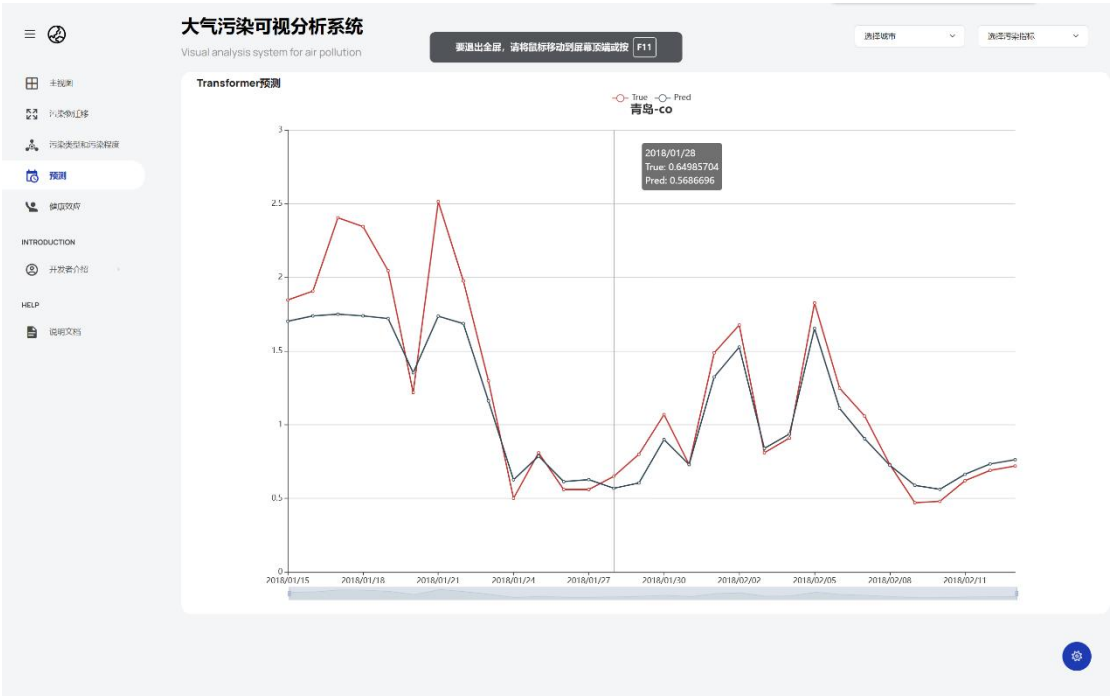
cosSquareFormer

A linear operation with decomposable non-linear cosine-square based re-weighting mechanism instead of a standard non-linear softmax operation - Alleviates quadratic time and space complexity!

$$s(\tilde{Q}_i, \tilde{K}_j) = \tilde{Q}_i \tilde{K}_j^T cos^2\left(\pi \frac{i-j}{2M}\right) = \frac{1}{2}\left[\tilde{Q}_i \tilde{K}_j^T + \tilde{Q}_i \tilde{K}_j^T cos\left(\pi \frac{i-j}{M}\right)\right]$$

This re-weighting mechanism weights the neighbouring tokens more (compared to cosine) with respect to the far-away ones.

为了根据已有信息对未来污染情况进行预测，我们使用 transformer 和模型，基于 2016 和 2017 两年各个城市的数据建模，并对不同城市 2018 一个月的数据进行预测，并绘制折线图如下：



拟合较好。

八 健康效应

1、健康效应研究背景：

随着中国快速的经济增长和工业发展，大气污染问题日益突出。 PM2.5 是中国主要的大气污染物之一，也是长期暴露研究中结果最一致的、健康危害最明显的大气污染物之一。在我国大部分地区，PM2.5 的浓度一直较高，2014~2016 年分别有 95%、87%和 81%的人群暴露于超过国家二级标准(GB 3095-2012)规定的 35 μg·m-3 的环境中，几乎没有人的暴露环境能够达到世卫组织(WHO)空气质量指导(air quality guidelines, AQG)建议的 10 μg·m-3。大量的流行病学研究表明，长期暴露于高浓度的 PM2.5 环境中，可显著增加人群，尤其是生活在中低收入地区人群的多种疾病发病率和死亡率。2015 年全球疾病负担(global burden of disease, GBD)研究报告显示，PM2.5 已经成为全球第五大致死因素。2015 年全球可归因于 PM2.5 暴露的死亡人数为 420 万人， 其中中国为 110 万人，占比超过四分之一。

2013 年，国务院颁布了《大气污染防治行动计划》来治理大气污染，改善空气质量。治理的重点主要集中在中国东部和中部地区，要求与基准年(2013 年)相比，2017 年京津冀、长三角和珠三角地区的 PM2.5 浓度分别下降 25%、20%和 15%

2、

使用 Burnett 等人的综合暴露响应模型(integrated exposure-response, **IER**)计算 25 岁及以上成年人群的 PM2.5 相关健康负担。IER 模型于 GBD2010 报告中首次提出后，已被广泛应用于 PM2.5 健康效应的相关研究。选取缺血性心脏病(ischemic heart disease, IHD)、慢性阻塞性肺病(chronic obstructive pulmonary disease, COPD)、肺癌(lung cancer, LC)和中风(stroke, STK)这 4 个成人主要致死疾病作为健康终点。

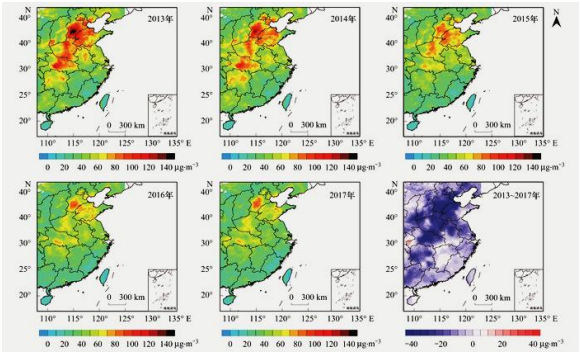
通过公式进行计算相对风险(relative risk, RR) ：

$$RR_{IER}(Z)=\begin{cases}1, & Z\leqslant Z_{cf}\\ 1+\alpha\{1-\exp[-\gamma(Z-Z_{cf})^{\delta}]\}, & Z>Z_{cf}\end{cases}$$

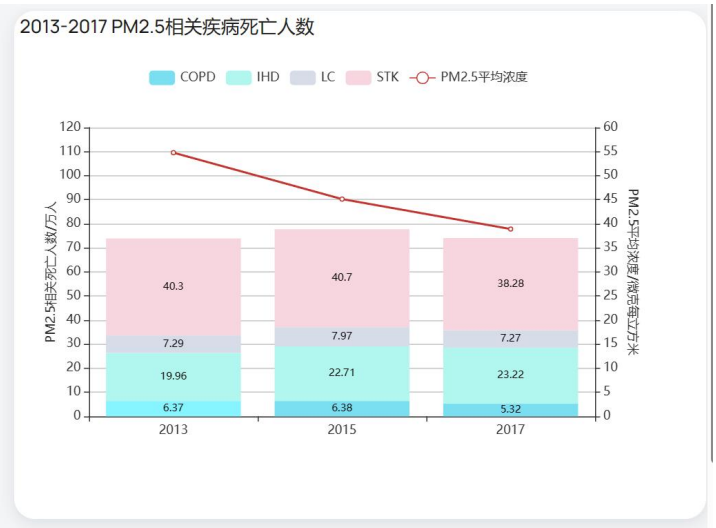
使用得出的相对风险(relative risk),计算归因死亡数ΔMor

$$\Delta Mor = y_0 \times Pop \times \left(\frac{RR-1}{RR}\right)$$
$$= y_0 \times Pop \times AF$$

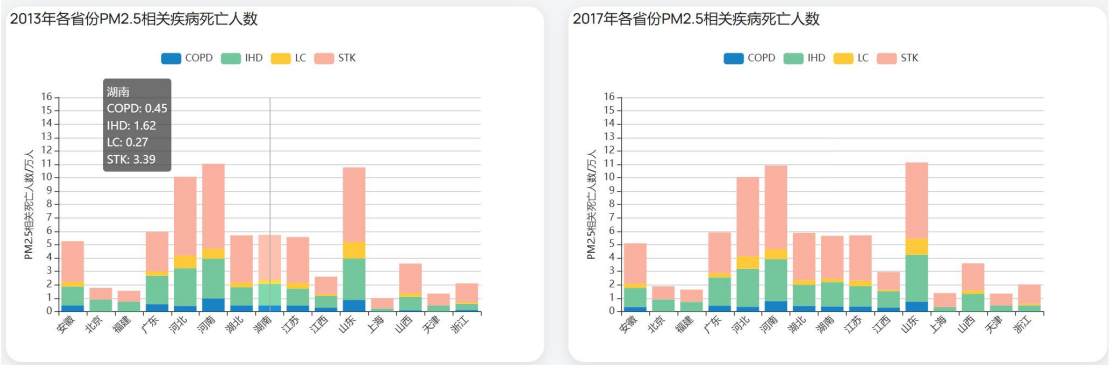
3、健康效应评估结果：



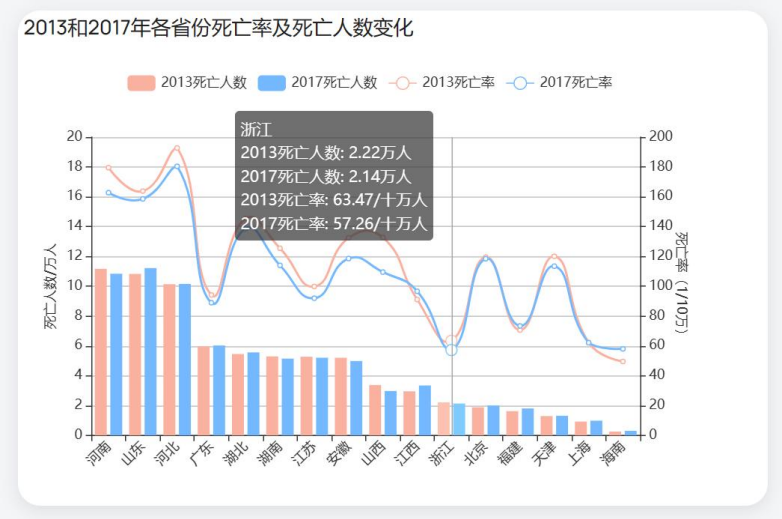
PM2.5 浓度呈现出显著下降的趋势。2013、2014、2015、2016 和 2017 年的 PM2.5 人口加权平均浓度分别为 62.30、58.23、52.60、47.72 和 44.40 μg • m-3。且与 2013 年相比, 2017 年 PM2.5 人口加权浓度下降了 28.73%, 大多数省份的 PM2.5 污染都有不同程度的改善, 尤其是北京、天津、河北、河南和山西, 这些地区的 PM2.5 浓度降幅超过 30%, 海南、福建和广东部分地区的 PM2.5 年均浓度小于等于 35 μg • m-3.在此期间, PM2.5 污染的空间分布相对保持稳定, 重污染主要集中在北京、天津、河北、河南、山东和湖北的部分区域。



2017 年较 2013 年, 可归因于 PM2.5 的总死亡人数变化不大, 但死亡率由 127.68/10 万人降为 119.60/10 万人, 死于 IHD 的人数增加了, 死于 STK、COPD 和 LC 的人数均减少。其中，STK 中风占比仍然最大。



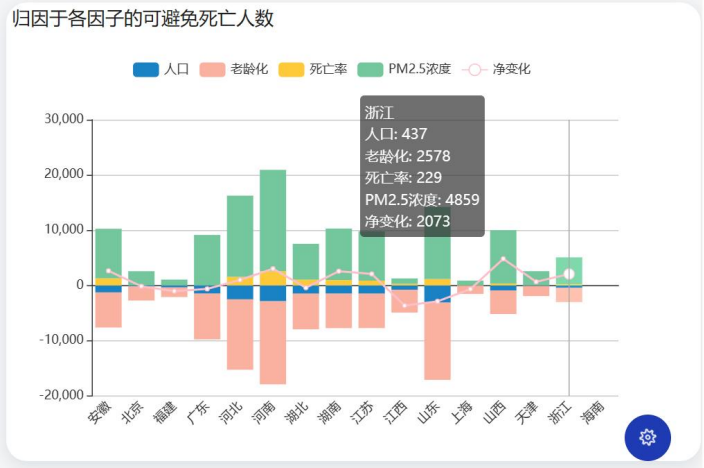
2013 年 PM2.5 相关死亡人数最多的 5 个省份是河南 11.17 万人、山东 10.83 万人、河北 10.14 万人、广东 5.97 万人和湖北 5.46 万人。这 5 个省份的死亡人数分别占到了选取地区总死亡人数的 15.16%、14.70%、13.76%、9.10%和 7.41%, 共计 58.94%。这些省份中, 除广东省外, 其他省份均为 PM2.5 污染严重区域。广东省 2017 年常住人口达到了 1.1 亿, 是中国常住人口最多的省份, 因此健康负担较重。2013 年 PM2.5 相关健康负担最小的 5 个地区是海南 0.26 万人、上海 0.93 万人、天津 1.30 万人、福建 1.63 万人和北京 1.89 万人。 2017 年 PM2.5 相关死亡人数最多的 5 个省份仍然是河南、山东、河北、广东和湖北, 与 2013 年相比 PM2.5 相关死亡人数减少超过 1 千人的省份是山西、河南、安徽和湖南。



如果考虑到人口因素，结果显示，虽然广东的总死亡人数较高，但死亡率和单位面积死亡人数较低，健康负担相对较轻。河北的死亡率最高，是海南的 3 倍。

上海、天津和北京的总死亡人数较少，但单位面积死亡人数很高。这 3 个直辖市是中国经济实力雄厚的大都市，人口密度极高，上海的人口密度甚至可以达到其他省份的 10 倍以上。

同时由于 $PM_{2.5}$ 污染的空间分布与人口分布之间存在很强的正相关关系，因此在人口稠密的特大城市加强 $PM_{2.5}$ 污染控制是非常必要的，更有利于减轻居民健康负担

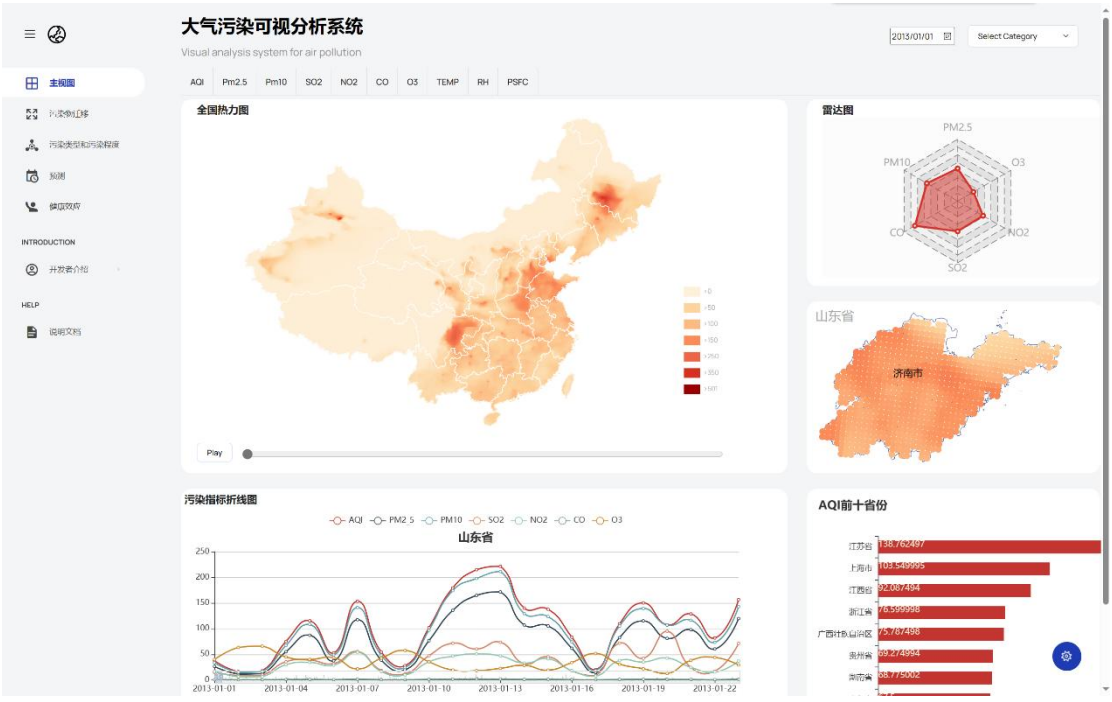


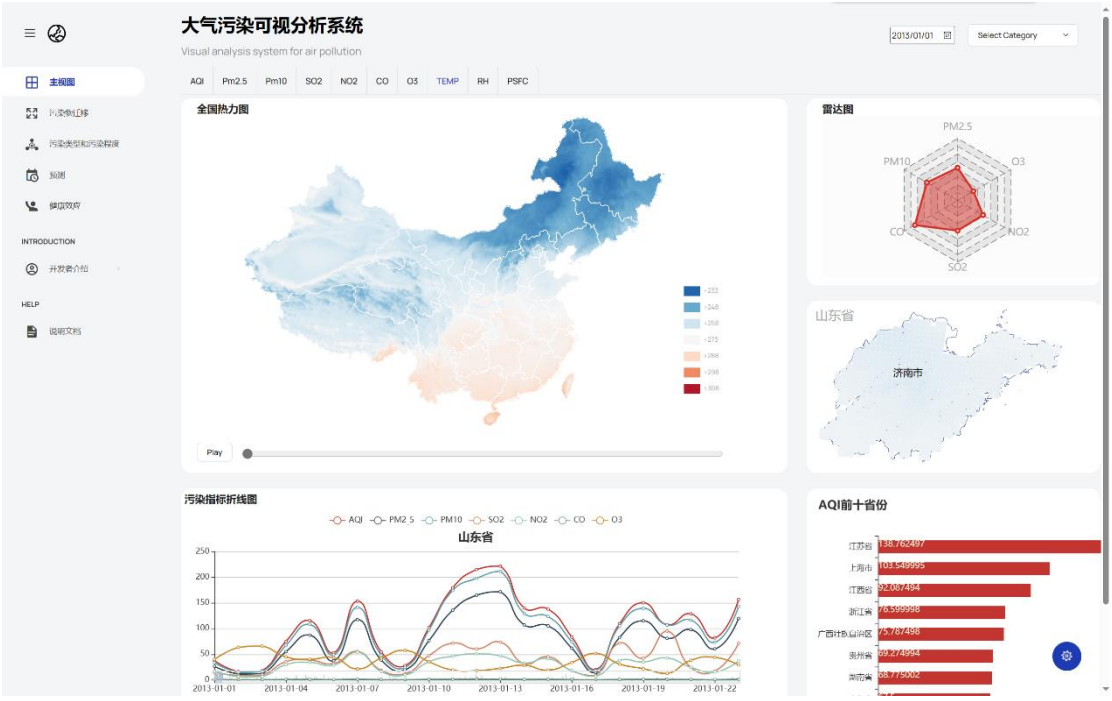
PM2.5 健康效应的估算结果不仅与空气质量变化有关，还与人口总量、老龄化程度及基线死亡率等社会因素有关。死亡率的下降和 PM2.5 浓度的降低减轻了健康负担，然而这些正面作用被人口增长和人口老龄化的加剧所抵消，因此虽然 PM2.5 污染不断改善，但归因死亡人数并没有明显减少，部分地区还发生了增加。尽管人口因素的变化不利于减轻公众健康负担，但控制污染物排放、降低 PM2.5 浓度仍是减轻健康负担最主要最有效的途径。

九 可视化界面设计：

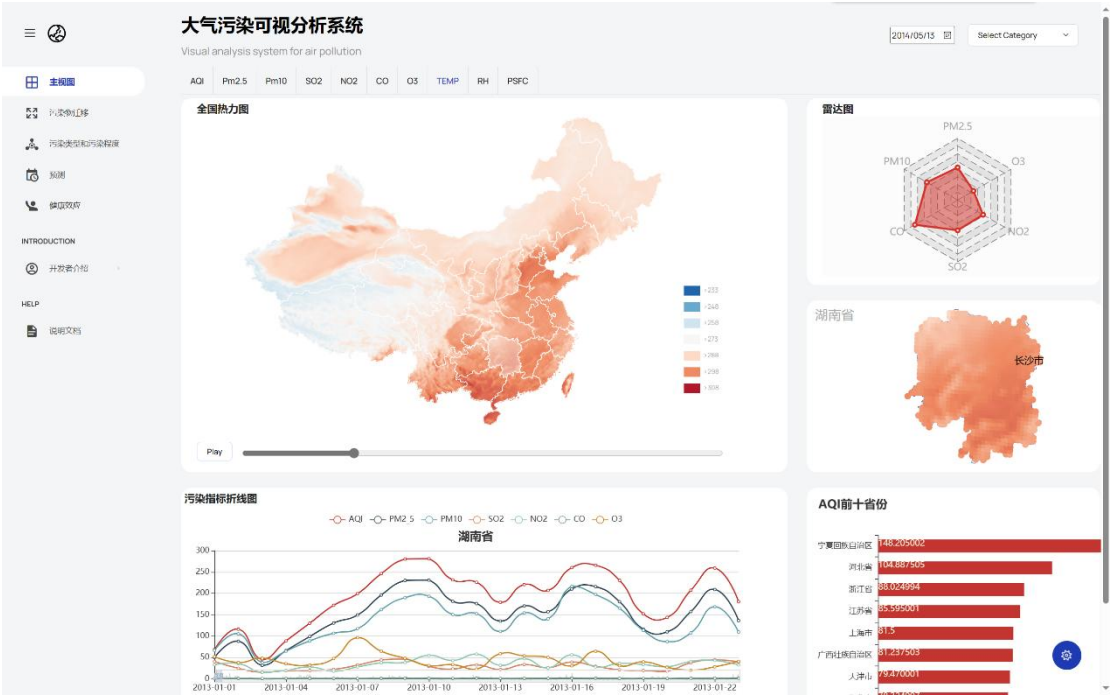
1. 主视图：

可以时空交互，切换指标。



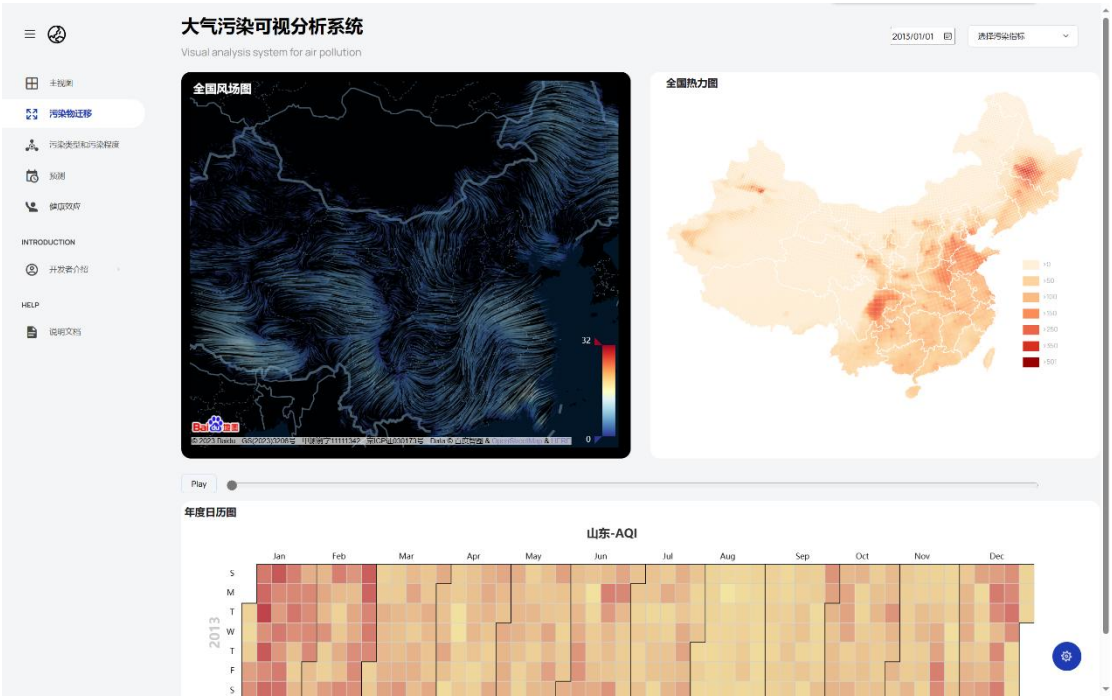


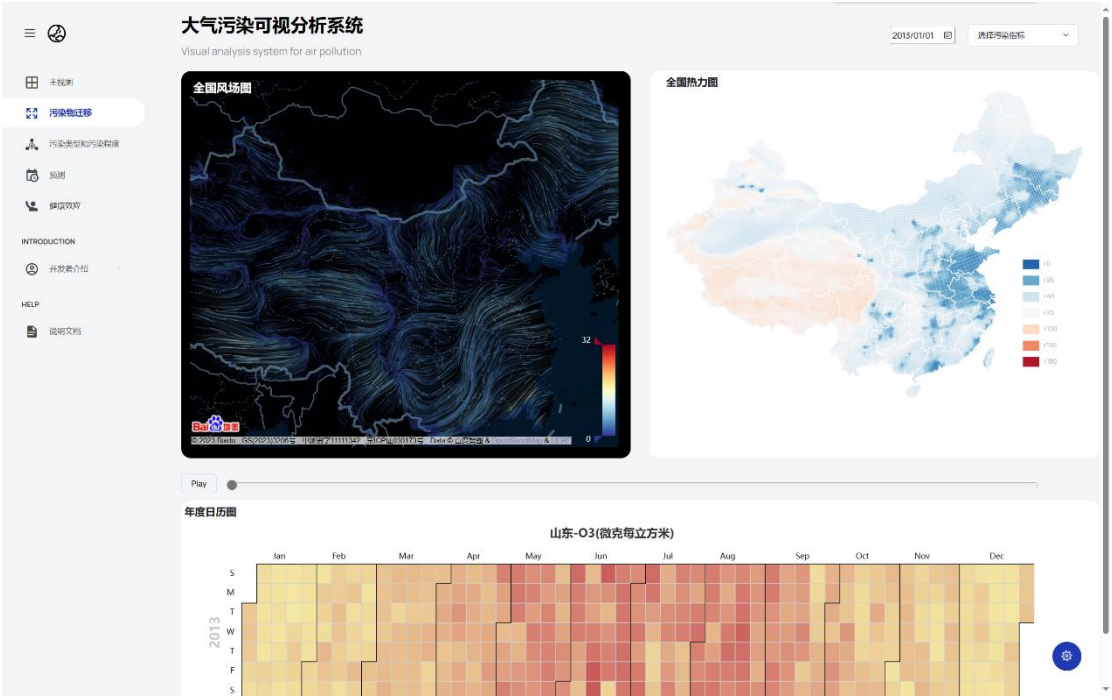

大气污染可视分析系统
Visual analysis system for air pollution
2013/01/01
Select Category
主视图
污染物迁移
污染类型和污染程度
预测
模型效用
INTRODUCTION
开发者介绍
HELP
说明文档
AQI
Pm2.5
Pm10
SO2
NO2
CO
O3
TEMP
RH
PSFC
全国热力图
Play
雷达图
PM2.5
PM10
O3
CO
NO2
SO2
河南省
郑州市
污染指标折线图
AQI
PM2.5
PM10
SO2
NO2
CO
O3
河南省
2013-01-01
AQI: 66.32
PM2_5: 39.98
PM10: 82.639999
SO2: 47.290001
NO2: 29
CO: 1.43
O3: 32.279999
2013-01-01
2013-01-04
2013-01-07
2013-01-10
2013-01-13
2013-01-16
2013-01-19
2013-01-22
AQI前十省份
江苏省
上海市
江西省
浙江省
广西壮族自治区
贵州省
湖南省


大气污染可视分析系统
Visual analysis system for air pollution
2014/05/13
Select Category
主视图
污染物迁移
污染类型和污染程度
预测
模型效用
INTRODUCTION
开发者介绍
HELP
说明文档
AQI
Pm2.5
Pm10
SO2
NO2
CO
O3
TEMP
RH
PSFC
全国热力图
Play
雷达图
PM2.5
PM10
O3
CO
NO2
SO2
湖南省
长沙市
污染指标折线图
AQI
PM2.5
PM10
SO2
NO2
CO
O3
湖南省
2013-01-01
2013-01-04
2013-01-07
2013-01-10
2013-01-13
2013-01-16
2013-01-19
2013-01-22
AQI前十省份
宁夏回族自治区
河北省
浙江省
江苏省
上海市
广西壮族自治区
天津市

2、污染物迁移视图

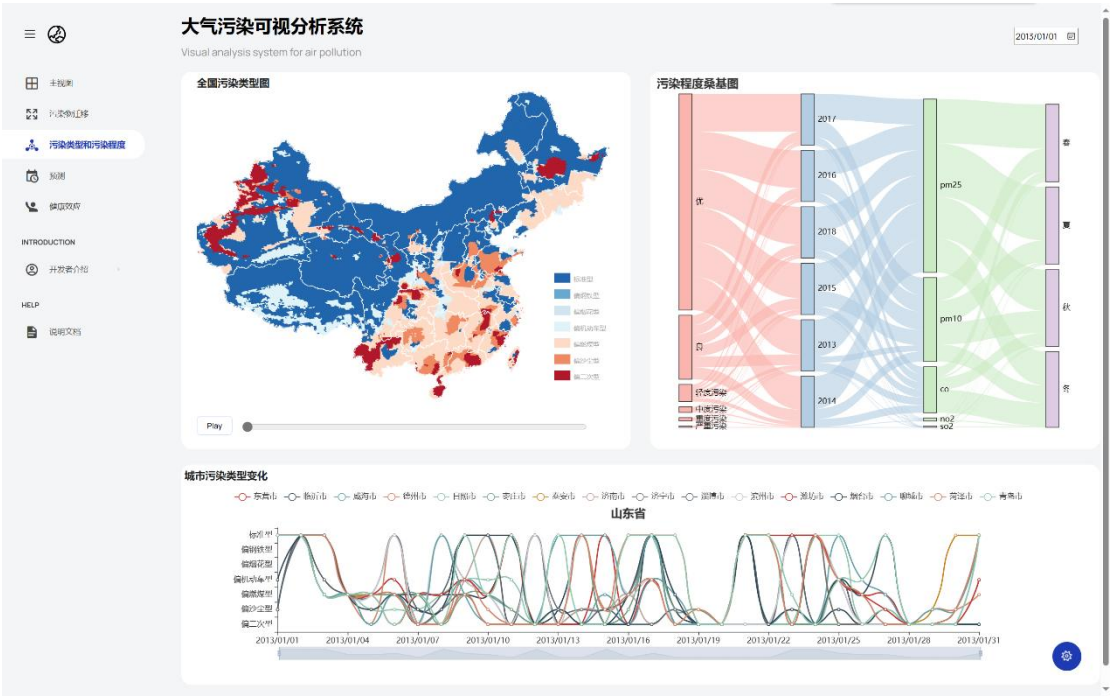
可以时空交互，切换指标。

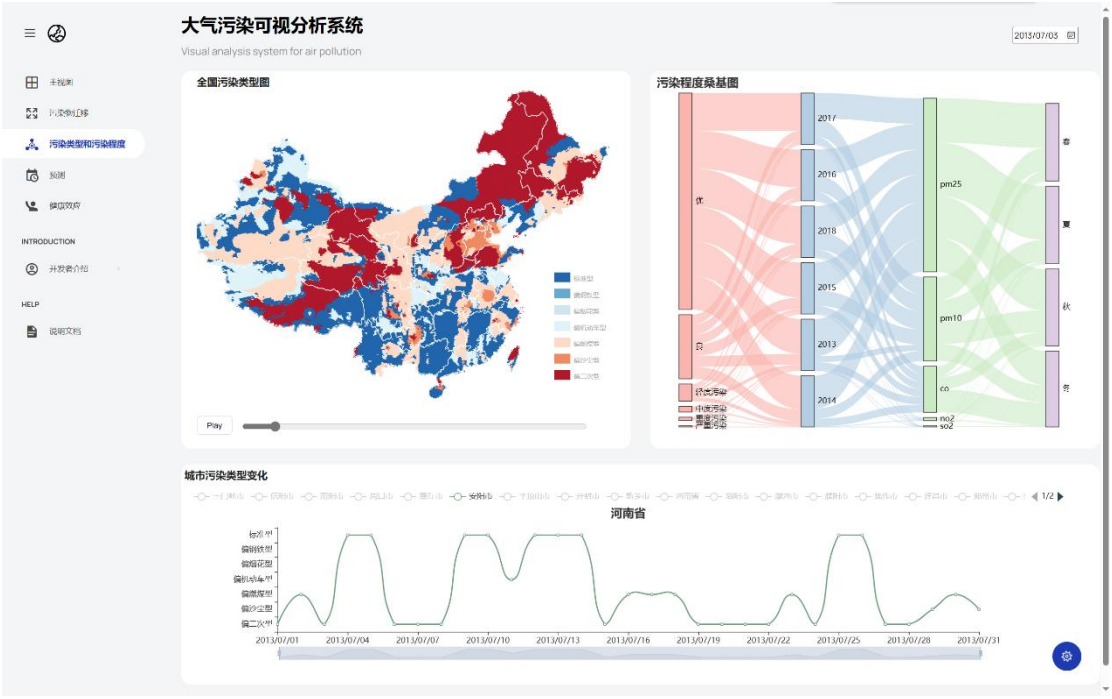




3、污染类型和污染程度视图

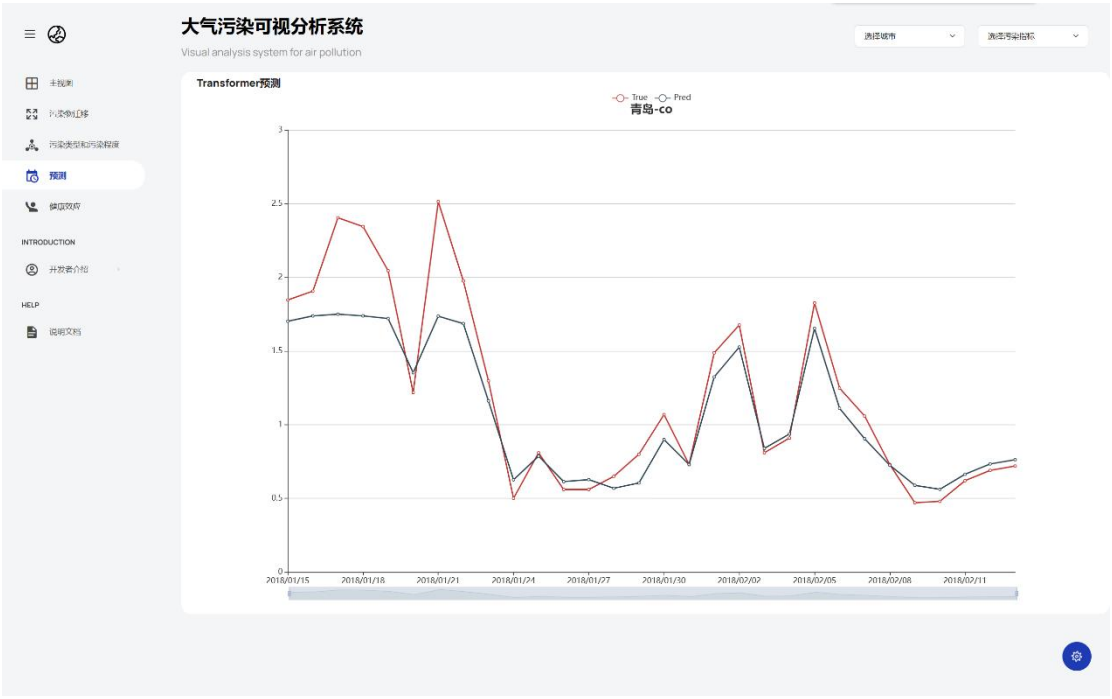
可以时空交互，切换指标。

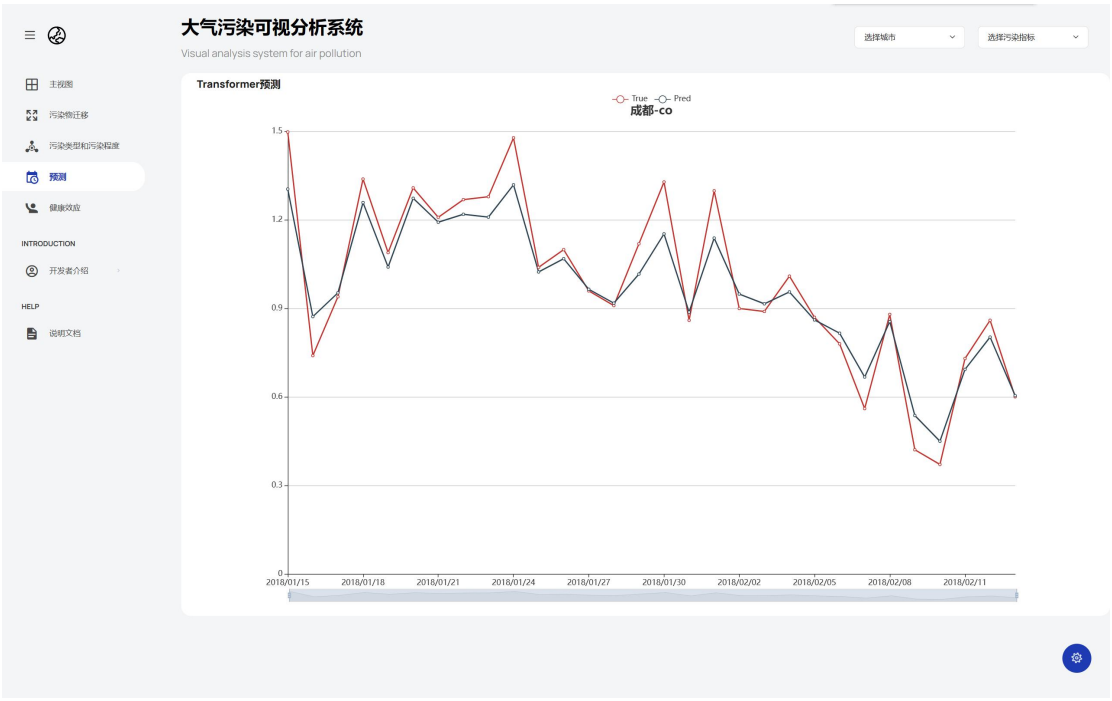




4、预测视图

可以切换指标





5、健康效应视图

十 实验总结：

本可视化系统清晰全面地展示了大气污染时空分布模式与时空演变态势，并拓展至探究污染物与气象以及污染物之间相关性的时空变化以及大气污染物变化趋势的空间分布聚集性，具有较强的可探索性。

本系统在能够完成可视分析任务的基础上，对视觉编码、视图创新与交互上做了相对细致的设计。但对于高维时序大气污染数据，还有众多可以切入的角度，以及与相关领域知识和数据结合，做更深入的分析。

圆满完成实验目标。