

2021 年第八届中国可视化与可视分析大会
数据可视分析挑战赛

(ChinaVis Data Challenge 2021)

作品说明文档

参赛队名称：西交利物浦大学-张智超

作品名称：大气污染时空分布及预测可视化

作品主题关键词：大气污染分析、大气污染预测

团队成员：张智超，西交利物浦大学，

zhichao.zhang20@student.xjtlu.edu.cn，队长

刘钰，西交利物浦大学，yu.liu@xjtlu.edu.cn

徐宁宁，西交利物浦大学，ningning.xu2002@student.xjtlu.edu.cn

石天磊，西交利物浦大学，tianlei.shi18@student.xjtlu.edu.cn

韦兴波，xingbo.wei18@student.xjtlu.edu.cn

俞凌云，西交利物浦大学，yingyun.yu@xjtlu.edu.cn，指导老师

Hai-Ning Liang, 西交利物浦大学，HaiNing.Liang@xjtlu.edu.cn, 指导老师

团队成员是否与报名表一致（是或否）：是

是否学生队（是或否）：是

使用的分析工具或开发工具（如果使用了自己研发的软件或工具请具体说明）：D3，
python, html，AutoGluon，ProPhphet

共计耗费时间（人天）：45 人天

本次比赛结束后，我们是否可以在网络上公布该文档与相关视频（是或否）：是

一、 作品简介：请围绕作品主题、要解决的问题\场景、目标用户\读者、应用价值等方面简要介绍作品

1.1 作品主题以及要解决的问题

大气污染是目前当今人类面临的重要公共卫生问题之一。有研究表明，大气污染的加重会带来心理或生理多个方面的疾病，比如呼吸和心血管疾病（Sun et al.2019）或对皮肤或头发造成损害（Rajput 2015; Chua et al. 2019）。同时，糟糕的空气还会影响生态环境和农业污染。比如因为臭氧有强氧化性，当浓度较高时，会直接危害植被的生长从而会影响生态环境质量，进而降低农业产量和质量（Feng et al.; Fuhrer et al.）。

解析空气污染的时空分布特征以及气象条件对城市空气质量的影响，对空气污染政策的制定以及有效措施的设计有着及其重要的作用。同时精准的空气质量预测对政府环境保护相关政策的制定有着重要意义。所以，我们基于气象数据和空气空气污染治理数据设计并实现了一个探索式分析系统（如 Fig 1 所示），其目的是为了帮助用户

- 1. 明确污染全年走势
- 2. 明确全国主要城市的污染情况
- 3. 了解全国范围内气象数据与各类污染物的关系
- 4. 了解局部地区气象因素改变造成的可能的污染物影响

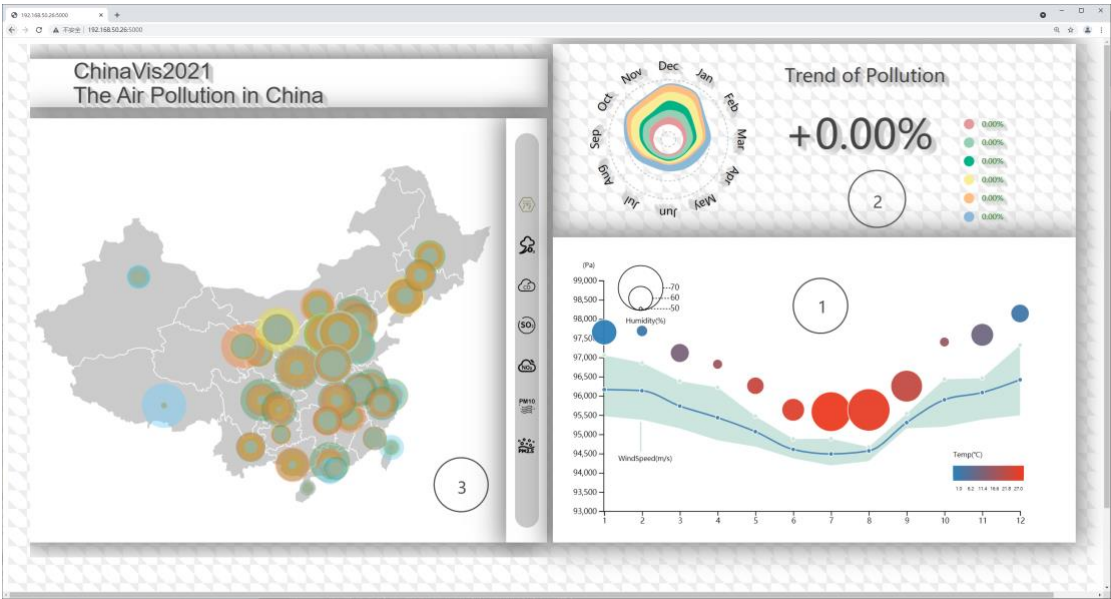


Fig.1. 系统界面。模块 1：全年温度，湿度，风速和压强的走势（预测值）；模块 2：全年六种污染物的走势（预测值）；模块 3：全国六种污染物的严重程度

本工作最大的亮点在于：随着数据量和复杂度的增长，用户很难通过抽象数据获取数据隐藏的信息，也很难快速地判断预测模型的有效性。我们的探索式交互系统通过数据驱动来做预测，通过可视化帮助人们快速判断预测模型的有效性：通过雷达图（模块 2），我们以简洁、有效的方式帮助用户快速地了解该气候状况下污染物大致情况。若用户需要进一步具体分析，可以将预测模型放置到某一城市（或某一区域），在地图上（模块 3）观察具体地点的污染情况。

1.2 目标用户和应用价值

本作品**目标用户**广泛，主要可以分为：大众、决策者以及各领域的学者。通过与各视图的交互与分析，探索全国污染走势，了解气象因素和污染物之间的关系。**面向大众**，可帮助其快速了解污染全年走势和空间分布；**面向决策者**，可帮助其了解措施可能导致的污染走势，并采取合理的调整；**面向各领域学者**，可帮助其对与科研所需的污染和气象数据进行可视化分析，对污染治理进一步探索。

二、 数据介绍：请围绕数据来源、数据格式、数据严谨性、数据清洗等方面简要介绍

2.1 数据集

本项目采用的是全国 34 个主要城市地区的大气污染和气象数据。数据来源于大赛官方提供的全中国每日的分析数据（2013 年-2018 年）。其中包含来自 42249 个不同地点的数据，本项目根据百度地图 API 提供的全国 34 个省会及行政区的经纬度信息，根据欧式距离从 42249 个地理数据源中筛选出距离最接近的作为各个主要城市的代表。数据以 csv 文件格式存储，数据格式均为浮点数且不包括缺失值。其中包括：

（1）大气污染数据

污染数据包括 PM2.5, PM10, O3, CO, SO2 和 NO 六种常见污染物。

（2）气象数据

气象数据包括了气温，风速，气压和相对湿度。

2.2 数据预处理

读取全部文件，筛选出 34 个主要城市代表数据，通过经纬方向风速特征计算出总风速值，同时根据文件名提取对应数据时间特征。根据年份和月份与城市名特征，借助透视数据表，获得各主要城市各类数据的月的均值。如 Fig.2 所示。

			CO	NO2	O3	PM10	PM2.5	PSFC	RH	SO2	TEMP	WindSpeed	lat	lon
year	month	name												
2013	1	上海	1.449677	66.449355	30.580000	101.112581	90.141290	102667.284516	75.934194	32.909355	277.891290	5.256495	31.20	121.53
		乌鲁木齐	1.495806	35.883548	48.030323	79.128065	57.960323	89615.903871	39.700000	16.470968	268.683871	9.461710	43.71	87.67
		兰州	1.848387	42.590323	25.941613	107.688710	75.612903	79745.470323	40.420645	37.440323	268.185161	2.441897	35.97	103.74
		北京	3.981935	90.510323	11.339677	235.911290	183.109032	102456.987419	55.208710	115.113548	266.699032	2.694278	39.97	116.38
		南京	1.959677	72.443548	19.823871	192.855484	128.501290	102544.700323	61.958387	70.261613	276.170645	3.483998	32.09	118.77
...
2018	12	重庆	0.984839	44.295806	9.166452	81.740645	58.921613	99178.401290	61.665161	33.239677	282.026452	2.036414	29.57	106.48
		银川	1.370645	43.009032	30.750645	137.705806	49.203548	89903.558387	37.227419	41.630645	267.567742	3.468528	38.41	106.30
		长春	0.867097	37.631613	26.793548	64.845484	39.370968	100005.130000	45.749355	24.990968	262.734516	5.195795	43.86	125.37
		长沙	1.062903	41.142903	17.377097	112.090323	88.697419	102197.059677	74.754194	19.595484	280.453226	4.844243	28.23	112.94
		香港	0.672581	37.224194	33.190000	47.898065	30.559032	101303.623226	72.902903	11.332581	289.591290	8.349340	22.46	114.08

2448 rows x 12 columns

Fig.2. 各个特征不同时间地点的月均值

由于不同污染物的物理化学形态不尽相同，所以对他们采用的评价指标具有不同的纲量和纲量单位。为了方便我们在同一堆叠面积图的极坐标系下综合比较变化趋势，我们运用的是最常见的 min-max 标准化对数据进行对数据进行线性变化，使结果映射到 0 和 1 之间。

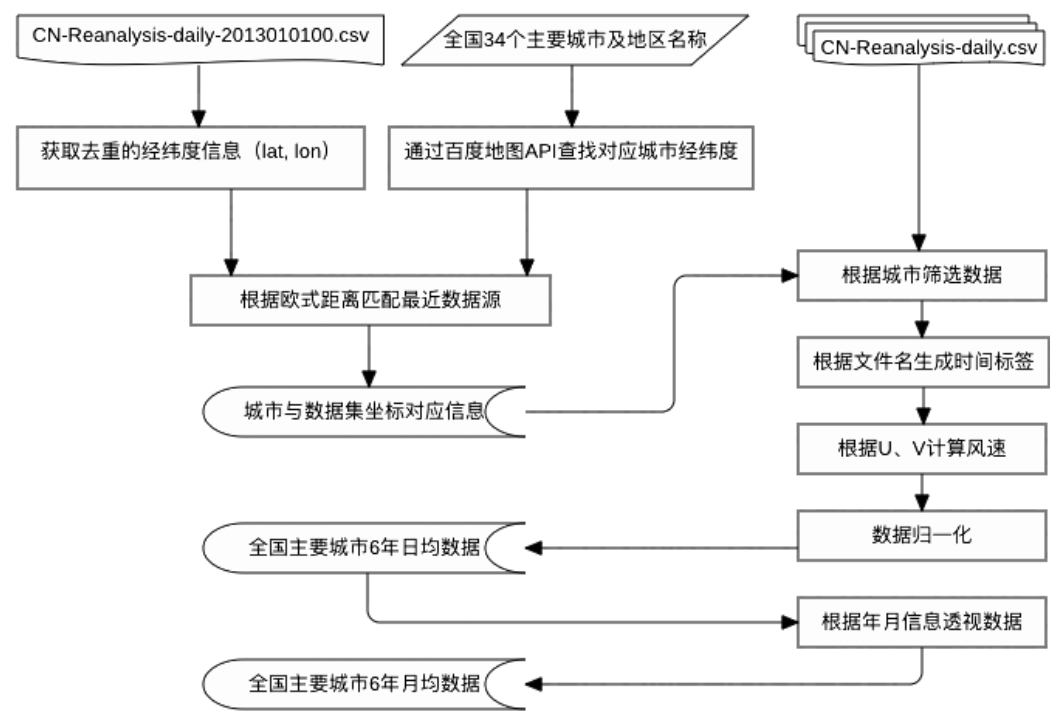


Fig.3. 数据预处理流程

三、 分析任务与可视分析总体流程

本作品以全国污染在中央和地方政府各项措施的努力下大气质量逐渐得到好转为背景，结合 2013-2018 这六年来的大气和污染数据与数据分析处理模型，通过可视交互界面，设计并开发实现了天气各项数据和污染物数据分析可视平台，开展相关可视分析工作，挖掘气象和污染物之间的相关关系，分析局部气象因素改变对一段时间一定区域内污染物的影响。本系统实现了以下分析任务，如 Fig.1 所示：

- 任务一：通过模块 1，了解全年温度，湿度，风速和压强的走势。
- 任务二：通过模块 2，了解全年六种污染物的走势。
- 任务三：通过模块 3，了解全国六种污染物的严重程度。
- 任务四：通过模块 1 和 2 的结合，了解气象因素改变对污染物的影响。
- 任务五：通过模块 1，2，3 的结合，分析单一区域某个时间气象变化带来的各类污染的影响

四、 数据处理与算法模型

首先，用户根据各类气象特征（气温，风速，气压和相对湿度）以往数据，预测其各自未来的发展趋势（数据处理与气象预测模型，Fig 1，模块 1）。基于预测的气象数

据以及某一污染物的实际数据，预测该污染物未来可能的发展趋势（污染物预测模型，Fig 1，模块 2）。

当用户调整了某个月的气象数据，系统将根据调整后的新气象数据作为训练集重新拟合，产生当年剩余月份的气象数据，最后得到全年新的气象数据。利用新的气象数据以及污染物之前的数据，预测用户调整气象后的污染情况（Fig. 6）。

(1) 数据处理

基于预处理的数据，进一步基于时间特征，通过透视表获取全国主要城市各类指标的每月的平均值，可用作对全国主要城市整体情况分析的数据源。

基于预处理的数据，按城市名称筛选就能获得该城市各类指标的每月的平均值，作为对某一特定城市分析的数据源。

数据初始划分方式以 2013 年至 2017 年低的所有数据作为训练集合，2018 年全年数据为测试集。

(2) 气象预测模型

仅基于时间特征通过透视表获取全国主要城市各类指标的每月的平均值，使用 Facebook 开源的时序预测软件 Prophet (TAYLOR et al. 2018) 为每一种气象数据进行单独建模。Prophet 是一种能够进行时间序列数据预测的工具，它最适用于具有强烈季节性影响和多个季节历史数据的时间序列。Prophet 对缺失数据和趋势变化具有稳健性，并且通常可以很好地处理异常值。Prophet 的模型输入须要有时间和需要预测的特征，它不能进行多元时序分析，但它在气象数据的预测方面表现良好，因此选用它搭建模型对气象数据建模。

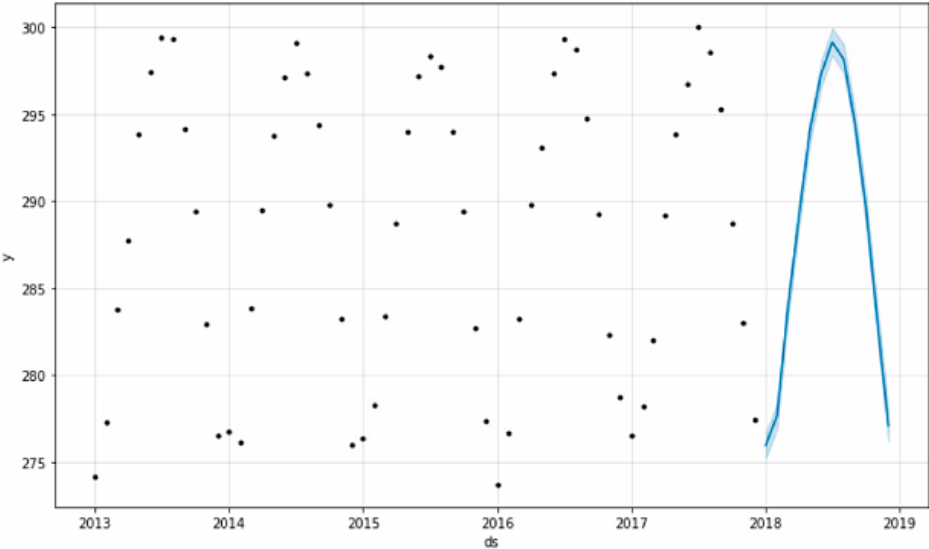


Fig.4. 全国主要城市总体月均气温分布以及预测

(3) 污染物预测模型

使用单一污染物以及全部种类的气象特征作为数据源，使用亚马逊开源机器学习框架 AutoGluon (Erickson et al. 2020)，进行多元回归预测。分析气象与污染物之间的关系。AutoGluon 实际上是一种 automl 框架，它能尽量在不需要人工的干预下，自动对数据抽取特征，选取合适的模型并进行训练。现有大部分 automl 都是基于超参数搜索来实现，模型往往需要数十或数百次的训练才能搜索到一个能够媲美人工调参的效果的超参

数。而 AutoGluon 的思路则是融合多个不需要超参数搜索的模型，以便在相同时间尝试更多种不一样的模型。它的实现方式包括，如 **stacking**, **k 折交叉的 bagging** 和多层 **stacking**。它将对训练集尝试多种不同的模型（如 **KNN**，树模型，核方法，神经网络等），最后用线性模型加权聚合。以便在一定资源的前提下，尽可能尝试多种不同的模型，并将其融合，得到效果很好的训练结果。

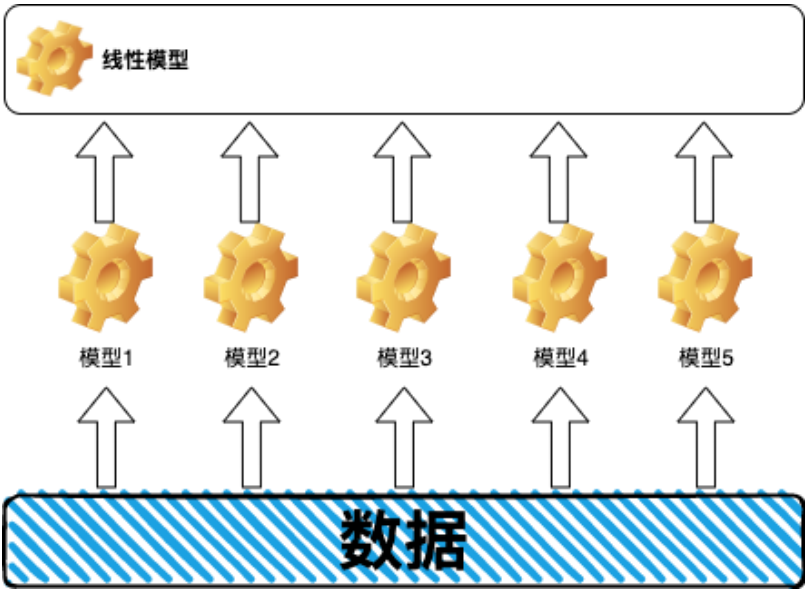


Fig.5. Stacking 模型示意图

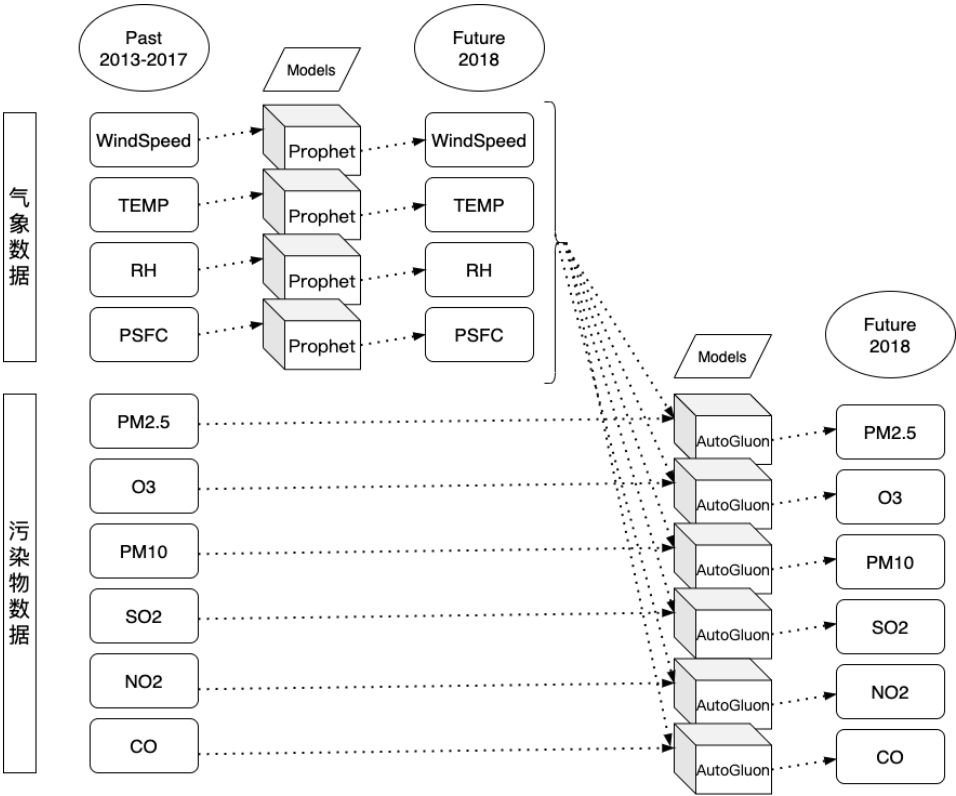


Fig. 6. 建模流程

五、 可视化与交互设计

本作品通过三个模块对大气污染天气状况相关数据进行了可视化，实现了视图内、视图间的交互操作，便于用户进行可视化分析，了解污染的时空分布以及相关性。本章将对该作品的三个主要可视化模块和一个文字模块进行详细的介绍。

1. 2018 全年气象数据的可视化

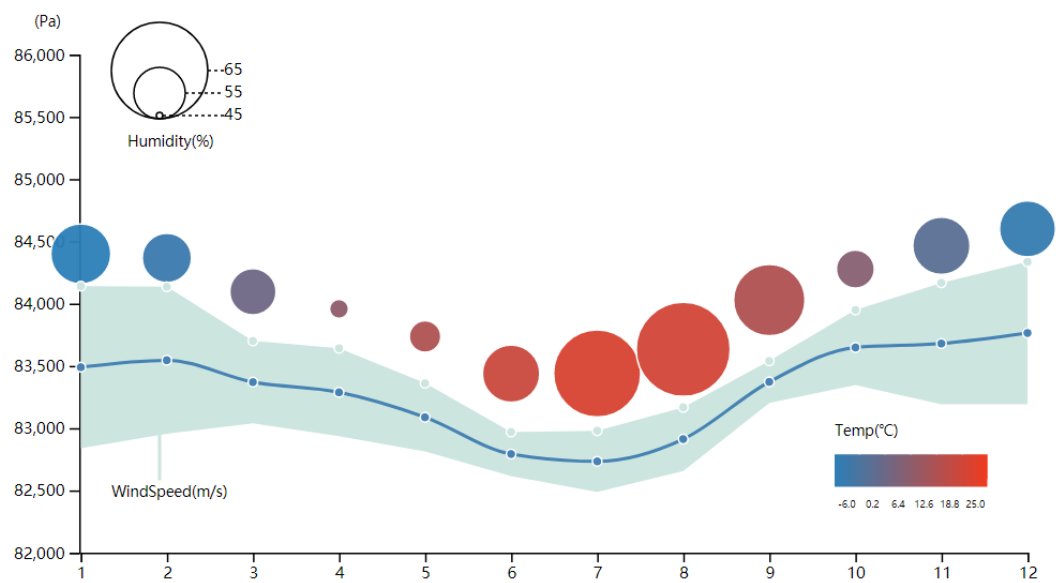


Fig.7. 可交互的气象数据可视化

通过 2013-2017 年的训练数据，我们预测了 2018 年的气象数据，如图 Fig. 7 所示。





气象数据	可视化元素	可视化元素说明	用户交互
风速	浅蓝色宽线 	线条 width 编码风速	选中小圆点进行上下拖拽
压强	深蓝色折线 	折线 Y 值编码压强	选中线上边界的小圆点进行上下的拖拽
湿度	气泡大小 	气泡大小编码湿度	通过鼠标滚轮进行大小的变化
温度	气泡颜色 	气泡颜色从蓝到红编码温度的从低到高	左键点击鼠标上下移动调整温度

Fig.8. 四种气象数据编码方式

可视化：我们通过可视化（折线图和气泡图的结合，Fig. 8），展示了四种气象因素全年的走势，需要注意的是，可视化显示的是气候的预测结果（当前以 2018 年为例）。可视化的结果可以让用户大致了解气候的发展趋势。在可视编码元素中，位置、大小和

颜色是最有效的可视化编码元素，因此我们有效地利用了这些元素，将气候和污染物的预测模型快速地展现。

交互：我们也支持用户根据预测的气候和污染情况，做一些改善措施（例如人工降雨等）。当用户改变了气候预测值，通过 linked view 污染物的预测值（在下面污染物的可视化中介绍）也会相应发生改变。用户可以调节湿度大小，温度高低，以及风速大小和压强高低来探索这些因素对六种污染物的影响。为了确保用户在合理区间内对数据进行调整，系统通过分析 2013-2018 的数据，设定了可以调整的上下限（分别是这六年来当月均值的最大和最小值）。

2. 2018 全年 6 种污染物的可视化

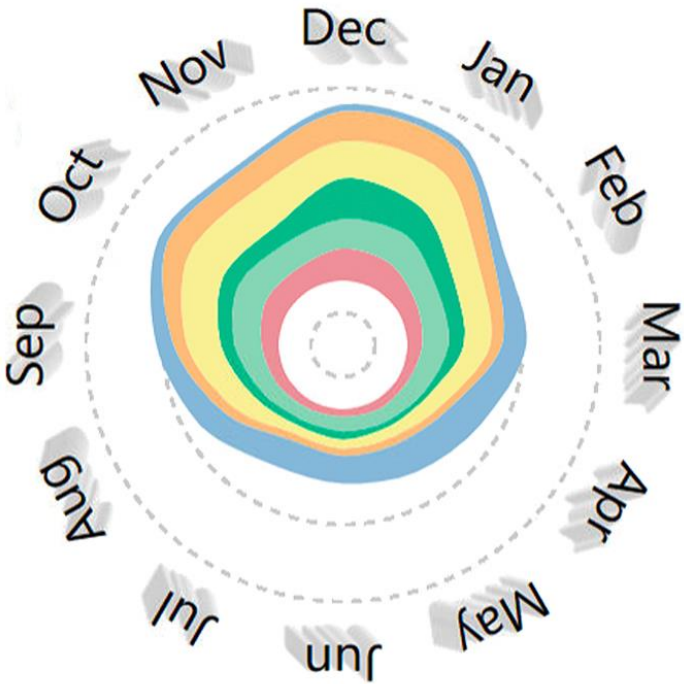


Fig.9. 可交互的污染物数据可视化

通过堆叠的雷达图我们展示了六种污染全年的走势。需要说明的是，为了支持用户通过简洁、有效、直观的可视化，快速地观察、判断调整后的气象数值对污染物的影响情况，因此，我们并不期望得到各污染物的具体数值，而是突出视觉效果。若用户通过比较，确定调整后的气象数值是一个好的模型，他们可以将之应用于某一地区，进一步观察全国主要城市在应用了调整后的气象数值后的污染情况。

为方便对某一种污染物走势的探索，我们支持用户通过鼠标悬浮高亮单独一种污染物全年的走势，如 Fig.5，10。

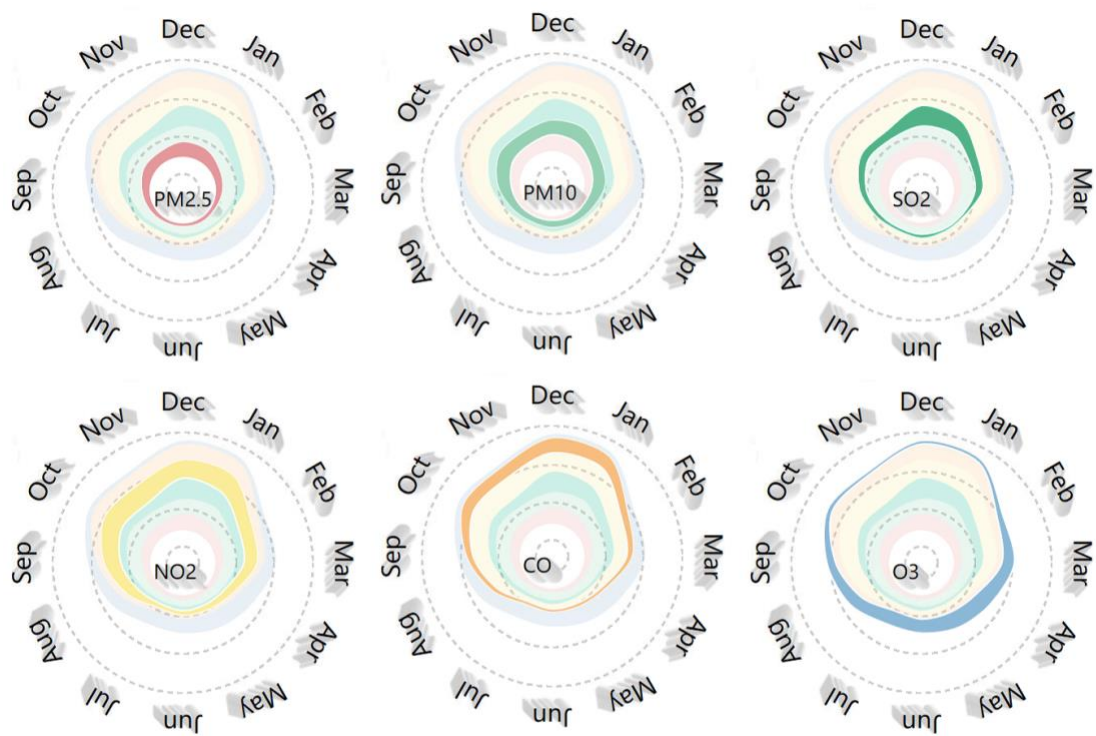


Fig.10. 六种污染物的全年浓度走势

3. 2018 全年全国主要城市的污染的可视化

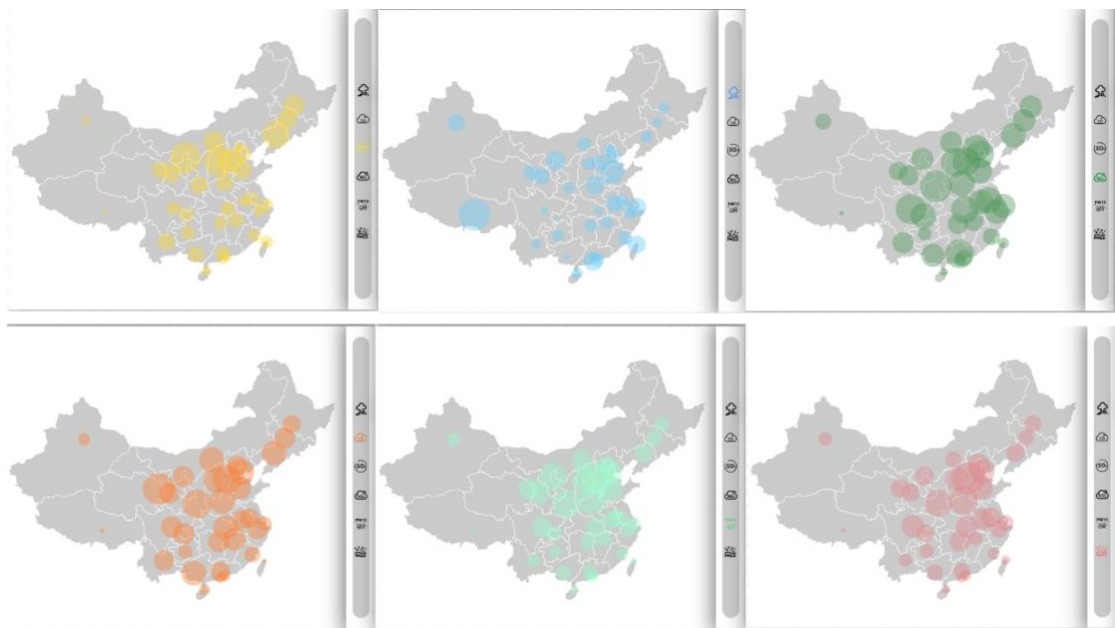


Fig.11. 可交互的污染物浓度可视化地图

通过地图展示了我国主要城市的污染情况。通过选择不同污染物，可探索不同污染物在我国主要城市的分布，如 Fig. 11。

4. 污染趋势

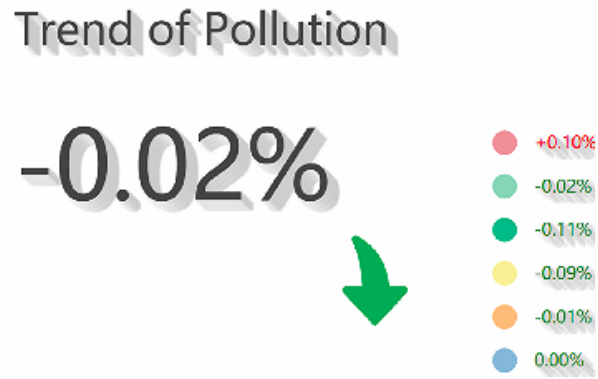


Fig.12. 污染变化数值

通过计算各污染物的变化数值，便于在污染物变化及其微小时洞察各类污染物的变化，如 Fig. 12。

六、 实验\案例\场景分析

案例 1：探索我国污染物的时空分布。

通过堆叠的雷达图，我们可以得到六种污染的全年的走势。污染物呈现季节性变化规律：二氧化氮，二氧化硫和一氧化碳均呈现秋冬严重而春末和夏天缓和。PM2.5 和 PM10 则呈现春冬严重而夏秋轻微。臭氧与其他污染物全年走势差异较大，呈现春夏严重而秋冬缓和。

- 不同污染物之间的相关性：整体来看，中国 PM2.5、PM10 呈现极显著正相关关系。
- 污染地域性特征：太原，银川沈阳是全国省会城市中 SO2 污染最为严重的城市。京津冀地区的 PM2.5/PM10 最高。拉萨，济南的在 34 个重要城市中全年臭氧污染最严重，如 Fig. 13。

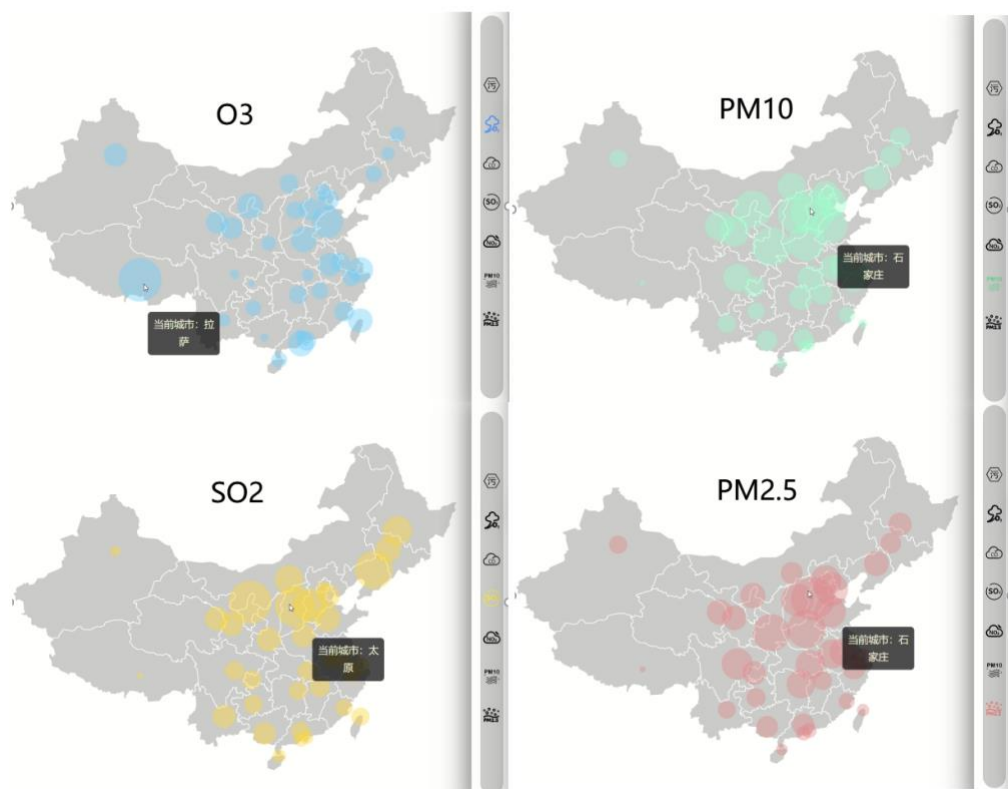


Fig.13. 全国 34 个重要城市中臭氧，PM2.5，SO2，PM10 全年平均浓度最高的城市

案例 2: 探索气象因素在全国范围内对各种污染物的影响。

通过模块 1 和 2 的结合，能很快的明确气象因素对污染物之间制约的关系。通过在合理范围内调整风速，压强，温度和湿度，可以迅速看到模块 2 相关月份的污染物的变化。从而甄别改变这个气象因素是不是与污染程度有紧密的联系从而选取最相关的因素来运用到局部地区。本文根据气象学方法以春季（3—5 月）、夏季（6—8 月）、秋季（9—11 月）、冬季（12—2 月）为标准进行季节划分。

通过调节三种常见的气象数据我们得到：

- 通过调节温度，我们发现臭氧浓度与城市温度呈正相关关系。无论是在夏季（6 月）还是在冬季（1 月），当温度下降时，臭氧都呈现下降趋势。其他污染物在温度下降时，污染水平都呈上升趋势，如 Fig.14。

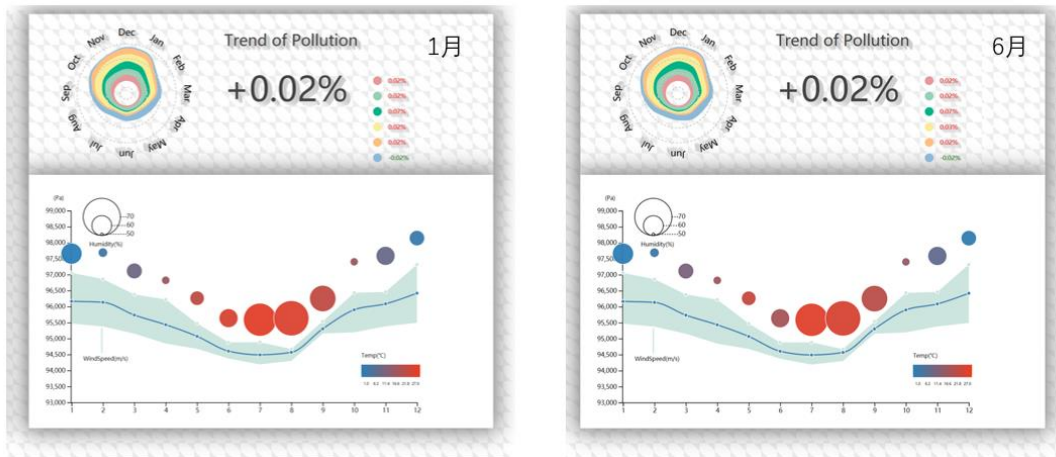


Fig.14. 全国平均气温下降后不同月份的污染物变化

- 在调节风速时，我们发现：夏季（6月）风速上升时，除了二氧化硫其他都成下降趋势，二氧化硫呈上升趋势。冬季（1月）风速上升时，除了二氧化硫基本维持不变外，其他污染物都呈下降趋势。所以我们可以预测风速与整体大气污染呈负相关，如图Fig.15。

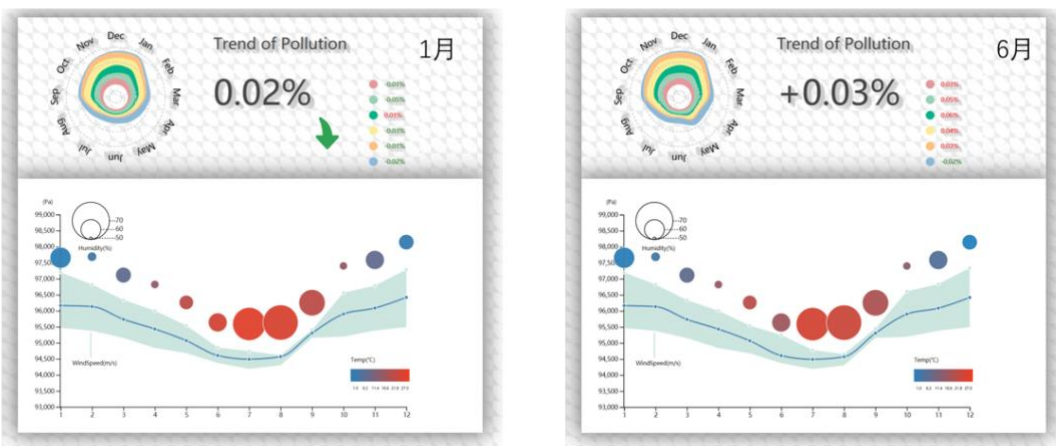


Fig.15. 全国平均风速上升后不同月份的污染物变化

- 在调节湿度时，我们发现在一些季节湿度的影响是有限的。在冬季（1月）增加湿度时，我们发现所有污染物呈下降趋势。但在夏季（6月）调节湿度时，所有污染物没有明显变化，如图Fig.16。

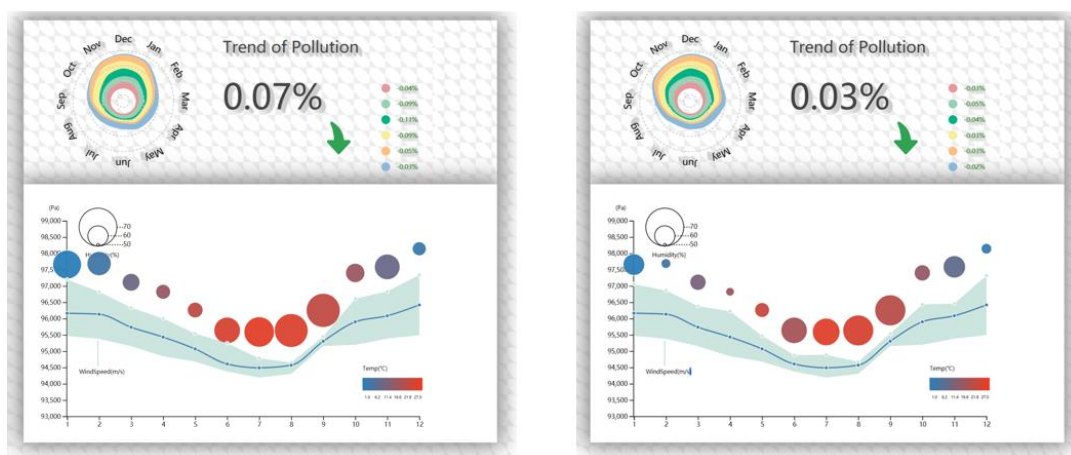


Fig.16. 全国平均湿度上升后不同月份的污染物变化

案例 3：探索在局部区域的气象数据和污染物浓度之间的关系。

通过案例 1 将关键气象因素甄别出来后，我们可以在模块 3 的地图中选中局部地区，来观察局部气象因素改变后污染物的走势以及地理上的影响。从案例二的探索中，我们认为风速和温度将是十分重要的影响污染的因素。而湿度可能在个别季节影响有限。所以我们着重观察风速和温度的影响，其次再探究湿度是否在局部地区也影响有限。

选取北京、南京和广州作为我国京津冀、长三角和珠三角地区代表性城市，通过调节各城市的空气质量状况对污染物浓度的变化进行对比分析，分析结果如下：

- 风速对三个城市不同季节的影响：在冬季（1 月），当风速增加时，北京和南京除二氧化氮外其他所有污染物都呈下降趋势（Fig.17）；而广州无明显变化。在夏季，当风速增加时，北京六种污染物都无明显变化（Fig.18）；南京的 PM2.5 和 PM10 以及二氧化硫的浓度都呈上升趋势；在广州，二氧化硫浓度有所下降，其他污染物均无明显变化。

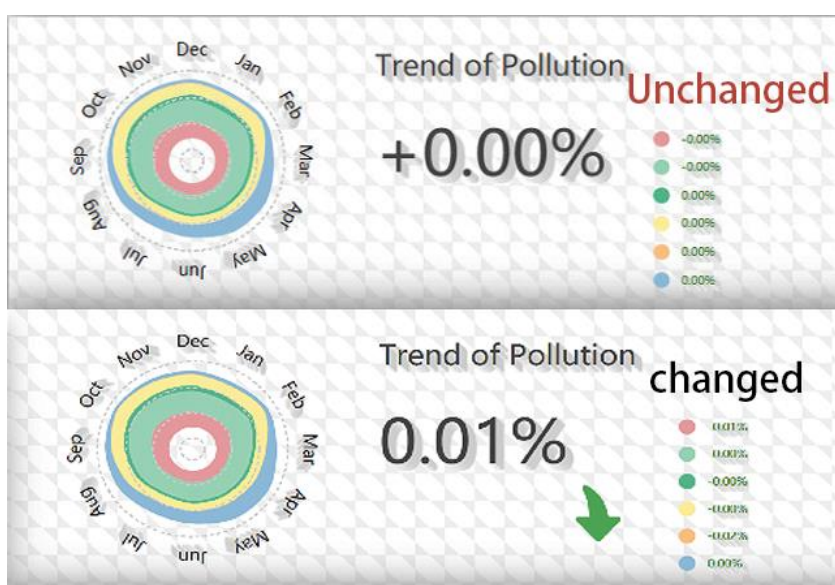


Fig.17. 一月北京风速增大前后的污染情况

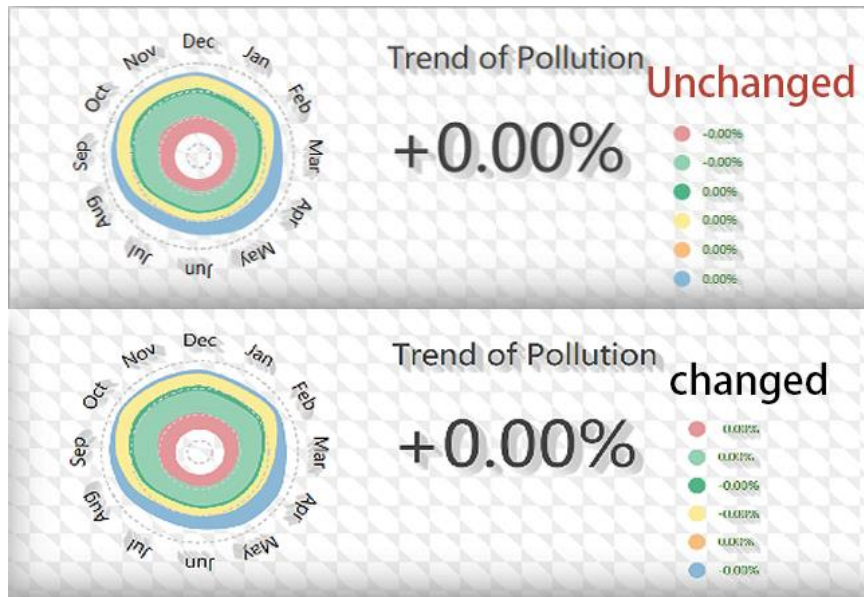


Fig.18. 六月北京风速增大前后的污染情况

- 温度对不同城市不同季节的影响：在冬季（1月），当北京温度下降时，一氧化碳会呈现上升趋势，其他均无明显变化；在南京，温度的下降会导致 PM10，二氧化硫，一氧化碳和二氧化硫的上升；在广州对污染物几乎无影响。

在夏季，北京的二氧化碳浓度会增加，臭氧会减少；南京除了臭氧减少外其余污染物浓度都增加（Fig.19）；在广州降温会造成臭氧的减少，PM2.5，PM10 以及一氧化碳，二氧化氮的增加，对二氧化硫没有影响。



Fig.19. 南京六月平均温度下降前后污染物的变化

- 湿度对不同城市不同季节的影响：在冬季（1月），当湿度增加时，北京的一氧化碳，臭氧，PM2.5 和 PM10 都增加。其他无明显变化； 南京在湿度增加时，二氧化氮和一氧化碳都减少，臭氧增加；广州除了一氧化碳增加外，其他都减少（Fig.20）。在夏季，在北京，PM2.5 和一氧化碳都增加，臭氧减少，其他无明显变化；在南京，当湿度增加时，二氧化氮和一氧化碳减少，臭氧增多，其他无明显变化；广州的 PM2.5，PM10 和二氧化硫，臭氧都减少。



Fig.20. 广州 1 月平均湿度增加前后污染物的变化

七、讨论与总结

7.1 讨论

当我们有能力获取和存贮越来越多的数据时，以数据驱动来做决策显得愈发迫切，因此展示数据的方式比以往变得更加重要。无论是为了验证猜想还是展示研究成果，可交互的可视化平台都是一个很好的选择。有效的可视化能帮助非专家用户了解原本晦涩难懂的数据，也可以帮助专家用户去发现大量数据中隐含的潜在关系。我们以大气数据和污染物数据为基础的可视化平台能迅速帮助用户了解不同地区不同时间的污染水平同时可以探索两类数据的关系。

7.2 总结

本作品通过围绕天气情况和污染物进行了时间和空间的可视化和分析，得到了如下重要结论：

1. 六种污染物的全年分布均存在季节相关性，臭氧和其他污染污染物呈相反的趋势。
2. 不同城市不同污染物严重情况差异明显。
3. 不同地区大气数据对不同污染物的影响有明显差异。

简言之，我们认为该系统能很好的帮助相关领域的专家和学者了解全年全国的污染状况，并能快速看到模型的预测结果以及了解可能采取的污染措施的优劣。基本达到了设计目标，它可以被广泛运用到与时间空间相关数据预测模型的展示上去。

参考文献：

Feng, Z. and Kobayashi, K., 2009. Assessing the impacts of current and future concentrations of surface ozone on crop yield with meta-analysis. *Atmospheric Environment*, 43(8), pp.1510-1519.

Fuhrer, J., 2009. Ozone risk for crops and pastures in present and future climates. *Naturwissenschaften*, 96(2), pp.173-194.

Chua, S.Y., Khawaja, A.P., Morgan, J., Strouthidis, N., Reisman, C., Dick, A.D., Khaw, P.T., Patel, P.J. and Foster, P.J., 2019. The relationship between ambient atmospheric fine particulate matter (PM_{2.5}) and glaucoma in a large community cohort. *Investigative ophthalmology & visual science*, 60(14), pp.4915-4923.

Rajput, R., 2015. Understanding hair loss due to air pollution and the approach to management. *Hair Ther Transplant*, 5(133), p.2.

Sun, Z. and Zhu, D., 2019. Exposure to outdoor air pollution and its human health outcomes: A scoping review. *PloS one*, 14(5), p.e0216550.

S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37-45, 2018.

TAYLOR, SEAN, J., LETHAM & BENJAMIN 2018. Forecasting at Scale. *American Statistician*.

Erickson, Nick, et al. "AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data." *arXiv preprint arXiv:2003.06505* (2020).