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：　浙江工业大学-王攸妍

作品名称：　AirPuVis

作品主题关键词：　NFM 污染变化模式挖掘 污染事件检测 重污染分析

团队成员：　王攸妍，浙江工业大学，ping_o96@qq.com，队长

许珂，浙江工业大学，921143363@qq.com

王思超，浙江工业大学，553436947@qq.com

盛祎琛，浙江工业大学，1146833328@qq.com

汤颖，浙江工业大学，7215525@qq.com，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，echarts，pycharm，VScode，Vue, Element-ui

共计耗费时间（人天）：　120 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：

空气污染是普通民众无可奈何，却又与人民生活息息相关的问题。在新冠疫情出现前，只有严重的空气污染，才能成为人们纷纷带上口罩的决定性因素，左右着群众的生活方式。位卑未敢忘忧国，不止国家，我们普通民众也十分关注空气污染问题，也由衷希望我们都能实现呼吸自由。

为探索区域污染物随时间的变化和变化的模式，我们设计并实现了 AirPuVis，以真实的大气污染数据集为基础，向用户呈现了全国大气污染的整体分布态势和污染物的变化模式，为空气污染的分析与治理提供了分析平台。

二、数据介绍：

1、总体说明

使用的大气污染数据集为中国科学院大气物理研究所等单位发布的中国高分辨率大气污染再分析数据小时值数据集，包括我国《环境空气质量标准》中的六项常规污染物的网格化数据。数据集为同化融合中国环境监测总站（CNEMC）的国家环境空气质量监测网和嵌套网格空气质量预报模式（NAQPMS）的再分析数据。

小时分析数据：小时值数据

时间范围	2013 年-2018 年，每年 1 月份	共 6 个 zip 文件，分别为 2013~2018 年 1 月的小时数据。采取 csv 文件格式，分隔符为逗号。每个 csv 文件以所在的小时命名，包含该小时污染物浓度和气象要素值。
空间范围	中国	
空间分辨率	15 公里	
位置变量	网格中心纬度（Lat）、网格中心精度（Lon）	
污染物变量	细颗粒物（$PM_{2.5}$）、可吸入颗粒物（PM_{10}）、臭氧（O_3）、一氧化碳（CO）、二氧化硫（SO_2）、二氧化氮（NO_2）	
气象变量	纬向风速(U)、经向风速（V）、温度（TEMP）、相对湿度（RH）、地面气压（PSFC）	

2、变量具体说明

每个 csv 文件总共包含 13 列，分别给出了我国 6 项常规污染物、5 个常用气象要素以及所在网格点的经纬度值。数据集中包含的变量及其具体说明如下表所示：

变量名	释义
$PM_{2.5}$	空气动力学直径小于 2.5 微米的颗粒物，单位：微克每立方米。 形成灰霾污染的关键污染物。
PM_{10}	空气动力学直径小于 10 微米的颗粒物，单位：微克每立方米。 在环境空气中长期飘浮的悬浮微粒，对大气能见度影响很大。
O_3	臭氧，单位：微克每立方米。 光化学污染的关键污染物。
CO	一氧化碳，单位：微克每立方米。 含碳物质燃烧不完全时的产物，高浓度时能使人出现不同程度中毒症状。
SO_2	二氧化硫，单位：微克每立方米。 PM2.5 的重要前体物之一，可在大气中被氧化形成硫酸盐气溶胶。
NO_2	二氧化氮，单位：微克每立方米。 PM2.5 和 O3 的重要前体物，一方面可在大气中被氧化形成硝酸盐气溶胶，另一方面参与大气光化学反应形成臭氧。
U	纬向风速，单位：米每秒。 影响大气污染物浓度变化的关键气象要素之一。
V	经向风速，单位：米每秒。 影响大气污染物浓度变化的关键气象要素之一。
TEMP	气温，单位：开尔文（K）。 影响大气污染物浓度变化的关键气象要素之一。
RH	相对湿度，指空气中水汽压与相同温度下饱和水汽压的百分比。 影响大气污染物浓度变化的关键气象要素之一。
PSFC	地面气压，单位帕斯卡（Pa）。 影响大气污染物浓度变化的关键气象要素之一。
Lat	网格中心纬度（北纬），单位：°
Lon	网格中心经度（东经），单位：°

三、分析任务与可视分析总体流程

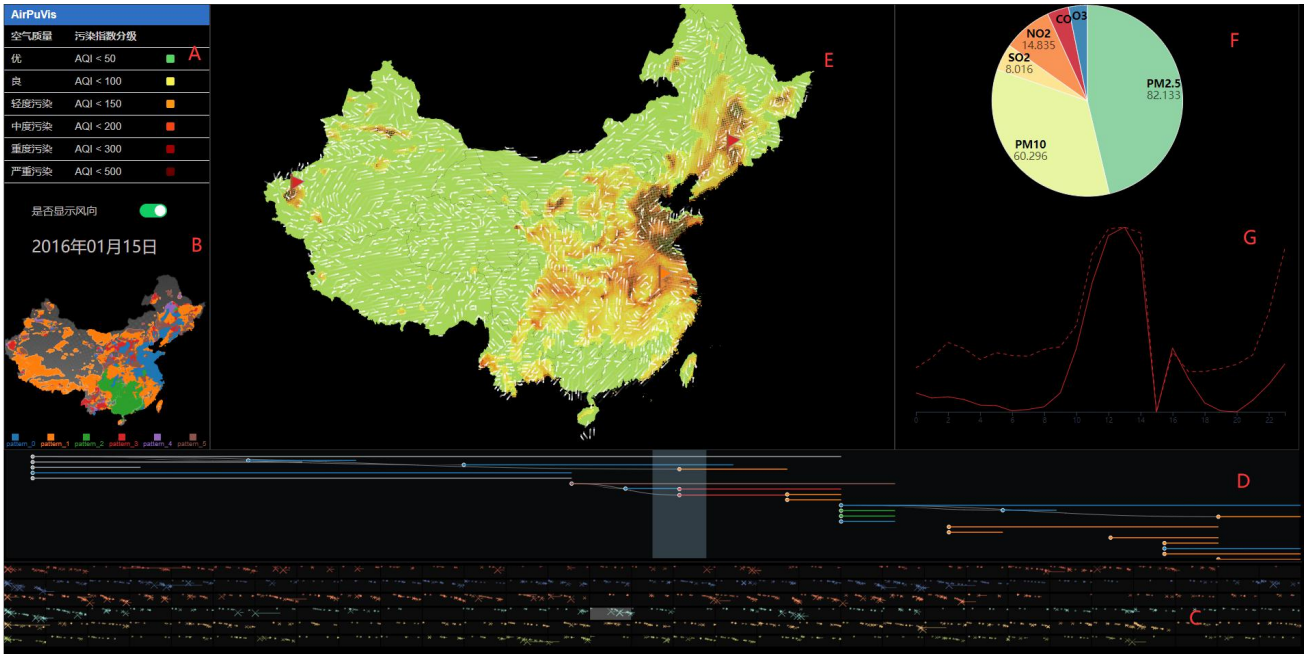


图 1　AirPuVis 空气污染可视分析系统的总体概览。该系统由 5 个部分组成，视图 A 为空气污染指数导览，视图 B 为污染变化模式分布地图，视图 C 为污染事件时间域，视图 D 为污染细节时间带，视图 E 为空气污染分布图，视图 F 展示了某时刻的主要污染源，视图 G 展示了相应 AQI 变化模式的趋势曲线以及该地区实际 AQI 变化曲线，以验证本文对污染模式提取的有效性。

分析任务：

任务 1. 利用可视分析技术，从大量空气污染数据中检测污染事件发生的时空信息

任务 2. 通过非负矩阵分解（NFM），挖掘重污染区域的空气污染变化模式

可视分析流程：

本系统以普通民众最为关注的重污染区域作为分析入口，旨在观察污染变化及其变化模式。通过层次化的时间探索，用户可以从本系统中发现感兴趣的重污染区域，并查看其 AQI 在 24 小时内的变化过程。

对于重污染区块，我们记录其出现时点与存续时间，同时关注其区块大小和污染严重程度，以块状时间域的形式呈现每天的污染事件。

用户通过鼠标交互的方式，可在视图 C 中选取感兴趣的时间域，如污染事件频发的时间域或重污染区块长期存续的时间域，我们将在视图 D 中为其呈现更为详实的 24 小时细节时间带，不同的颜色标记重污染区块的污染变化模式，并在时间带上透明化相邻时间段不同重污染区块间的可能联系，例如 n+1 时刻的重污染区块 q，可能是由时间段 n 的重污染区块 p 移动得到的，或是由 p 在扩展过程中变化形成的。

接着，用户可在时间带上选择具体的时间段，将在视图 B 中展示该时刻的污染分布地图，此时 AirPuVis 的主视图，即视图 E 将会呈现所选时刻的全国空气污染分布图，动态变化的蓝色路径绘出了风的流向，并通过地图上的小旗子标记出了该时刻出现的重污染区块，旗帜颜色代表了区块的污染变化模式。

点击旗帜，左侧的分析视图区域（视图 F，视图 G）将会呈现该重污染区块在 24 小时内的实际污染变化曲线（虚线表示）和该区域所属模式的变化趋势曲线（实线表示），以及该污染区域的各项污染物组成。

四、数据处理与算法模型

1.原始数据处理：

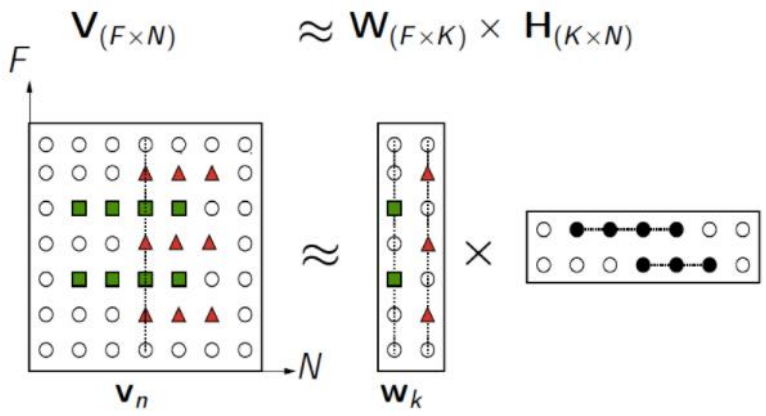
将原始数据集中的空气污染物浓度统一转化为 AQI（空气污染指数），并标记其主要污染物，为了提取重污染区域在不同时间下的相互关系和绘制全国风向流图，我们对每一区块分别计算其与上一时刻的 AQI 差值判断污染变化情况，以及计算风向和风速来提取风的流动方向和风在该方向上的速度。

2.非负矩阵分解

提取每个区域的特征时，我们希望将每个区域按照 AQI 数值进行聚类，因此我们采取了非负矩阵分解方法。不同于传统的聚类算法， NMF 是一种软聚类方法，也就是一个区域可以被分为多种类型，而不是非此即彼。

由于气象数据在以天为单位的情况下变化过大，无法有效提取其变化规律，因此，我们考虑按照 24 小时来挖掘其特征。首先构造 V=id*hour（42249*24）矩阵，用 NMF 分解后得到 W（42249*k）和 H（k*24）两个矩阵。其中 k 为需要人为指定的模式数。从数据拟合和模型复杂性来考虑，k 的取值应当适中，且以数据实际情况为参考，最终 k 取值为 6。

将 W（区域-模式）和 H（模式-小时）分别解释为每个区域 AQI 在 6 种模式下的强度，和每种模式随小时变化的趋势。在画出对应的视图并检查了一年中 50%的区域-模式分布后，我们发现每个模式都有一定区域的重合，以 pattern1 包含区域最大。因此我们提取了 pattern1 的强度做统计，呈正态分布趋势，算出其平均值，并过滤小于平均值的区域；为了让 pattern1 分布更加均匀，为每个区域设置主模式与副模式，通过计算每个区域主副模式的差值的平均来划分主模式为 pattern1 的区域。最终将模式均匀划分且符合模式-小时曲线分布。



3. 计算相对位置

对于每一个区块，我们通过经纬度计算其与其他区块的距离和方向，保存每一区块周围相邻的 8 个区块信息，以便于提取相互关系和后续计算工作。

4. 生成风向路径

为了保证风向提取的公平性以及合理提取风向的源头，我们首先从地图上的 42249 个小区块中随机选取 1 块，向其风向的反方向追溯该风向路径的源头，再从源头沿风向搜索至终点，标记整条路径上的区块作为风的流向，并获得风的起始和终点，以此来绘制动态的风向轨迹。另外，为了降低地图中风向的密集程度，减少视觉干扰，我们对各路径周围的区块进行风向聚类。

5. 划分重污染区块

筛选所有 AQI 大于 150 的区块，通过上述相对位置的标记信息，根据其 AQI 变化范围将相连区块划分为同一个污染区域，并记录其区块数量、AQI 均值、最大最小经纬度和中心点等信息.

6. 记录重污染区域变化

对每一时刻的重污染区域，根据中心点和区块数量等信息，判断其与上一时刻相应重污染区域的联系（出现、消失、移动、扩散），并对变化类型进行标记，记录区域的存续时间。

五、可视化与交互设计

如图 1 所示，本系统主要分为 3 个部分，以及 7 个视图，以下将进行该可视化系统的详细介绍。

1. 污染事件概览图

该部分主要由视图 C（污染事件时间域）和视图 D（污染细节时间带）组成，是层次化时间探索的入口，用户可根据其兴趣选择某年的某一天，在从树图时间带中选择具体时刻进行进一步分析。

① 污染事件时间域（视图 C）：

该视图分为 6 行，一行代表一年，每行又分为 31 个小格，代表了具体的日期。每个小格中的标记×表示一个污染事件的发生；×的大小表示污染覆盖区域的面积大小；×的透明度表示该区块空气污染的严重程度；×的密集程度表示在该时间段内污染事件的出现频率，即×越密集，说明这一天出现新污染区域越频繁。为了便于比较，我们对该视图中的每个小格提供了交互放大功能，鼠标悬停在小格上方，将会高亮，并在污染细节时间带上放大其细节，如图 2 所示，详细展示了污染事件的出现与存续。点击小格，将进一步显示该日期的污染细节时间带和模式分布地图。

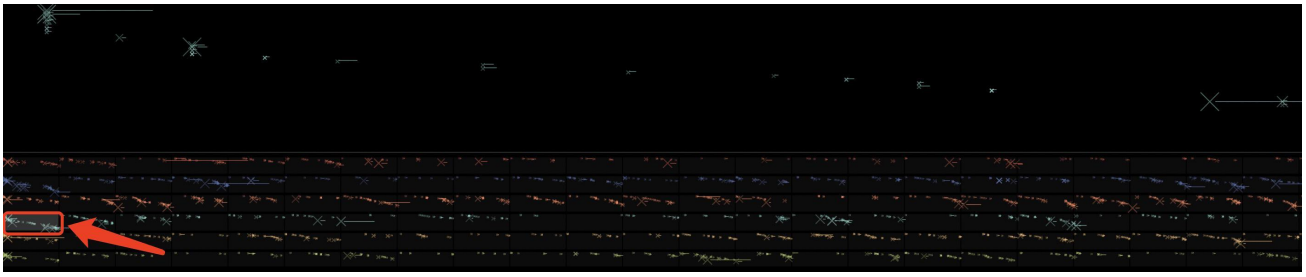


图 2 污染事件时间域及其交互放大展示

② 污染细节时间带（视图 D）：

该树图的横轴为 24 小时，每个圆点代表了一次污染事件，即新污染区域的生成；每条线段表示其存续时长，即污染区域持续的时间段；线段颜色表示该区域的污染变化模式，具体模式分布可结合视图 B 进行查看。另外，为了展示各污染事件之间的联系（例如，由一个重污染区域扩散为多个区域），我们用树形结构隐喻了其扩散关系。在该视图中，用户可点击小时区

域，系统将会在主视图，即视图 E 中显示这一时间段的全国空气污染分布状态。

图 3 污染细节时间带

2. 主视图

本系统的第二个部分为主视图，该部分共包括视图 A（空气污染指数导览），视图 B（污染变化模式分布地图），以及视图 E（全国污染情况分布地图）。

① 空气污染指数导览（视图 A）：

根据空气污染分级情况，以列表的形式向用户展示了各分级区间以及该区间的颜色分段，以便于用户结合全国污染分布情况来理解全国污染程度。

② 污染变化模式分布地图（视图 B）：

该视图展示了通过非负矩阵分解算法提取得到的 24 小时全国污染变化模式，当用户点击污染事件时间域中的某一天，视图 B 将在地图上显示这一天的六种污染变化模式所覆盖的区域，不同的颜色代表不同的模式。另外，用户若想了解相应模式的变化趋势，可在第三部分的折线图中进行详细查看。

③ 污染分布地图（视图 E）：

该视图展示了全国 AQI 数据的热力图，颜色的渐变表示 AQI 的数值大小，与污染指数导览中的分级相对应。另外，该视图提供了动态的风向轨迹，用户可结合风向与污染分布情况对污染产生原因进行初步分析；提供了污染变化模式的标记，即旗帜，该旗帜的坐标是所属重污染区域的重心坐标，旗帜的颜色表示该污染区域所对应的模式，用户可点击旗帜，在第三部分的折线图和饼图中查看对应污染区域的 AQI 变化以及各污染源的占比情况。

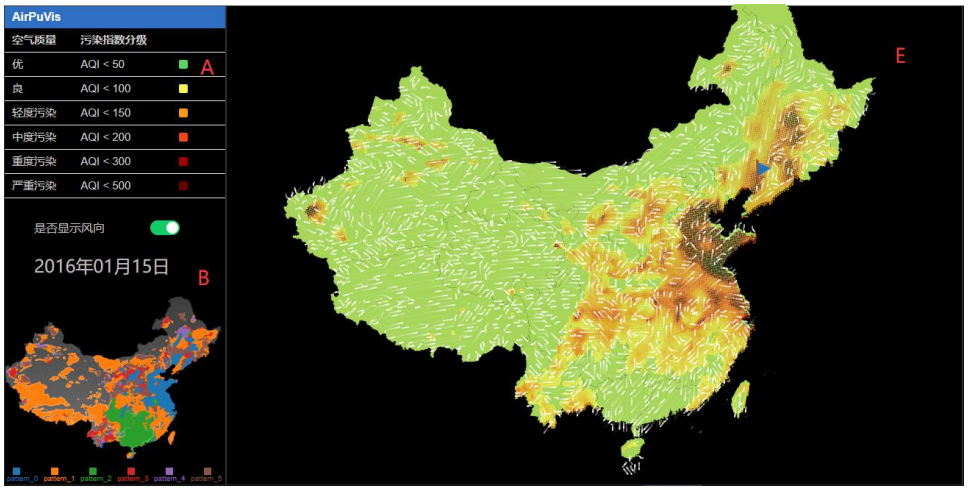


图 4 主视图

3. 细节图

该部分主要由折线图和饼图构成。

① 空气污染源占比（视图 F）：

饼图作为最常见，应用最为广泛的视图之一，能够一目了然的向用户展示所选时刻的污染物源占比情况， 了解当前时刻主要的污染源细节信息。

② 空气污染变化模式趋势图（视图 G）：

该视图展示了用户选择旗子区块所对应的 AQI 污染变化模式（实线）以及该区块真实的 AQI 数据变化（虚线）趋势，以便用户从数值变化角度比对基于 AQI 分解出的模式变化趋势与真实污染变化趋势的一致性。另外，用户可在该视图中点击不同时刻，以探索不同时刻的空气污染源占比及其变化，详细了解该时刻污染物的详细信息。

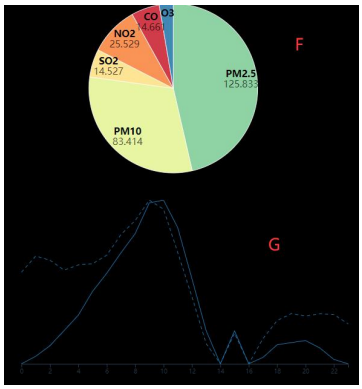


图 5 折线图，饼图

六、案例分析

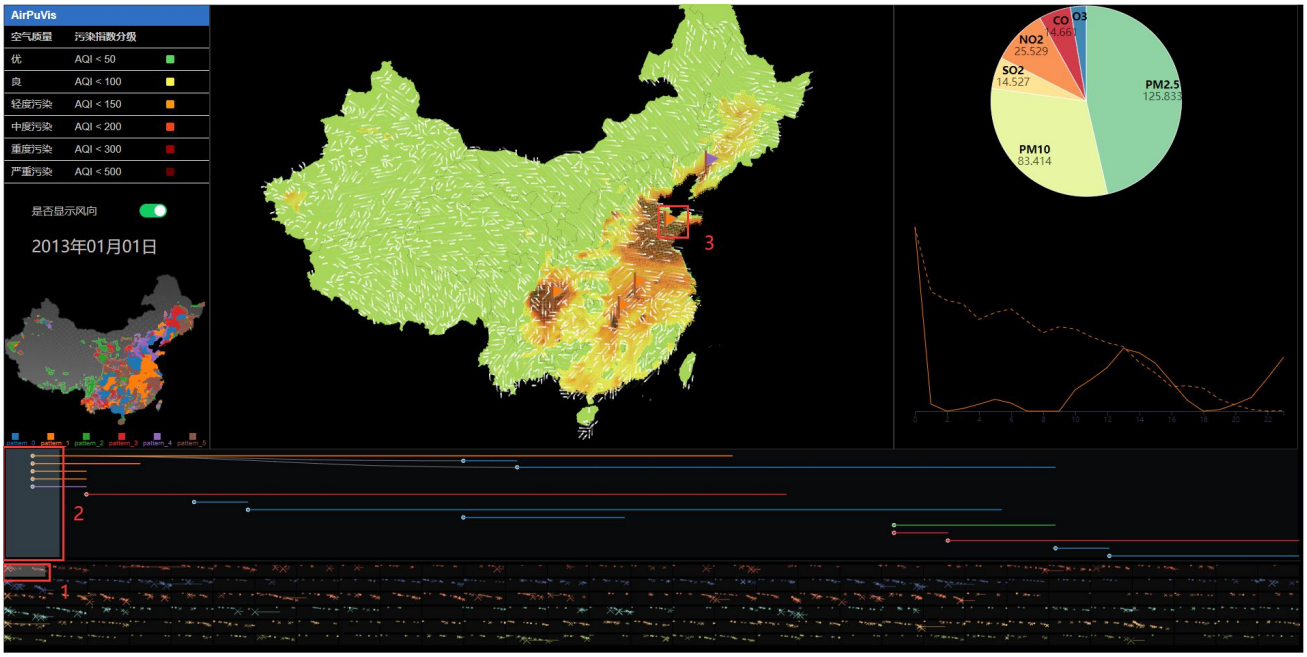


图 6 案例分析 1 操作演示及其相应视图

案例分析 1：

第一位用户是一名普通大学生，关心空气污染的影响但不了解空气污染的相关知识，也不曾接触可视化。

在进行系统体验前，我们向用户简单介绍了系统的界面以及界面中各视图的意义。用户表示想了解最早的一天的空气污染信息。在我们的指导下，他首先点击位于视图最下方区域的时间域，选取了代表 2013 年 1 月 1 日这一天的时间戳。时间域上方区域出现该日的小时时间带，呈现了该日 24 小时内每一时刻出现的重污染区块，用其所属污染变化模式的颜色标识代表区块出现节点的“X”及其延伸直线的颜色。

用户点击 0 点时刻，看到地图上出现了一些小旗。这些小旗插在这一时刻出现的重污染区域的中心，小旗的颜色代表了这个区块所属的污染变化模式。在点击第一个小旗后，用户观察分析视图区域，在饼图中发现该时刻区域的主要污染物是 PM2.5，其次是 PM10，其他污染物占比较少。在折线图中发现，两条曲线的变化趋势在 1 点至 9 点相去甚远。我们向其解释，这是因为 2013 年 1 月 1 日 0 点是数据集的第一个时刻，这一时刻出现在地图上的重污染区块的真实出现时间在更早之前，而由上一时刻存续下来的重污染区块在经过变化后其分布范围发生变化，区块内的污染变化模式也趋于复杂，范围过大的区块由于其内部包含多种污染变化模式，

其整体的污染变化也将呈现出不规则性，需要更多的结合一些诸如风向等气象因素来综合分析其区块污染变化。

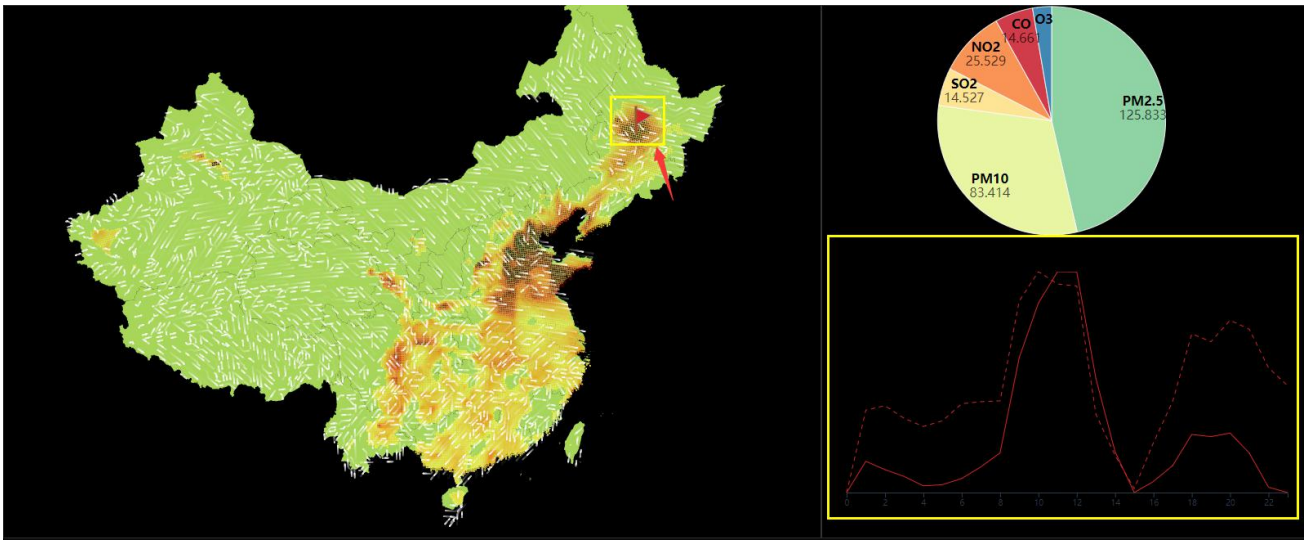


图 7 用户点击 1 时刻后，相应的视图展示

用户点击 1 点时刻，相应视图信息如图 7 所示，发现地图上出现了新的小旗，可知这一时刻出现了新的重污染区块。点击小旗，发现这一时刻两条折线的变化趋势基本吻合。我们向其解释，与上一时刻的折线不重合相比，这次折线吻合是因为该区域属于新生成的重污染区域，其规模往往处在一个较小的量级，此时这一区块在后续时间的污染变化将会呈现出明显的模式特征，用户可结合污染变化模式分布地图对区块污染变化进行针对性的观察和分析。对此，用户表示该视图确实能够在一定程度上发现污染区域变化情况。

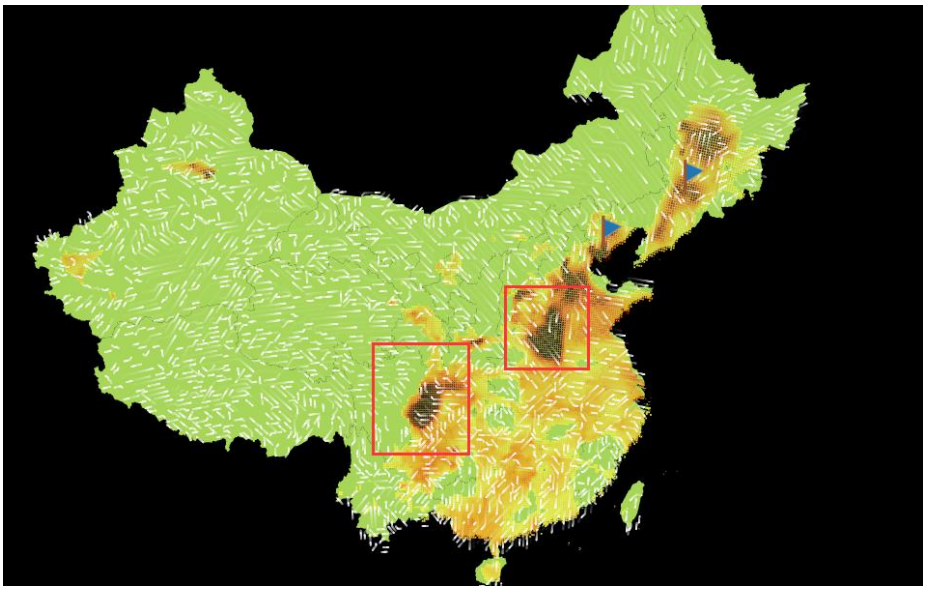


图 8 风向与污染区域移动关系

接着，用户关闭了风向路径显示，并逐一时刻的观察全国污染分布情况，发现地图中部的重污染块位置基本不变，而地图左侧的重污染块却在整体移动。此时，用户又回到了最早的时刻点并打开了风向路径进行逐一观察。如图 8 所示，发现地图中部的重污染区块刚好处于两条相向的风向路径的交汇处，而左侧的重污染块的移动轨迹基本与所在地的风向路径相合。由此，用户作出猜想，风向与污染块的成形和移动有着较大的关系，对立的风向交汇容易造成污染物成块堆积，而若是风向之间没有冲突，污染块则会沿着风向移动。

我们查阅了空气污染传输的相关文献，发现风向这一因素对空气污染传输有着紧密的联系。

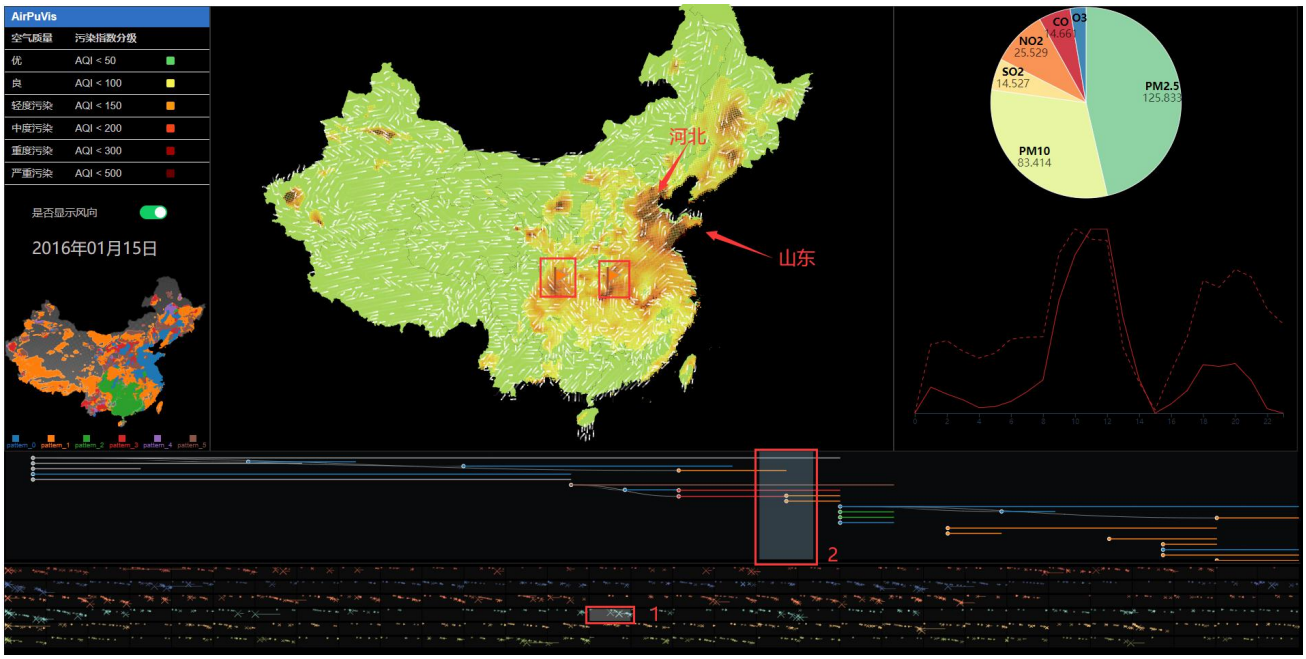


图 9 用户 2 操作流程及其相应视图信息

案例分析 2：

第二位用户是从事可视化研究工作的研究生，操作情况以及相应信息如图 9 所示。该用户首先从污染事件概览图中选择了污染标记较大，颜色较深的区块，因为该区块可能是污染较为严重的一天。接着，系统将在污染树图中展示这一天的污染区域随 24 小时的变化情况。用户发现该视图中的节点较多，且线段跨度较长，可一目了然地获知这一天新出现的污染区域较多且污染持续时间较长的隐喻。

随后，用户将节点与 AQI 变化模式分布图结合起来进行分析，选择了分布区域较大的黄色节点进一步查看，在该时刻的污染分布地图中，用户发现污染较重的区域基本集中在黑河—腾冲线的下方，也就中国的东部地区，尤其是渤海周围的山东和河北两个城市，而黄面小旗所标记的新出现的污染区域则次之。结合风向来看，用户猜测小黄旗区域的污染物应该是本地产生的，因为该区域的上风地区污染物较轻，而下风方向的地区污染较重。

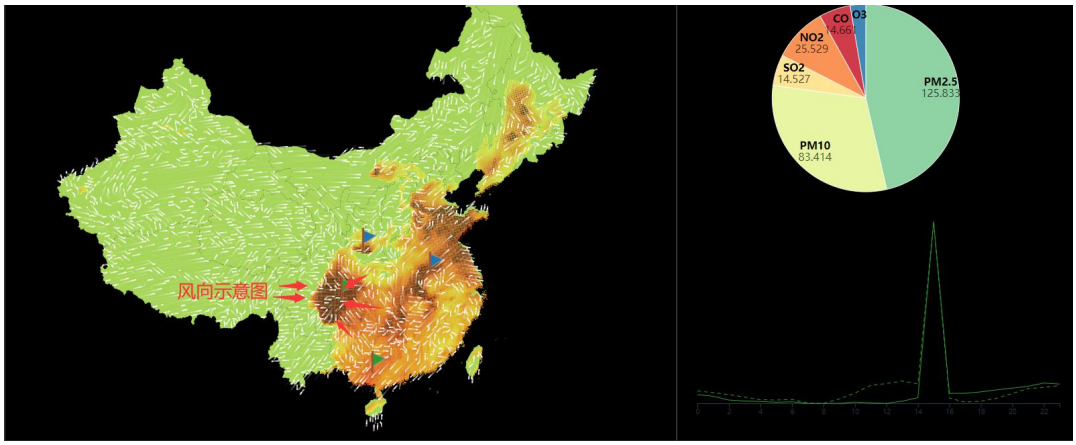


图 10 风向与污染传输关系探索

为了进一步验证这个猜想，用户选择了该时刻之后的多个时间点进行查看，发现往后的一个小时里，该区域周边城市的污染情况均比上一时刻严重，如图 10 所示，尤其在绿色小旗标记的四川中部，污染浓度急剧上升，仔细查看风向后，用户猜测四川中部的污染物应该不属于本地产生，而是属于上风地区传播堆积所致，因为四川中部地区左右两边的风向刚好相反，而上一时刻的污染区域逐渐消失。最后，用户查看了四川中部地区这个绿色小旗子的污染模式，发现其真实 AQI 变化曲线（虚线）与我们通过 NFM 算法所挖掘出的模式变化曲线（实线）几乎一致，侧面证明了我们的模式挖掘方法是有效的。

七、讨论与总结

本次参赛作品 AirPuVis 以真实的大气污染数据集为基础，实现了重污染区域的空间定位以及存续时间的检测，并将优化后的非负矩阵分解算法应用于空气污染变化模式的挖掘和提取。

基于以上分析方法，我们设计并实现了一个重污染事件检测的可视分析系统，向用户展示了全国大气污染的整体分布态势、全国风向流动情况以及重污染区域的时空变化模式，通过层次化的时间探索，用户能够高效地发现六年中污染区域较大，污染事件频发的日期，并通过污染细节时间带进一步查看 24 小时的污染变化情况，以及该时间段内的主要污染物成分。另外，用户可将全国重污染分布地图与风向流图进行结合分析，探索污染物的来源以及污染事件发生的原因。

最终，我们通过两个测试用户的案例分析，证明了该可视化系统通过 NFM 算法提取重污染区域的污染变化模式的有效性以及大气污染时空分布变化情况的可行性，并通过多个视图联动，在一定程度上分析大气污染事件产生的原因。