基于城市群的空气质量数据的可视分析方法研究

孙国道, 胡亚娟, 蒋 莉*, 姜晓睿, 梁荣华

(浙江工业大学信息工程学院 杭州 310023)
(jl@zjut.edu.cn)

摘 要: 空气污染正迅速成为一个重要的社会问题, 并越来越受到公众和科学界的关注和重视. 不同城市间的空气质量如何相互影响? 不同城市的空气质量属性间有哪些相似性和差异性, 以及哪些因素会影响空气质量的相互作用? 为了回答这些复杂的问题, 提出一个全面的基于城市群分布的可视分析系统以探索城市群的动态演变过程. 该方法包括 Voronoi 图以展示城市群在空间上的分布以及在空间上的演变过程, 嵌入式线条堆栈图以展示城市群在时间上的演变过程, 平行坐标视图以展示城市群的污染情况. 文中方法已用于城市空气质量的真实数据分析中, 该系统可以用来探索城市群的演变, 并为城市群的联合污染防治提供必要的基础.

关键词: 可视分析; 城市群; 空气质量数据
中图法分类号: TP391.41

Urban Agglomerations-Based Visual Analysis of Air Quality Data

Sun Guodao, Hu Yajuan, Jiang Li*, Jiang Xiaorui, and Liang Ronghua

(*College of Information Engineering, Zhejiang University of Technology, Hangzhou* 310023)

Abstract: Air pollution has gradually become an important issue, and has attracted more and more attention from the general public and the scientific community. How do different cities interact with each other with respect to the air quality issue? What is the similarity and difference among the air quality attributes of distinct cities, and what factors may influence the interaction? We answer these intricate questions by proposing a comprehensive visual analysis system to explore the evolution of an urban agglomeration. The method proposes a Voronoi diagram to show the spatial distribution of urban agglomeration and the evolution of urban agglomeration in space, a stacked graph with threads embedded to show urban agglomeration in the time evolution process, and a parallel coordinates diagram to show the pollution situation of urban agglomerations. We have applied our method to the real urban air quality data. Our system was used to explore the evolution of urban agglomeration and provide a necessary foundation for united pollution prevention of urban agglomerations.

Key words: visual analysis; urban agglomeration; air quality data

随着我国经济、工业化和城镇一体化的快速发展, 空气中污染物的浓度不断增加, 污染范围逐渐扩大[1-2]. 2015 年 12 月, 全国空气污染城市达 6 成, 多地首发红色警报. 中国大部分城市经常笼罩在雾霾之下, 在过去的几年里, 导致空气质量达标天数呈现下降趋势. 雾霾已经成为中国一个急需

收稿日期: 2016-11-06; 修回日期: 2016-11-18. 基金项目: 国家自然科学基金项目(61402412, 61602409); (LY14F020016, LR14F020002); 孙国道(1988—), 男, 博士, 讲师, 主要研究方向为信息可视化; 胡亚娟(1990—), 女, 硕士研究生, 主要研究方向为信息可视化; 蒋 莉(1976—), 女, 硕士, 副教授, 论文通讯作者, 主要研究方向为图形与图像处理; 姜晓睿(1983—), 男, 博士, 讲师, 主要研究方向为数据挖掘、信息可视化、可视分析、科学计量学等; 梁荣华(1974—), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究方向为可视化和计算机图形图像处理.

解决的问题，是当今中国最严重的环境问题之一[3]. 雾霾不仅威胁人体的健康[4]，还阻碍工业、农业的发展[5]. 对空气污染的研究不仅可以提高人类的环保意识，而且有助于国家和地方政府作出相应的决策来改善环境. 为了改善空气质量，环保部门和政府部门制定了相关的法律法规. 然而，影响一个地区或城市空气质量的因素非常复杂[6-7]，包括人类活动、地理、地形和气象等因素. 而且，城市的空气质量也受周边城市空气质量的影响. 例如，风从 PM2.5 浓度低的地区吹来，可能会降低目的区域的 PM2.5 浓度；而如果一个地区的 PM2.5 浓度很高，那么，其周围城市的空气质量很可能会受到该地区的影响而变差. 同时，风向、湿度、降水和温度等的季节性变化，也给空气质量的分析带来了挑战.

研究人员已经提出了不同的分析和可视化方法来研究空气质量数据的分布[8-9]、影响空气质量潜在的原因[9-10]和空气质量数据的评估[11]. 然而，这些工作忽视了不同城市间空气质量复杂的相互作用，或者不能直观地解释分析结果.

在本文提出了城市群的概念，即把所有的城市作为一个整体来研究城市的动态演变和不同城市间空气质量的相互影响. 首先采用来自城市群的领域知识，基于地理距离和空气质量属性，如用空气质量指数(air quality index, AQI)和 PM2.5 等的时间相关性来对不同城市进行聚类. 用 Voronoi 图来可视化城市群的聚类情况，不同城市群用不同的颜色编码. 由于不同的集群可能会动态地收敛或发散，为了避免在不同的时间内相同的城市群用不同的颜色编码，进一步提出了颜色一致性方案. 为了对城市群的收敛或发散行为提供一个概览，提出了一个嵌入式线条堆栈图来直观地可视化城市群随时间的演变和相应城市的模式转换. 因此，空气污染的防治可以从单个城市的防治规划上升到城市群联合进行防治，以达到改善城市空气质量的目的.

本文的主要贡献如下：

- 以城市群为单位来研究城市空气质量问题，并且对城市群的动态模式转变提供了一个全面的可视化，以更好地检测不同城市间的相关性.
- 提出了一个可视分析系统来研究城市群的演变. 该系统包括基于 Voronoi 的地理视图、嵌入式线条堆栈图和平行坐标图.
- 对城市群的动态演变以及与空气质量数据的相关性提出了深刻的见解.

1 相关工作

近年来，随着空气污染问题的加剧，许多研究者都对空气污染问题进行了大量研究，他们所使用的方法包括数据挖掘分析方法和可视分析方法. 本节概括了关于上述 2 种方法的相关工作.

1.1 空气质量数据挖掘与分析

数据挖掘和机器学习技术已广泛应用于空气质量数据的分析. 李文杰等[12]从季节与月平均处理、不同空气污染级别分别处理和空气污染过程选取处理 3 个角度处理空气污染指数(air pollution index, API)和对应的气象要素数据，借助 SPSS17.0 软件，采用相关分析法探究 API 与气象要素数据之间的关系. 袁博等[13]根据矢量距离和地理距离，采用逐步聚类方法对各城市空气 API 指数进行分析，以季节为时间尺度分析聚类形成的城市群的季节变化特点. Wang 等[14]对中国城市 PM2.5 数据进行处理，以月、季节为时间尺度，研究 PM2.5 在时间上的变化规律；以月、季节、年为时间尺度，研究 PM2.5 在空间上的分布规律. Zheng 等[15]结合空气质量数据、气象、交通流、人类的移动性、道路网络的结构和感兴趣的点，提出了一种基于协同训练框架的半监督学习方法，它包括 2 个独立的分类器. 上述所有技术都支持对空气质量数据的智能与自动分析，并在时间和空间方面找到特定的模式. 然而，上述的相关工作都没有对结果作出直观的解释、对所有的模式作出时空方面的概览. 本文工作不仅能根据地理距离和空气质量数据的不同属性自动分析不同城市间的关系，而且关于模式提供了一个直观的可视化视图.

1.2 空气质量数据的可视分析

对于空气质量数据的分析有很多不同的可视分析方法. Qu 等[8]提出了采用几个可视化工具分析香港的雾霾问题，例如，S 形的平行坐标、嵌入圆形像素条的极坐标系统和加权完全图. Li 等[9]结合空气质量数据和气象数据提出了一个新的多维视图，可视分析了中国的雾霾问题；还提出了一个相关性探测视图，对空气质量的变化形式和不同城市的气象数据属性进行了可视化. Liao 等[16]提出了一个基于网络的可视化分析系统来监测北京市的空气质量数据，综合利用不同的视图，如 GIS

视图、散点视图和平行坐标视图来提供多维分析. Engel等[10]对可视分析的设计选择和矩阵分解在空气质量研究中进行了验证. 然而，大多数相关的工作只集中在可视分析一个单一的城市或空气质量本身，没有把不同的城市作为一个整体研究，或没有考虑不同的城市之间的相互影响. 本文工作以城市群为单位来研究城市空气质量数据，可从全局上帮助理解空气质量演化的动态模式.

2　方法概览

2.1　数据源

我们抓取的空气质量数据以小时为单位，每小时的数据组成一个文件，该数据是从 2013 年 12 月到 2016 年 9 月，共包括 312 个城市和 1000 多个监测站. 数据每小时实时地发布于 PM25.in 网站①. 此网站的数据来源于国家环境保护部网站，提供的是按新的《环境空气质量标准》(GB3095-2012)发布的环境空气质量指数(AQI)相关数据. 表 1 概括了空气质量数据的属性，包括 15 个维度，其中“_24 h”代表了某个属性 24 h 的滑动平均值，“_8 h”代表了某个属性 8 h 的滑动平均值，“_8h_24h”代表了某个属性日最大 8 h 的滑动平均值. 城市的地理位置由城市的经纬度坐标可视化，一个城市的经纬度坐标是这个城市内所有监测站经纬度坐标的平均值，而监测站的经纬度坐标是在 PM25.in 网站①爬虫获取的.

表 1　监测站收集的数据的属性

序数	属性	单位
1	AQI	μg/m^3
2	PM2.5	μg/m^3
3	SO_2	μg/m^3
4	SO_2_24 h	μg/m^3
5	NO_2	μg/m^3
6	NO_2_24 h	μg/m^3
7	PM10	μg/m^3
8	PM10_24 h	μg/m^3
9	CO	mg/m^3
10	CO_24 h	mg/m^3
11	O_3	μg/m^3
12	O_3_24 h	μg/m^3
13	O_3_8 h	μg/m^3
14	O_3_8 h_24 h	μg/m^3
15	PM2.5_24 h	μg/m^3

① http://www.pm25.in/api/

2.2　可视分析任务

在对城市群的研究中，对于空气质量数据的分析有不同的分析任务，主要分为以下 4 类:

1) **全局概览.** 城市群在空间特征上的全局概览. 示例问题如地理位置彼此靠近的城市肯定属于同一个集群吗? 一个大城市所属的集群有多大?

2) **时序分析.** 城市群在时间特征上的动态演变. 示例问题如一个城市会一直属于一个集群吗? 何时一个城市会转换到另一个集群?

3) **模式检测.** 定位一个特定的模式，例如，在其发生的空间和时间位置的转换模式. 一个例子是当许多城市从一个集群收敛或发散时，检测其空间分布.

4) **模式比较.** 比较城市群的详细空间特征. 示例问题如随着时间的推移，城市群会保持稳定吗?

3　可视化技术

本节主要介绍了本文系统中使用的可视化技术.

3.1　基于 Voronoi 地理视图的城市群可视化

3.1.1　基于城市群的聚类

为了分析城市群的动态演变过程，并为城市群的污染联合防治提供一定的基础，我们对中国的 312 个主要城市进行聚类分析.

城市聚类之前，需要计算不同城市间的“距离”. 本文根据下列结果计算不同城市间的“距离”:

$$d(c_1,c_2)=w_1\mathrm{dist}(c_1,c_2)+w_2 r_{\mathrm{PM2.5}}(c_1,c_2)+w_3 r(c_1,c_2) \quad (1)$$

- 根据 2 个城市的经纬度坐标计算城市间的地理距离.
- 2 个城市的 PM2.5 和其他 14 个空气质量属性的时间序列的皮尔逊相关系数.

其中，$\mathrm{dist}(c_1,c_2)$是 2 个城市间的地理距离，$r_{\mathrm{PM2.5}}(c_1,c_2)$是 2 个城市间 PM2.5 时间序列的皮尔逊相关系数，$r(c_1,c_2)$是 2 个城市间其他 14 个空气质量属性时间序列的皮尔逊相关系数之和. 对应于空气质量属性有 15 个维度，默认情况下，参数 w_1 的值是 15，w_2 和 w_3 的值都是 1；还可以动态调节参数 w_1，w_2 和 w_3，以便找到使聚类效果更好的参数.

皮尔逊相关系数的范围是[–1,1]，我们把它转换为[0,1]. 地理距离的单位是 km，因此，需要对地理距离进行一定的变形操作才能将其范围调整为[0,1]，本文采用的映射关系为

$$f(z)=\frac{z-d_{\min}}{d_{\max}-d_{\min}} \tag{2}$$

其中, z 为要转换的 2 个城市间的地理距离; $d_{\max}$ 和 $d_{\min}$ 分别为所有城市中 2 个城市间地理距离的最大值和最小值. 以小时为单位实时地抓取数据, 每个小时抓取的数据组成一个文件; 以周为时间粒度来计算城市间各个属性时间序列的皮尔逊相关系数, 一周内, 每个属性的数据组成一个时间序列, 所以此序列的长度是 7×24. 由于一些数据属性中包含零值或缺失值, 这会影响计算城市间“距离”的准确性. 因此为了准确计算城市间的“距离”, 数据属性中数据值是零的或缺失的将被 2 个前后最相邻的不为零的数据的平均值所代替.

在计算不同城市间的“距离”后, 采用 K-means 算法对所有城市进行聚类. 然而, 在 K-means 聚类算法中, K 值的确定是一个棘手的问题. 为了解决这个问题, 我们从气象学、地理科学和资源研究领域提取领域知识, 因为这些领域同样也研究城市群的形成和演化.

根据中国科学院地理科学与资源研究所编制的《2010 中国城市群发展报告》的研究结果, 我国目前正在形成的城市群有 23 个. 表 2 展示了中国前十大城市群及其所包含的城市[17], 它源于地理科学和资源研究领域. 由于前十大城市群基本包含了中国大部分主要城市, 因此, 根据城市之间的污染相似性, 我们将城市聚类划分为 10 类, 即取 K=10.

除此之外, 聚类初始中心的选择是另一个挑战. 聚类中心的随机选择可能会导致一个不稳定的聚类结果. 为了提高传统的 K-means 聚类算法, 在聚类初始中心的确定上, 本文加入了语义信息. 在第一个时间段, 将表 2 中每个城市群所包含的所有城市的地理几何中心设定为每个城市群的初始聚类中心. 为了每个城市的聚类划分都具有很高的相似度, 对下一个时间段城市的聚类来说, 每个初始聚类中心设为前一个时间段内每个城市群聚类结果的地理几何中心. 它可以降低初始中心的随机性, 确保城市的每一次聚类划分都具有很高的相似性. 表 2 中不同颜色的文字代表不同的城市群, 它们的颜色编码和图 1 相同.

3.1.2 空间上聚类的可视化

经过聚类, 每个城市都隶属于一个集群, 每个集群就是一个城市群. 已有很多种可视化方法来展示城市群以及各个城市的归属, 一个简单的方法是根据城市的经纬度坐标在地图上每个城市的位置放置不同颜色的点或圆, 即用不同颜色的点或圆在地图上表示每一个城市, 其中不同的颜色编码不同的城市群. 然而, 这可能会带来一些问题: 如果点或圆太小, 许多地图空间将被浪费, 用户可能不容易发觉视觉效果. 除此之外, 点或圆并不能代表城市周围的地区. 例如, 在上海及周边地区有不同强度的城市, 如果点或圆太大, 可能会引入遮挡问题.

表 2 我国十大城市群及其包含的城市

主要城市群	包含城市
长三角城市群	上海、南京、无锡、常州、苏州、南通、扬州、镇江、泰州、杭州、宁波、嘉兴、湖州、绍兴、舟山、台州
珠三角城市群	广州、深圳、珠海、东莞、佛山、中山、惠州、江门、肇庆
京津冀城市群	北京、天津、石家庄、唐山、秦皇岛、保定、张家口、承德、沧州、廊坊
山东半岛城市群	济南、青岛、烟台、潍坊、淄博、东营、威海、日照
辽中南城市群	沈阳、大连、鞍山、抚顺、本溪、丹东、辽阳、营口、盘锦、铁岭
海峡西岸城市群	福州、厦门、漳州、泉州、莆田、宁德
中原城市群	郑州、洛阳、开封、新乡、焦作、许昌、平顶山、漯河
关中城市群	西安、咸阳、宝鸡、渭南、铜川就、商洛
长江中游城市群	武汉、黄石、鄂州、黄冈、孝感、咸宁、随州、荆门、荆州、信阳、九江、岳阳
川渝城市群	重庆、成都、自贡、泸州、德阳、绵阳、遂宁、内江、乐山、南充、眉山、宜宾、广安、雅安、资阳

为了解决上述问题, 本文提出了一种基于 Voronoi 视图[18]的城市群可视化. 图 1 是可视化的结果.

每个 Voronoi 代表相应城市的周边地区, 不同的颜色代表不同的城市群; 这样可以很直观地看到城市所属的城市群. 如果某个地区有很多城市, 那么在最终的可视化结果中就会出现很多的 Voronoi; 而且每个 Voronoi 内的红色圆点位置是由城市的经纬度坐标决定的. 每个 Voronoi 内任何一个点到此 Voronoi 内红色圆点的地理距离是它到所有 Voronoi 内红色圆点的地理距离的最小值. 基于 Voronoi 视图的可视化可以帮助确定受相应的城市影响的地区以及各城市群中城市的分布.

每个时间段城市聚类一次, 我们将在不同的

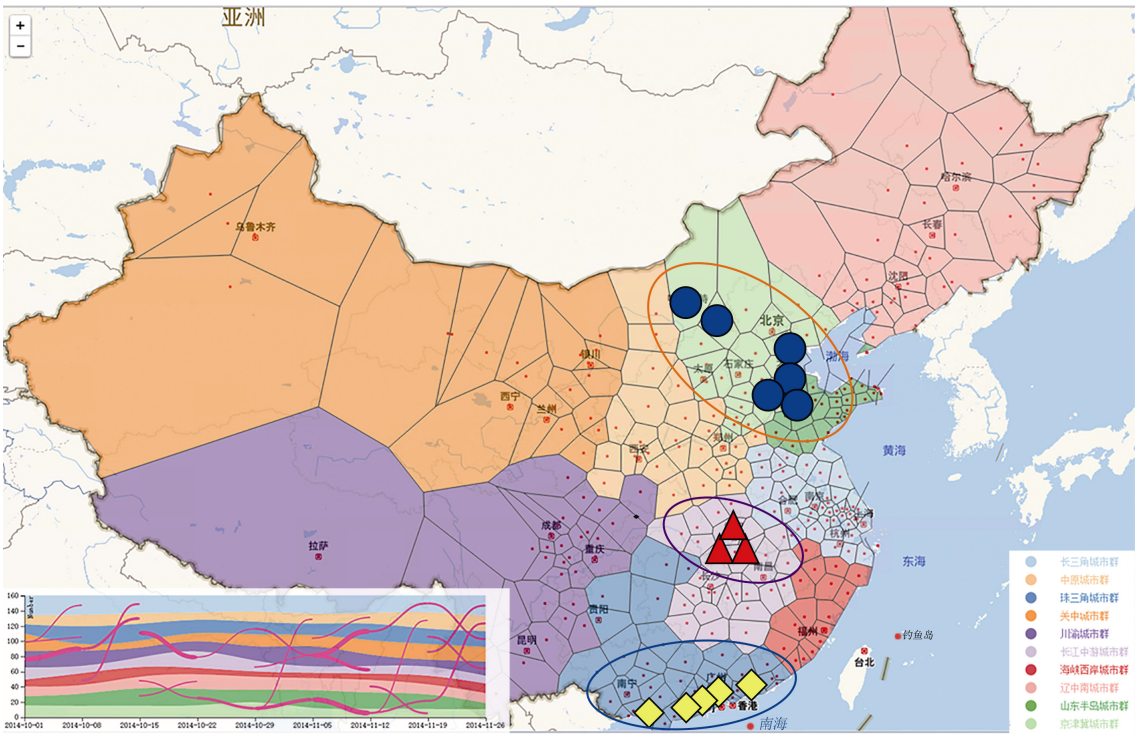
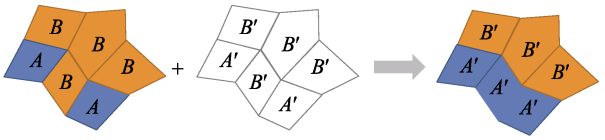


图 1　基于 Voronoi 视图的城市群聚类结果的可视化

时间段获得不同的城市群聚类结果. 为了更直观地观察和研究城市群在空间上的演变，在不同的时间段内，保持同一个城市群的颜色恒定是必需的. 然而，不同的城市可能会收敛和发散于不同的集群，这就使得色彩配置更加困难. 因此，本文提出了颜色一致性编码方案. 在第一个时间段聚类后，10 种不同的颜色代表 10 个城市群. 接下来其他时间段的城市群的颜色由

$$d = \frac{\#\text{overlap}(T_1, T_2)}{\#\max(T_1, T_2)} \tag{3}$$

决定. 其中，#overlap(T_1,T_2)是 2 个时间段 T_1 和 T_2 内 2 个城市群间所包含城市的交集个数；#max (T_1,T_2)是 2 个时间段 T_1 和 T_2 内 2 个城市群所包含城市数量的最大值. 当比值 d 最大时，T_2 时间段内城市群的颜色设定为 T_1 时间段内城市群的颜色. 这样，就可以更有效和直观地可视和分析城市群的演变. 例如，如图 2a 所示，在 T_1 时间段，有 2 个 Voronoi 为集群 A，用蓝色表示，有 4 个 Voronoi 为集群 B，用橙色表示. 如图 2b 所示，在 T_2 时间段，这 2 个集群稍有变化，集群 A'和 B'都有 3 个 Voronoi. 对于 T_1 时间段中的集群 B 和 T_2 时间段的集群 B'，共有的 Voronoi 的个数是 3，集群 B 和集群 B'所包含的最大 Voronoi 的个数是 4，运用式(3)，d 的值是 3/4. 同理，对于 T_1 时间段中的集群 A 和 T_2 时间段的集群 B'来说，d 的值是 0. 所以集群 B' 的颜色设为 T_1 时间段的集群 B 的颜色，即橙色；同样的过程也适用于集群 A. 图 2c 是颜色一致性方案的输出.



a. T_1色彩配置结果　b. T_2色彩未配置时　c. T_2色彩配置结果

图 2　相邻 2 个时间段内 Voronoi 的色彩配置

3.2　嵌入式线条堆栈图

3.1 节描述了系统的地理可视化部分，然而，这只能呈现一个时间段内城市群演化的一个快照，用户也想考察分析城市群如何随着时间而演变，一个城市如何转换其所属的城市群.

因此，本文提出了一个嵌入式线条堆栈图来可视化城市群在时间上的演变. 图 3 展示了可视化的结果；其中水平轴表示时间，从 2014-10-01—2014-11-26，以周为单位，这样城市群的聚类划分更加稳定；纵轴表示在某时间段内某个城市群所包含城市的个数. 堆栈图由 10 个条带组成，每个条带代表一个城市群；堆栈图的高度表示某个时间段内所有城市群所包含的城市的总数，而且不同条带的高度随时间而变化，其可视化了每一个城市群含有的城市个数的时间变化趋势. 在图 3 的

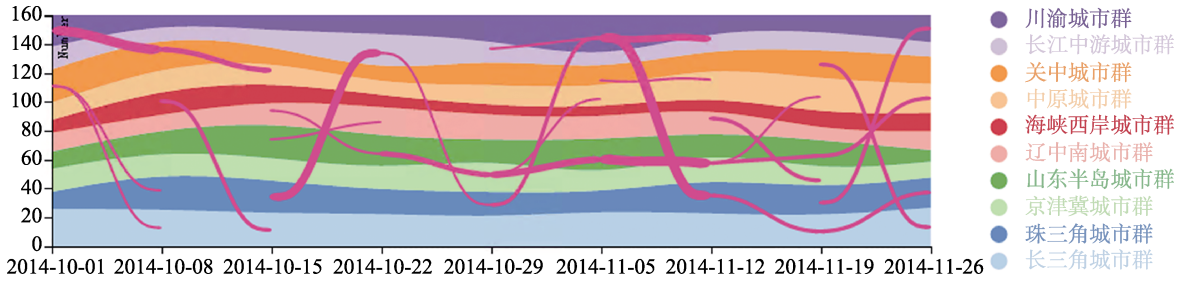


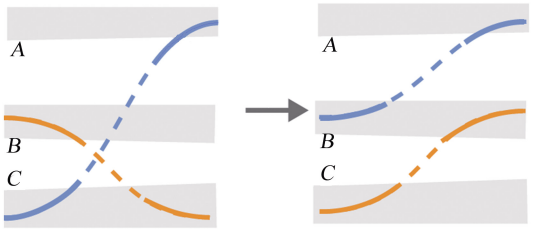
图 3 对代表城市群的条带和代表城市转换的线条重新排序前的堆栈图

右侧，圆的颜色和文本信息分别表示城市群所对应的颜色和名称. 城市群的颜色有助于我们更直观地辨别不同的城市群，这个颜色也对应于 Voronoi 视图中城市群的颜色. 用户点击堆栈图中感兴趣的任何时间段区域，就可以在 Voronoi 视图上看到这个时间段的城市群的分布，即实现了从时间到空间上的交互.

虽然传统的堆栈图[19]可以展示城市群所包含的城市数量随时间的变化，但是却不能展示城市群中具体的城市变迁. 所以，我们进一步绘制了 Bézier 曲线来编码不同城市群间城市的转换. 如图 3 所示，线条的宽度可视化了在相邻的 2 个时间段内从一个城市群转换到另一个城市群的城市数量，线条越宽，转换的城市越多.

然而，直接在不同的条带上绘制 Bézier 曲线可能会导致线条交叉. 为了缓解这一问题，我们采用了基于重心(barycentre)的方法[20]对代表城市群的条带和代表城市转换的线条重新排序，条带用灰色表示，前一个时间段条带的排序索引决定下一个时间段条带的顺序. 例如，如图 4a 所示，在第 2 个时间段条带 A 的索引是第 1 时间段内的索引 3；相似的，条带 B 和 C 的索引分别是 0 和 2. 根据索引 3, 0 和 2 对条带 A, B 和 C 按降序进行重新排序，如图 4b 所示.

图 3 是对代表城市群的条带和代表城市转换的线条重新排序之前的堆栈图，可以看到线条交叉比较严重. 用上述方法对条带和线条重新排序，经过多次重新排序，找出线条交叉数量最少的情况作为重新排序后的最终结果. 图 5 所示为重新排序后线条交叉最少的堆栈图，其中线条交叉数减少了 28%，显著地减少了视觉杂乱的现象.



a. 重新排序前视图 b. 重新排序后视图

图 4 条带和线条的重新排序

3.3 平行坐标视图

基于 Voronoi 的地理视图和嵌入式线条堆栈图着重可视化了城市群在空间和时间上的演变，但并不能展示城市群空气污染的情况. 因此，我们使用了平行坐标视图[21]来可视化城市群空气污染的情况. 平行坐标将高维数据的各个维度用一系列相互平行的坐标轴表示，每个数据项是一条数据集，包括图 6 中平行坐标轴所代表的 8 个维度的数据，每个数据项是由一个和每个轴相交的折线表示，每个折线与坐标轴交叉点的值是某时间段内某城市群所包含城市的属性值的平均值；每一条

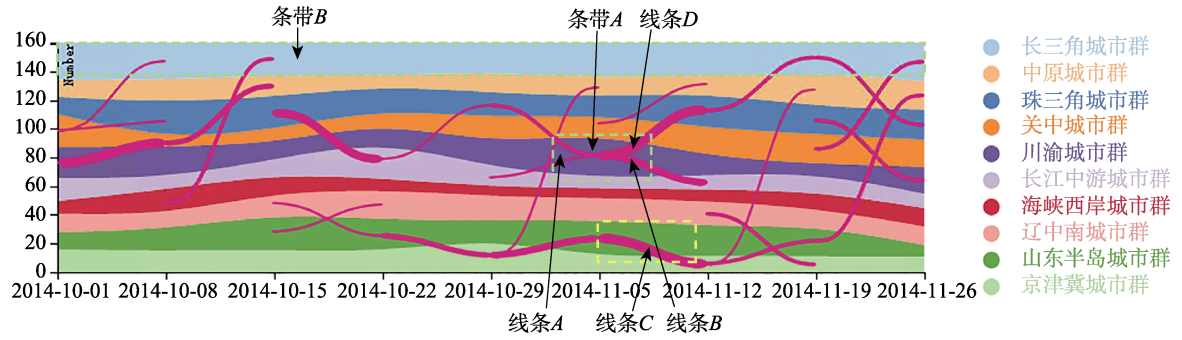


图 5 对代表城市群的条带和代表城市转换的线条重新排序后的堆栈图

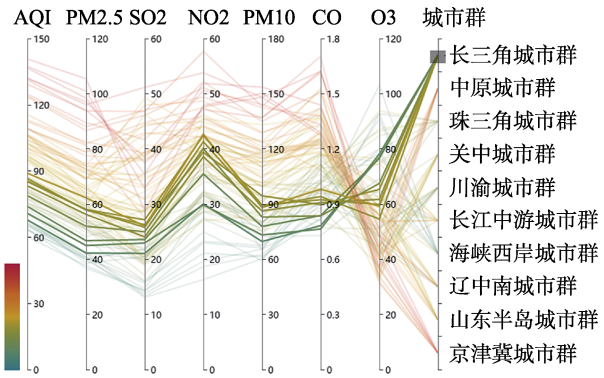


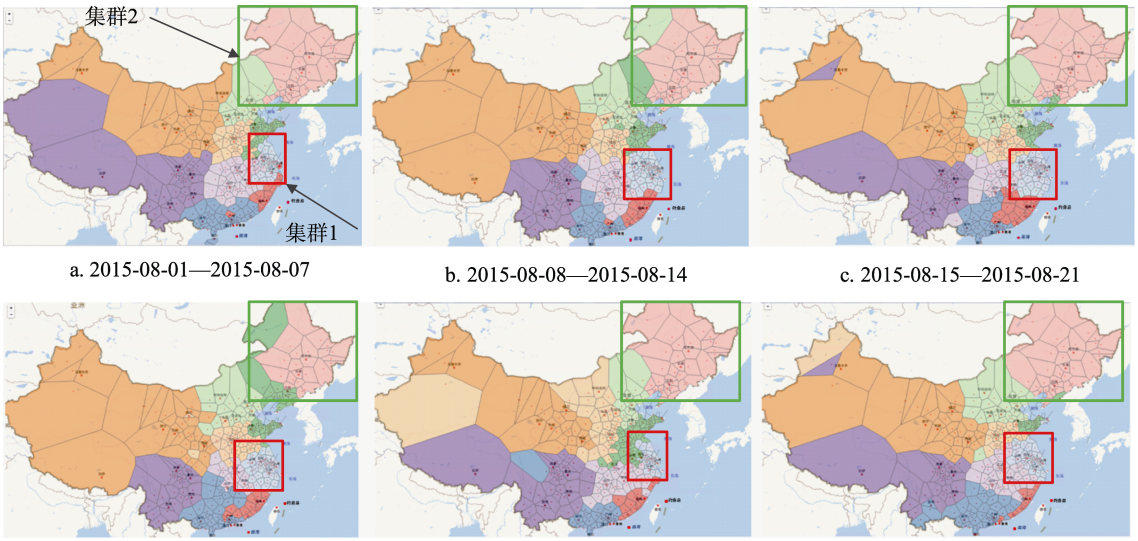
图 6　根据 PM2.5 的值映射折线颜色的平行坐标视图

折线的颜色根据 PM2.5 的值进行映射，值越小，映射颜色更接近绿色；值越大，映射颜色更接近红色. 图 6 所展示的数据是从 2014-10-01—2014-11-26，以周为单位，共 9 个时间段，所以延伸到每个城市群的折线有 9 条. 用户可以针对维度轴进行刷选，更清晰地了解局部数据的变化规律，更直观地观察感兴趣的数据，也可以只刷选某一个城市群，查看此城市群的空气污染状况. 如图 6 所示，通过刷选，我们可以看到延伸到长三角城市群的折线都比较集中，不会出现大的变动.

4　案例分析

4.1　城市群的空间聚类分析

本文系统可以帮助用户检测城市群的动态演变. 此时设式(1)中的参数 w_1=15, w_2=w_3=1，以平衡地理距离和空气质量属性所占的权重. 图 7 展示了从 2015-08-01—2015-09-11 的聚类结果，其中不同颜色代表不同城市群. 其与表 2 中的中国城市群的分布相比较，有很多相似之处,尤其是红色框中代表长三角城市群的集群 1 和绿色框中代表辽中南城市群的集群 2. 结果表明，所计算的聚类结果验证了表 2 的分类结果，也反过来证明了该方法的有效性. 我们也可以看到，一些城市收敛或发散于不同的集群，例如，拉萨、克拉玛依和齐齐哈尔等；这可能为相应城市模式转换的潜在原因提供一个直观的提示.



a. 2015-08-01—2015-08-07　　b. 2015-08-08—2015-08-14　　c. 2015-08-15—2015-08-21

d. 2015-08-22—2015-08-28　　e. 2015-08-29—2015-09-04　　f. 2015-09-05—2015-09-11

图 7　相邻 6 个时间段内城市的聚类结果

4.2　长三角城市群的模式检测

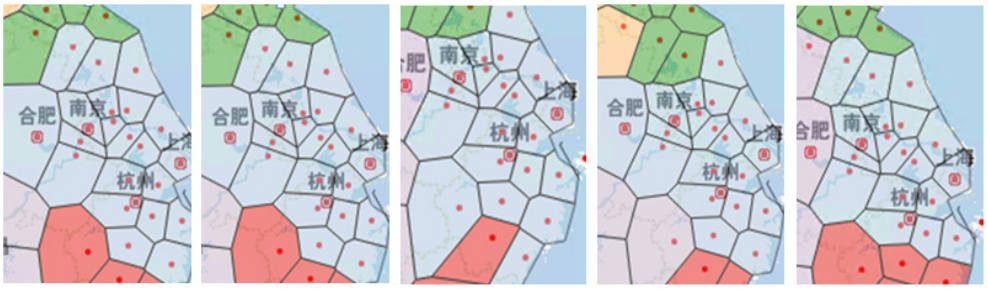
Voronoi 图可以帮助用户检测城市群的动态演变. 本文分析了长三角城市群的演变. 在图 8~10 中，长三角城市群用浅蓝色编码，展示的数据时间段为 2014-10-01—2014-11-04.

图 8 中，此时式(1)中的参数 w_1=w_2=w_3=1，各个属性的权重一样，来观察城市聚类结果. 根据式(1)计算的“ 距离 ”，长三角的一些城市聚集成一个城市群. 从图 8 所示可以看到，图 8a 中，南京、上海和杭州等周边城市聚集成长三角城市群，并和图 8b~8e 附近聚集成长三角洲城市群的中心大城市几乎是一样的，尤其是南京、杭州和上海. 我们推测这是因为这些城市在地理距离上是紧密联系在一起的，有相似的人类活动或工业模式.

为了方便进一步分析，我们可以调整参数 w_1, w_2 和 w_3. 式(1)中不同空气质量数据属性的时间

a. 10-01—10-17 b. 10-08—10-14 c. 10-15—10-21 d. 10-22—10-28 e. 10-29—11-04

图 8 式(1)中 $w_1=w_2=w_3=1$ 时 2014 年不同时间段长三角城市群的聚类结果



a. 10-01—10-17 b. 10-08—10-14 c. 10-15—10-21 d. 10-22—10-28 e. 10-29—11-04

图 9 式(1)中 $w_1=15$, $w_2=w_3=1$ 时 2014 年不同时间段长三角城市群的聚类结果

a. 10-01—10-17 b. 10-08—10-14 c. 10-15—10-21 d. 10-22—10-28 e. 10-29—11-04

图 10 式(1)中的参数 $w_1=w_3=0$, $w_2=1$ 时 2014 年不同时间段长三角城市群的聚类结果

序列的皮尔逊相关系数的范围是[0,1]，总和的范围是[0,15]，所以设 $w_1=15$，以对应 15 个空气质量属性的时间序列的皮尔逊相关系数，当然设置为很大的数值也可以，$w_2=w_3=1$. 长三角城市群的聚类结果如图 9 所示，在图 9a 中，长三角城市群主要包括南京、上海、杭州和合肥. 与图 9b~9e 相比，长三角城市群的聚类结果基本稳定，特别是杭州、上海和南京. 而且，与图 8 相比，长三角城市群的聚类结果变化不大，特别是杭州和上海. 也就是说，地理距离的大幅度增加并没有对城市群的聚类结果造成很大影响. 它证明了长三角城市群的空气质量状况是相似的，这可以为长三角城市群空气污染的联合防治提供必要的依据. 但是一些周边城市，如合肥、南昌和其他小城市，聚类结果动态地收敛或发散于不同的城市群. 我们推断这是因为城市到集聚中心的距离变大，影响了城市空气污染的传播.

PM2.5 是 GB3095-2012《环境空气质量标准》新增的监测指标，因其对雾霾贡献率较大，对人体健康危害严重，所以被视为重点监测指标之一[2]. 接下来，我们研究 PM2.5 对长三角城市群聚类的影响. 将式(1)中的参数 w_2，即 PM2.5 的时间序列的皮尔逊相关系数的参数值设为 1，设剩余参数 $w_1=w_3=0$，长三角城市群的聚类结果如图 10 所示. 图 10a 中，长三角的一些城市聚类为一个集群，包括南京、上海、杭州和合肥. 与图 10b~10e 相比可以看到聚类结果变化很小，聚集形成长三角城市群的中心附近的大城市几乎是相同的，尤其是杭州和上海.

如图 6 所示，在这相邻的几个时间段内，延伸到长三角城市群的折线比较集中，也就是长三角城市群内城市的平均 PM2.5 的值相差不大，更加

证明了长三角城市群的污染情况比较稳定，因此长三角城市群的聚类变化不大．与图 8 和图 9 相比，长三角城市群的聚类结果变化也不大，我们推测 PM2.5 是影响长三角城市群空气质量的一个重要因素，而且这些城市可能有类似的生活模式，比如，正常工作日大部分 8 点左右去上班，6 点左右下班等．这些证据可能有助于国家和地方作出决策，例如，空气质量整治可能需要考虑整个城市群而不是单一的个别城市，不同政府部门之间应该协同合作等．

4.3　城市群的时间演变

嵌入式线条堆栈图可以帮助用户研究城市群中的城市转换，并进一步研究城市群在时间上的动态演变过程．设式(1)中的参数 $w_1=w_2=w_3=1$，聚类结果如图 5 所示，它所用到的数据是从 2014-10-01—2014-11-26．

如图 5 所示，绿色虚线框中的条带 B 可视化了从 2014-10-01—2014-11-26 浅蓝色条带(长三角城市群)所包含的城市数量，它的宽度基本上是恒定的，这意味着长三角城市群所包含的城市数量是相对稳定的．我们推测这是因为这些城市是紧密联系在一起的，并且有类似的工业模式或生活模式．这一点已经在案例 1 分析过，此处不再重复．

从图 5 还可以看到，绿色虚线框中的条带 A 比两边都宽，它代表的含义是 2014-11-05 将有一些城市变换到紫色的条带(川渝城市群)，并在 2014-11-12 有一些城市变换出紫色的条带(川渝城市群)．

通过把绿色虚线框中的线条 A 和线条 B 中包含的城市可视化到地图上，发现有 3 个城市，分别是湘潭、株洲和长沙．如图 1 所示，在紫色椭圆中用红色三角形标记这 3 个城市的地理位置，这些城市在 10 月 29 日从浅紫色的条带(长江中游城市群)变换到紫色的条带(川渝城市群)；然后在 11 月 12 日又变换回浅紫色的条带(长江中游城市群)．这 3 个城市都位于湖南省，而且它们在地理上是相邻的．通过在网站[①]查询这 3 个城市的天气信息，发现这 3 个城市的风向在 10 月 29 日是北风，11 月 5 日是南或东南风，然后又转为北风．这种模式与城市群中的城市的地理流方向有关，我们推断这是因为风向在影响空气质量的传播，以及动态变化起着重要的作用．

如图 5 所示，通过把绿色虚线框中的线条 D 和线条 E 中包含的城市可视化到地图上，还发现有 5 个城市，分别是广州、北海、肇庆、云浮和梅州，如图 1 所示用深蓝色椭圆里的黄色的菱形在地图上标记这些城市的地理位置，其中北海位于广西壮族自治区，其他 4 个城市位于广东省．这些城市在地理上相邻，在 10 月 29 日从蓝色的条带(珠三角城市群)变换到紫色的条带(川渝城市群)，然后下个时间段又变换到蓝色的条带(珠三角城市群)．通过在网站[①]上查询这些城市的天气信息，发现这些城市没有固定的风向，推测是周边城市影响了这些城市所属城市群的变化．

如图 5 所示，黄色虚线框中的线条 C 和绿色虚线框中的线条 D 是所有线条中最宽的，也就是说，城市变换的数量最多．通过进一步研究，把线条 C 所包含的城市映射到地图上，发现这些城市在北京附近，如图 1 所示橙色椭圆中用蓝色的圆标记的一些城市的位置，包括天津、沧州、德州、济南、呼和浩特和大同市；我们推测其与会议期间对空气质量的整治相关．

通过查询线条 D 所包含的城市，这些城市是广州、北海、韶关、梅州等．这些城市上面已经讨论过，并推测这些城市的城市群的变化可能是受到周边城市的影响．

5　结　语

本文提出了一个全面的基于城市群的可视化方法，以分析在大量空气质量数据中存在的时空模式．本文可视分析系统中使用了基于城市群的 Voronoi 视图、嵌入式线条堆栈图以及平行坐标视图．为了保证有效的可视化和避免认知负担，本文系统中进一步采用了颜色一致性编码方案．

本文原型系统可以用来研究城市群的演变，并为国家和地方政府作出相应的决策及进一步改善环境状况提供帮助．关于空间和时间方面分析的 3 个案例研究证明了本文可视化方法的有效性．在将来，计划将继续我们的工作，并研究更多的气象数据对城市群的演变的影响，如风速、风向和温度等．

① http://tianqi.2345.com

参考文献(References):

[1] Wang Y, Sun M, Wang R, *et al*. Promoting regional sustainability by eco-province construction in China: A critical assessment[J]. Ecological Indicators, 2015, 51: 127-138

[2] Zhang Z, Xue B, Pang J, *et al*. The decoupling of resource consumption and environmental impact from economic growth in china: spatial pattern and temporal trend[J]. Sustainability, 2016, 8(3): 222

[3] Ma J, Xu X, Zhao C, *et al*. A review of atmospheric chemistry research in China: photochemical smog, haze pollution, and gas-aerosol interactions[J]. Advances in Atmospheric Sciences, 2012, 29(5): 1006-1026

[4] Chen R, Zhao Z, Kan H. Heavy smog and hospital visits in Beijing, China[J]. American Journal of Respiratory and Critical Care Medicine, 2013, 188(9): 1170-1171

[5] Chen J, Chen H, Zheng G, *et al*. Big smog meets web science: smog disaster analysis based on social media and device data on the web[C] //Proceedings of the 23rd International Conference on World Wide Web. New York: ACM Press, 2014: 505-510

[6] Li X, Yang Y, Xu X, *et al*. Air pollution from polycyclic aromatic hydrocarbons generated by human activities and their health effects in China[J]. Journal of Cleaner Production, 2016, 112: 1360-1367

[7] Cai Yijing, Li Taiping. An empirical analysis on the factors influencing the urban air quality[J]. Environmental Protection and Circular Economy, 2015, 35(2): 65-68(in Chinese)
(蔡怡静, 李太平. 城市空气质量影响因素的实证分析[J]. 环境保护与循环经济, 2015, 35(2): 65-68)

[8] Qu H M, Chan W Y, Xu A, *et al*. Visual analysis of the air pollution problem in Hong Kong[J]. IEEE Transactions on Visualization and Computer Graphics, 2007, 13(6): 1408-1415

[9] Li J, Zhao X, Zhao H, *et al*. Visual analytics of smogs in China[J]. Journal of Visualization, 2016, 19(3): 461-474

[10] Engel D, Greff K, Garth C, *et al*. Visual steering and verification of mass spectrometry data factorization in air quality research[J]. IEEE Transactions on Visualization and Computer Graphics, 2012, 18(12): 2275-2284

[11] Quinan P S, Meyer M. Visually comparing weather features in forecasts[J]. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(1): 389-398

[12] Li Wenjie, Zhang Shihuang, Gao Qingxian, *et al*. Relationship between Temporal-Spatial Distribution Pattern of Air Pollution Index and Meteorological Elements in Beijing, Tianjin and Shijiazhuang[J]. Resources Science, 2012, 34(08): 1392-1400 (in Chinese)
(李文杰, 张时煌, 高庆先, 等. 京津石三市空气质量指数(API)的时空分布特征及其与气象要素的关系[J]. 资源科学, 2012, 34(8): 1392-1400)

[13] Yuan Bo, Xiao Sulin, Jiang Dahe. Air pollution of city clusters in China and its characteristics on seasonal change[J]. Environmental Science and Technology, 2009, 22(A01): 102-106(in Chinese)
(袁 博, 肖苏林, 蒋大和. 我国城市群空气污染及其季节变化特点[J]. 环境科技, 2009, 22(A01): 102-106)

[14] Wang Z, Fang F, Xu G, *et al*. Spatial and temporal variation of the concentration of PM 2.5 in Chinese cities in 2014[J]. Journal of Geography, 2015, 70 (11): 1720-1734

[15] Zheng Y, Liu F, Hsieh H P. U-Air: When urban air quality inference meets big data[C] //Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2013: 1436-1444

[16] Liao Z F, Peng Y N, Li Y, *et al*. A web-based visual analytics system for air quality monitoring data[C] //2014 22nd International Conference on Geoinformatics. Taiwan: IEEE Press, 2014: 1-6

[17] Zhang Jiansheng. Research on environmental performance differences and its influential factors of Chinese main urban agglomerations[J]. Reform of Economic System, 2016(1): 57-62 (in Chinese)
(张建升. 我国主要城市群环境绩效差异及其成因研究[J]. 经济体制改革, 2016(1): 57-62)

[18] Wakamiya S, Lee R, Sumiya K. Crowd-sourced cartography: measuring socio-cognitive distance for urban areas based on crowd's movement[C] //Proceedings of the 2012 ACM Conference on Ubiquitous Computing. New York: ACM Press, 2012: 935-942

[19] Byron L, Wattenberg M. Stacked graphs–geometry & aesthetics[J]. IEEE Transactions on Visualization and Computer Graphics, 2008, 14(6): 1245-1252

[20] Sun G, Wu Y, Liu S, *et al*. EvoRiver: Visual analysis of topic coopetition on social media[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12): 1753-1762

[21] Johansson J, Forsell C. Evaluation of parallel coordinates: Overview, categorization and guidelines for future research[J]. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(1): 579-588