硕 士 学 位 论 文

MASTER'S DISSERTATION

论文题目　区域大气污染物关键传播路径和重要节点挖掘方法研究

作者姓名　吴瑶

学科专业　计算机科学与技术

指导教师　黄国言教授

2018 年 5 月

中图分类号：TP393　　学校代码：10216
UDC：621.3　　密级：公开

工学硕士学位论文

区域大气污染物关键传播路径和重要节点挖掘方法研究

硕 士 研 究 生：吴瑶
导　　师：黄国言教授
申 请 学 位：工学硕士
学 科 专 业：计算机科学与技术
所 属 学 院：信息科学与工程学院
答 辩 日 期：2018 年 5 月
授 予 学 位 单 位：燕山大学

A Dissertation in Computer Science and Technology

RESEARCH ON KEY PROPAGATION PATHS AND IMPORTANT NODE MINING METHODS OF REGIONAL AIR POLLUTANTS

by Wu Yao

Supervisor: Professor Huang Guoyan

Yanshan University

May,2018

燕山大学硕士学位论文原创性声明

本人郑重声明：此处所提交的硕士学位论文《区域大气污染物关键传播路径和重要节点挖掘方法研究》，是本人在导师指导下，在燕山大学攻读硕士学位期间独立进行研究工作所取得的成果。论文中除已注明部分外不包含他人已发表或撰写过的研究成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。本声明的法律结果将完全由本人承担。

作者签字：　　日期：　　年　　月　　日

燕山大学硕士学位论文使用授权书

《区域大气污染物关键传播路径和重要节点挖掘方法研究》系本人在燕山大学攻读硕士学位期间在导师指导下完成的硕士学位论文。本论文的研究成果归燕山大学所有，本论文的研究内容不得以其它单位的名义发表。本人完全了解燕山大学关于保存、使用学位论文的规定，同意学校保留并向有关部门送交论文的复印件和电子版本，允许论文被查阅和借阅。本人授权燕山大学，可以采用影印、缩印或其它复制手段保存论文，可以公布论文的全部或部分内容。

保密□，在　　年解密后适用本授权书。

本学位论文属于

不保密☑。

(请在以上相应方框内打“√”)

作者签名：　　日期：　　年　　月　　日

导师签名：　　日期：　　年　　月　　日

摘 要

随着我国经济的快速发展，大部分地区的空气质量状况日趋严重。严重的空气污染对人民群众的身体健康和生活质量造成了严重的影响。由于我国地域辽阔，无法建立全方位覆盖的空气质量监测站网络，导致不能及时全面的反映当前的空气质量状况。因此，本文提出一种大气污染物传播网络模型，并在此基础上挖掘出污染物关键传播路径和重要节点，有效的支撑了空气污染交互影响和演变的研究。

首先，本文充分分析影响污染物扩散的因素和机理，提出了一种基于污染物传播代价的传播模型，挖掘出实时污染物传播路径，构建污染物传播网络模型，并在污染物传播网络基础上展开深入研究。

其次，污染物关键传播路径是区域内污染物频繁传播的路径，因此为了挖掘这些污染物关键传播路径，本文基于污染物传播网络模型提出一种图矩阵方法。通过矩阵存储每条边在每个子图中出现情况，筛选出符合频繁阈值的子图序列，对各子图序列评分并排序，取 Top-K 子图序列，将其中的路径片段拼接成路径序列，从而挖掘出污染物关键传播路径。

再次，重要节点是区域内极易受到污染物影响的站点，因此为了挖掘区域内的重要节点，本文提出一种大气污染物传播网络重要节点挖掘算法—StationRank 算法。该算法不仅考虑了站点之间污染物传播方向，也考虑了站点间传播的权值。采用 StationRank 算法对大气污染物传播网络中的站点进行评分排序，从而挖掘出大气污染物传播网络中的重要节点。

最后，在京津冀空气质量监测站以及气象监测站数据下进行实验，挖掘出京津冀区域周期内大气污染物关键传播路径以及重要节点，并结合实际气象特征和地理特征验证了大气污染物关键传播路径和重要节点的重要性。

关键词：大气污染物传播；复杂网络；关键路径；图矩阵；PageRank；重要节点

Abstract

With the rapid development of China's economy, the air quality in most areas is becoming more and more serious. Severe air pollution has seriously affected people's health and quality of life. Because of the vast territory in China, it is impossible to establish a comprehensive coverage of the air quality monitoring network, resulting in a timely and comprehensive reflection of the current air quality. Therefore, in this paper, a model of air pollutant transmission network is proposed, and on this basis, the key propagation path and important node of the pollutant are excavated, which effectively supports the study of the interaction and evolution of air pollution.

First, this paper fully analyzes the factors and mechanisms affecting the diffusion of pollutants, and puts forward a propagation model based on the cost of pollutant transmission, excavates the path of real-time pollutant propagation, constructs a model of the pollutant transmission network, and further research on the basis of the pollutant transmission network.

Secondly, the key propagation path of pollutants is the path of frequent contaminants in the region, so in order to excavate the key propagation paths of these pollutants, a graph matrix method is proposed based on the pollutant propagation network model. Through the matrix storage, each subgraph appears in each subgraph, the subgraph sequence which meets the frequent threshold is selected, the sequence of each subgraph sequence is scored and sorted, the Top-K subgraph sequence is taken, and the path fragments are spliced into the path sequence, thus the key propagation path of the pollutant is excavated.

Thirdly, the important node is the site which is easily affected by the pollutants in the region, so in order to excavate the important nodes in the region, this paper proposes an important node mining algorithm of atmospheric pollutant propagation network - StationRank algorithm. The algorithm considers not only the direction of pollutant transmission between stations, but also the weight of stations. The StationRank algorithm is used to sort and rank the sites in the air pollutant transmission network, so as to dig out the important nodes in the air pollutant transmission network.

Finally, under the data of the Beijing Tianjin Hebei air quality monitoring station and

the meteorological monitoring station, the key propagation paths and important nodes of the air pollutants in the Beijing Tianjin Hebei region are excavated, and the importance of the key propagation path and important nodes of the air pollutants is verified in combination with the actual meteorological features and geographical features.

Keywords: air pollutant transmission; complex network; critical path; graph matrix; PageRank; important node

目 录

第 1 章 绪　论

1.1 课题背景及研究的目的和意义

1.1.1 课题研究背景

近年来，我国快速的经济发展和日益增长的城市化使我国大部分地区出现了严重的空气污染问题。大气污染对人类健康有着严重负面影响，而且严重影响了经济和社会的可持续发展，因此防治空气污染成为中国社会的一个亟待解决的问题[1]。

空气污染就像个“隐形杀手”。每年都有大量的人死于与空气污染有关的疾病，仅在美国，由于暴露在发电厂排放的微小颗粒环境中，每年就有 7500 至 5.2 万人死亡。空气污染也是肺癌的潜在病因，有大量的证据表明，它也可能增加了人类患上心脏病和中风的概率。最新发布的 2018 年环境绩效指数显示，空气污染已经成为公众健康最大的环境威胁。

近些年来，京津冀地区空气质量问题日趋严重，曾多次出现污染物浓度严重超标的情况，严重的影响了人们的身体将康和扰乱了人们的生活规律。有研究结果表明，北京地区与周边地区的大气污染物有着频繁的互相交换和输送关系，因此，周边地区的大气污染物排放对北京地区的空气质量也会造成一定程度上的影响[2]。由此可见我国大气污染呈现出明显的区域性复合污染的特征[3]，区域性复合污染的一个现象就是地理位置相近的城市颗粒污染物具有非常明显的一致性和同步性。区域性污染物特征另一个表征现象就是单纯控制单个城市的污染物浓度对污染物排放对大气中污染物浓度控制效果欠佳，必须在区域的尺度上协同控制才能解决区域性问题。

因此，一个地区的空气污染不仅仅与该地区自身产生的污染物有关，还与其周边地区的的大气污染物输送有着紧密的联系。所以，要想有效的治理大气污染，就需要从全方位来进行研究，不可仅限于对本地区污染状况的研究。

为了及时监测空气质量的变化并迅速对重污染天气采取出有效的应对措施，在每个地区的重污染频发的地区建立空气质量监测站。但是，由于我国国土辽阔，无法在每个地区都建立完善的空前质量监测体系[4,5]，因此，就会出现无法全面的采集污染物相关信息，从而导致无法正确对空气污染防治做出正确指导。为此，合理的建立监测站对采集污染数据以及有效的防治大气污染起着至关重要的作用。

1.1.2 课题目的和意义

通过大量研究发现，一个地区的大气污染状况不仅是受当地污染物源的影响，同时很大一部分是来自周边地区的大气传输以及交换，因此大气污染具有跨区域输送的特点[6,7]。因此，本文在空气质量监测站历史数据和气象监测站历史数据的基础之上，根据污染物传播机理，构建污染物传播模型，搭建污染物传播网络，最终挖掘出污染物关键传播路径和重要节点。

京津冀地区具有北临燕山山脉，西临太行山，东临渤海以及高度密集城市群的地理特征[8]。独特的地理特征、气象条件和各个地区不同的污染情况对污染物的传输都造成了显著的影响。目前，京津冀地区一共部署了大约 200 个空气质量监测站，其中北京地区大约有 30 个，天津地区大约有 20 个，从而在京津冀地区内形成了一个比较完善的污染物检测体系，可及时有效的获取各个地区的空气质量情况，并及时的加以防治。

通过研究发现，污染物的传输具有显著的区域性，往往一个地区受到污染后，相邻的区域很快也会受到不同程度的污染，因此研究污染的传输规律，挖掘出污染物关键传输路径对污染物的防治起到至关重要的作用。因此，空气质量监测站监测的数据就具有重要的意义，但是由于京津冀地区十分辽阔、昂贵的监测站建设成本以及维护费用，导致了不能在整个地区全方位的覆盖空气质量监测站。因此为了使用尽可能少的空气质量监测站来监测范围尽可能大的区域，从而获取经可能完善的数据，研究大气污染物区域内关键传播路径以及挖掘区域内重要节点是十分有意义的。

1.2 国内外研究现状

大气污染是当今世界最严重的环境问题之一。严重的大气污染导致人们的生活质量下降和生物圈的改变，迫使各国政府和国际组织为改善大气质量投入大量的人力物力来保护生物圈。大气污染不仅仅是控制排放源的这样的简单问题。大气污染防治因其过程中极其复杂、无法控制的因素而显得异常复杂，这些因素中最重要的是气象学，因此建立气象条件与空气质量之间的联系是十分必要的，可以帮助当局制定有效的环境保护战略来改善生活质量。近些年来，各国政府开始意识到保护生态环境的重要性，为了改善人们的生活质量，加大了对环境污染的监管和防治力度。

空气质量具有动态性、可变性和复杂性的特点，有效表征空气质量特征的动态关联关系，准确分析空气质量演变趋势还没有得到很好解决。目前，国内外空气质量研究方法主要采用数值分析方法、统计分析方法、气象场分析方法。

1.2.1 数值分析方法

数值分析模型依赖于气象预报、污染源清单和扩散模型等复杂技术体系。其代表性成果是美国环保局空气质量模式(Model-3/CMAQ)[9]。2007年王扬锋等人[10]通过采用新一代空气质量模式系统Models-3，对沈阳市冬季采暖期间大气污染物传播与化学反应进行了实验，发现实验值与实际值具有相似的变化趋势。2013年薛文博等人[11]基于CAMx空气质量模型，定量模拟了全国PM2.5及其化学组分的跨区域输送规律。2013年张稳定等人[12]利用嵌套网格空气质量模式系统(NAQPMS)较为合理的模拟出污染情况和输送过程。2015年吕炜等人[13]采用CMAQ模型系统获得了珠江三角洲主要污染物浓度时空分布和大气传输季节变化特征，并分析了长距离传输对珠三角区域空气质量的影响。

1.2.2 统计分析方法

统计分析模型的性能依赖于对于空气质量影响因素的深刻理解、因子筛选和模型设计，主要包括神经网络和灰色模型等研究方法。

神经网络分析方法具有非线性和自学习等特点，适合解决复杂和难以建模的问题。2009 年赵宏等人[14]提出了一种遗传算法与神经网络算法相结合的空气质量预测方法；2012 年 Perez 等人[15]提出了结合神经网络模型与近邻模型实现 PM10 浓度等级预测方法；2013 年 Ana 等人[16]使用基于随机变量的优化神经网络方法对空气质量进行预测，在一定程度上减少神经网络模型需要的输入数据量；2015 年 Shifeng Li 等人利用遗传 BP 神经网络对城市空气质量进行预测。

由于空气质量分析受时空限制使数据获取不完备，非线性回归和灰色理论模型能够利用少样本进行建模和预测。2011年Tzu-Yi等人[17]通过灰色理论GM(1,1)模型预测局地PM2.5和PM10的浓度变化趋势；2012年Carbajal-Hrnández等人使用模糊逻辑和自回归方法来评估和预测空气质量；2013年丁卉等人[18]基于灰色聚类和模糊评判方法建立了一种空气质量评价模型，并与API实测数据进行了对比分析；2013年司志娟

等人[19]将灰色GM(1,1)与神经网络组合，建立了灰色神经网络组合模型；2015年Shen J等人[20]建立了一种基于聚类与多元回归的空气质量预报模型。

1.2.3 气象场分析方法

气象场分析模型是通过对大气运动规律进行模拟从而从中发现出一定的污染物传输的规律。

2008年Chu等人[21]介绍了兰州每个季节SO2浓度与气象条件的关系；2011年Xu等人[22]连续观察了一个地区连续七个月的空气质量，揭示了该地区的空气质量与污染物特征、区域交通状况、气象条件相关。2012年Liu等人[23]通过对北京2008年奥运会期间大气污染物的大量检测，发现了控制污染物的排放对污染趋势有着显著的效果；2011年Barmpadimos等人[24]研究了气象因素对瑞士1991～2008年间PM10趋势和变化的影响，指出风和温度是两个最重要的因素；Lalas等人研究了复杂的海陆相互作用和特定区域不同地貌对大气污染物的影响，结果表明，海风环流影响Athens地表O3的日变化；2007年Maraziotis等人[25]发现PM10和PM2.5与水平风速呈负相关关系；SS Abdalmogith等[26]采用轨迹聚类模拟了大气运动对空气污染物长距离传输的影响；Garcia Menendez F [27]以三维欧拉输送模式研究了大气流动及预测了相应的空气质量变化；S Freitag等[28]结合HYSPLIT4模拟了西班牙气溶胶背向轨迹并对其进行聚类分析发现主要传输路径。

1.3 存在的问题

虽然目前国内外有多种方法可以对污染物的传输演变进行分析，但是这些方法中还存在着一些问题。

(1) Model-3/CMAQ模型存在质量不守恒、边界参数设置困难和模拟结果存在系统性误差等问题[29]。

(2) 神经网络方法通过对大量监测数据自学习来预测空气质量的演变趋势，但由于缺乏空气质量机理分析理论的有效支持，空气质量分析的约束条件和关联关系不充分，影响了监测数据分析的质量和效率。

(3) 灰色模型方法实现了利用少样本数据对空气质量的分析和预测，但理论研究基础较薄弱，缺乏对污染物关键传播途径和区域空气交互影响的分析，空气质量的预测的准确性不高。

复杂网络是一种比规则网络具有更复杂拓特征的网络，其主要特征包括：平均距离、度分布、簇系数、度-度相关性、多维度和多层次结构等，复杂网络已经成为系统科学、复杂性科学和统计学研究的有力工具，它作为一种研究模式，为表征和分析包括生物学、医学、计算机科学和社会科学等众多领域中的复杂问题提供了新的技术方法。目前，国内外对于复杂网络在空气质量建模和分析研究中尚未得到应用。然而，我国京津冀和长江三角洲等区域，已经建设了分布广泛、多层次空气质量监测系统，形成了具有一定规模的常规监测网络，空气质量监测站采集到的数据具有规模庞大，结构复杂的特点。如何深化理论研究，提高空气质量监测数据的分析和利用质量亟待解决[30]。

1.4 论文研究内容

本文采用京津冀区域内的空气质量监测站历史数据和气象监测站历史数据，基于污染物传播机理、复杂网络和 PageRank 改进算法挖掘大气污染物在传播过程中的关键路径和重要节点，为大气污染物的区域联防提供理论支持和帮助。本文的整体研究过程如图 1-1 所示。

(1) 本文通过采用分析污染物传播机理的方法，筛选出影响大气污染物扩散的气象因素和地理因素，从而建立污染物传播代价模型表征污染物传播关系。

(2) 基于污染物传播代价模型，通过分析污染物传播约束条件来确定站点之间的路径。将监测站点视为节点，将站点之间的路径视为边，污染物传播代价视为边上的权重，从而将污染物传播过程抽象成一个有向加权污染物传播网络，并对该网络进行了度分布的特性分析，发现污染物传播网符合复杂网络特征。

(3) 通过矩阵存储每条边在每个子图中出现情况，并筛选出符合频繁阈值的子图序列，统计出每个子图序列中 1 的个数和相同子图序列下的路径片段个数，由此计算出各个子图序列的评分并排序，取 Top-K 子图序列，将其中的路径片段拼接成路径序列，从而挖掘出污染物关键传播路径。

(4) 污染物传播网络中站点所处地区越是经常出现重污染天气，说明该节点越容易受到污染物影响，其重要程度越高。为了找到污染物传播网络的这些易受污染的站点，提出了一个基于 PageRank 算法的改进算法—StationRank 算法，该算法对传统算法中的节点数目以及转移概率进行了改进，充分考虑了边的权重以及相邻节点的重要程度。对所有监测站进行评分排序，评分越高，节点越重要，越易受到污染。

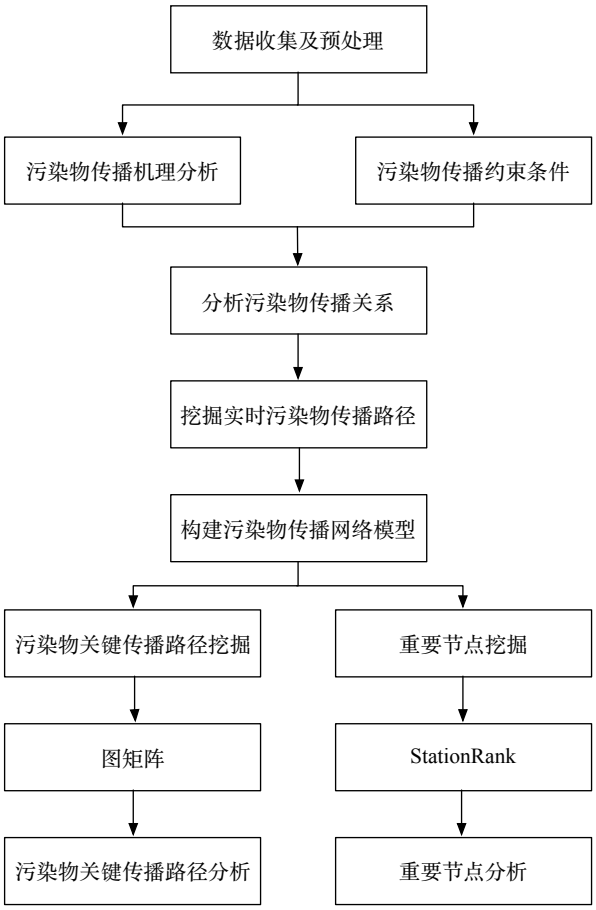

图 1-1 整体研究过程

1.5 论文的组织结构

论文共分 5 章，本章主要阐述了本文的研究背景、目的、意义和国内外研究现状。从第 2 章起，论文结构如下：

第 2 章首先介绍了空气污染指数和污染物传播机理等基本概念；然后，简要阐述了复杂网络的基本概念；最后简要介绍了网页排序算法 PageRank 的基本原理。

第 3 章首先对京津冀地区的气象数据和污染物数据进行数据预处理；然后根据污染物传播机理构建大气污染物传播模型；最后，根据污染物传播约束条件，依据构建的传播模型构建出污染物传播网络，并验证其复杂网络的特性。

第 4 章首先在污染物传播网络的基础上，挖掘出污染物传播子图。然后，在运用图矩阵的方法，存储每条路径片段的子图序列。最后，计算出各个子图序列的评分并排序，取 Top-K 子图序列，将其中的路径片段拼接成路径序列，从而挖掘出污染物关键传播路径，并对挖掘出的污染物关键传播路径进行了实验分析。

第 5 章首先对京津冀地区目前空气质量的现状进行了分析并阐述了网络节点重要性的评估方法；然后，在论述了传统 PageRank 算法应用于空气污染物传播有向加权网络的不适应性，并在此基础上提出 StationRank 算法；最后，详细分析了在京津冀空气污染物传播网络基础上挖掘关键节点的实验过程与结果。

最后总结了本文研究成果，并对本文研究工作做出了总结和展望。

第 2 章 相关理论和基础知识

2.1 基本概念

2.1.1 污染物传播机理

空气污染物传播是一个十分复杂的过程，在这个传播过程中会受到多方面因素的影响[31,32]。其中主要的三个方面是地形因素、气象因素以及污染物因素。地形因素包括：两个地区之间的水平距离、两个地区之间是否有山脉、两个地区之间的海拔差。气象因素包括：风速、风向、温度、湿度、降雨、雾、沙尘等。污染物因素包括：两地之间的浓度差。

水平距离：由于污染物传播具有远距离传输特性，所以在一定范围的地区，常常会与受污染的区域有着同步增长和减少的趋势，因此这些地区的相关性就会很高，这样就会导致当一个地区受到严重的污染时，与其相邻的其他地区的污染物浓度也会同步增加。

山脉：山脉对污染物的传播既有抑制作用，也有促进作用。当山脉位于两地之间时，这时山脉就会对污染物的传播起到一个阻挡和削弱的作用。当两个地区位于山脉的同侧时，污染物会沿着山脉的走势进行传播，此时山脉对污染物的传播就会起到促进的作用。

海拔高度：当两个地区的海拔高度相差不大，并且之间不存在山脉时，此时污染物的传播主要受气象因素的影响。如果两地之间的海拔高度差十分大时，地势低的地区对地势高的地区的影响很小，污染物就会产生汇聚现象，导致该地区的污染物浓度很高。

风向、风速[33,34]：风是影响一个地区大气污染物运动的重要气象因子。风向决定了污染物的传播方向，风向决定了污染物的传播速度、传播的范围的大小以及局地地区污染物浓度的大小。当风速较大时，会使局地地区的空气变得稀薄，污染物的沉降效果降低，因此局地地区的空气质量会很好，但是这样污染物影响的范围会扩大。当风速很低时，不仅使污染物很难得到扩散，而且对污染物的聚集起到了促进作用，很容易会造成局部严重污染。

温度：逆温是指低空区域温度比高空区域温度低形成比较稳定的形势，同时高空的暖温区与低空辐射降温相接触形成逆温层，不利于污染物的对流和垂直传播。

湿度：当湿度过高时，大气污染物会吸收空气中的水分而使自身重量增加，从而导致了大气污染物难以扩散，使得空气中的污染物浓度增加，污染状况加重。只有当降雨量较大时，一般是中雨、大雨的天气状况，才会使得污染物沉降下来，从而使得空气质量得到改善，降低空气污染。

降水：雨水可以稀释大气污染物，对大气污染物的有着净化的作用，但降雨的净化作用与降雨量和降雨持续时间相关，雨水量大，持续时间长，净化作用就越明显，反之，则没有净化作用，甚至会污染更加严重。

沙尘：沙尘天气会使空气污染加剧，强沙尘暴会在极短的时间内使空气中的污染物急剧增加，造成严重的颗粒物污染。空气中高浓度的悬浮颗粒物，在静风、逆温等特殊的气象条件相互作用下，易导致雾霾天气的发生，具体表现为区域性或大范围内的空气质量恶化。

污染物浓度差：一个地区的污染物浓度十分高时，它对周围地区污染物浓度有着很大的影响，根据气体扩散原理，污染物浓度高的地区的污染物会向污染物浓度低的地方进行扩散，到这整个区域受到污染，另外受到风等气象因素的影响，会导致整个区域污染程度的加剧，因此，两个地点之间的污染物浓度差也是影响空气污染物扩散的一个重要因素。

2.1.2 空气质量指数

AQI 即空气质量指数(Air Quality Index)，用来定量描述空气质量状况的无量纲指数。由于 AQI 评价的 6 种污染物浓度限值各有不同，在评价时各污染物都会根据不同的目标浓度限值折算成空气质量分指数 AQI。AQI 范围从 0 到 500，大于 100 的污染物为超标污染物。AQI 分优(绿色)、良(黄色)、轻度污染(橙色)、中度污染(红色)、重度污染(紫色)、严重污染(褐红色)6 种评价类别和表征颜色。当 AQI 类别为优或良时，一般人群都可以正常活动；为轻度污染以上，各类人群就需要关注建议采取的措施。

空气质量分指数计算公式如下：

$$IAQI_p=\frac{IAQI_{hi}-IAQI_{lo}}{BP_{hi}-BP_{lo}}\left(C_p-BP_{lo}\right)+IAQI_{lo} \tag{2-1}$$

式中 $IAQI_p$——污染物p的空气质量分指数；

C_p——污染物p的质量浓度值

BP_{hi}——与C_p相近的污染物浓度限值的高位值；

BP_{lo}——与C_p相近的污染物浓度限值的低位值；

$IAQI_{hi}$——与BP_{hi}对应的空气质量分指数；

$IAQI_{lo}$——与BP_{lo}对应的空气质量分指数；

空气质量指数计算方法如式(2-2)：

$$AQI=max\{IAQI_1,IAQI_2,\cdots,IAQI_n\} \tag{2-2}$$

式中 $IAQI$——空气质量分指数；

n——污染物。

当AQI超过 50 时，$IAQI$最大的污染物为重要污染物。若$IAQI$最大的污染物为两种或两种以上时，并列为重要污染物。

表 2-1 空气质量分指数及对应的污染物项目浓度限值

空气质量分指数(IAQI)	SO_2 (日均值)	SO_2 (小时均值)	NO_2 (日均值)	NO_2 (小时均值)	PM10 (日均值)	CO (日均值)	CO (小时均值)	O_3 (日均值)	O_3 (小时均值)	PM2.5(日均值)
0	0	0	0	0	0	0	0	0	0	0
50	50	150	40	100	50	2	5	160	100	35
100	150	500	80	200	150	4	10	200	160	75
150	475	650	180	700	250	14	35	300	215	115
200	800	800	280	1200	350	24	60	400	265	150
300	1600	(1)	565	2340	420	36	90	800	800	250
400	2100	(1)	750	3090	500	48	120	1000	(2)	350
500	2620	(1)	940	3840	600	60	150	1200	(2)	500

2.2 复杂网络特性

如果存在一个网络，它具有自组织、自相似、吸引子、小世界、无标度中部分或者全部特性，则称该网络为复杂网络(Complex Network)。随着网络的发展，复杂网络从始至终没有离开过国内外学者研究的视线。最早的网络研究可以追溯到对图

的研究，即数学家欧拉向圣彼得堡科学院递交的《哥尼斯堡的七座桥》论文，其开创了图论与几何拓扑学。随后的两百年间，各国的学者都致力于对简单的规则网络和随机网络进行抽像的数学研究。规则网络由于过于理想化并不能反映系统的复杂性，在 20 世纪 60 年代 Erdos 和 Renyi 提出了随机网络。进入到 20 世纪 90 年代，人们发现现实世界中的绝大多数网络既符合规则网络和随机网络的部分特征，同时还具有自己的不同于这两者的明显特征，于是提出了一些更符合现实的网络。与此同时，复杂网络正进入人们的视线，学者们争相加入到研究复杂网络的大军中。在最初的工作中，有两项比较重要的成果。一是复杂网络的小世界模型[35]的提出，Wats 和 Strogate 在享誉全球的 Nature 杂志上发表了重要文章，正式提出了此模型，他们指出该模型不仅具有类似随机网络的小的平均路径长度，同时具有规则网络的高聚类特性。二是 Barabasi 和 Albert 在 Science 杂志上发表文章，提出了无标度网络模型[36]。他们认为现实生活中的大多数复杂系统都是动态演化的，是开放自组织的，实际网络中的无标度现象来自于增长机制和优先连接机制这两个重要因素。复杂网络具有高度的复杂性，其特点的具体表现如下：

(1) 小世界(Scale-free)特性又称为六度分割理论，通俗来讲就是网络中任何一个成员和另一个成员之间间隔都不会超过六个人。复杂网络中小世界特性对网络中信息传播有着重要的作用。在这样的网络中，信息传播十分迅速，如果在网络中去除或者修改几个边时，就会导致整个网络性能的巨大变化。

(2) 无标度(Small-world)特征主要表现在网络中少数节点存在大量连接，而大量节点存在少数连接。这些节点的度数分布呈现幂律分布的特性。复杂网络的无标度性与网络的鲁棒性有着紧密的联系。无标度网络中的幂律分布的特性很大程度上提升了高度数节点存在的可能性，正是因为这一点，也极大的降低了复杂网络的鲁棒性，也就是说，如果恶意攻击者攻击网络中度数很高的一部分节点时，很可能会造成网络迅速的瘫痪。

很多研究成果发现现实世界中很多网络都可以进行复杂网络抽象化，应用复杂网络理论研究其网络性质，这也为更复杂的网络关系研究提供了新思路。因此，将现实网络抽象为具有某些特殊拓扑特性的复杂网络，使人们对此网络进行理论性研究，应用复杂网络的性质对其进行定性和定量分析。

图作为复杂网络的表示方法，是研究网络的重要工具。我们通常使用方式描述网络结构。表示网络中所有节点的集合，表示网络中所有边的集合，在网络中如果

边不存在固定的指向，则使用节点对与代表同一条边。称之为无向图。反之则是有向图，表示以节点为起始以节点为目标节点由指向的边，而则代表从节点指向节点的边。如果网络中每条边都存在着与其相对的权重，则我们称之为加权图。

通常可以使用邻接矩阵或邻接表的结构储存整个网络：

(1) 邻接矩阵：邻接矩阵是表示顶点之间关系的矩阵，用二维数组保存图中信息，数组下标表示图中顶点编号，用二维数组中对应位置的值来表示图中边上信息。图的邻接矩阵是唯一的，存储图使用邻接矩阵计算更简单方便。邻接矩阵不仅能存储无向图（对称阵），而且也能存储有向图（行号为初始点，列号为终端点）。邻接矩阵可以快速判断图中两个顶点之间是否有边，以及快速获取每个顶点的出入度及其邻接点等优点。例如给定两个顶点编号，能快速的在邻接矩阵中检索对应位置的值。但是，邻接矩阵也有着明显的缺点，当图中边很少，顶点很多的时候，矩阵会因为链接边的分布而变得极其稀疏，二维数组中存储的有效值很少，这样会造成极大的内存浪费。

(2) 邻接表：邻接表又被称为邻接链表，是数组和链表组合的存储方式。图中顶点用一维数组存储信息，每个顶点的所有邻接点组成一个线性表，由于邻接点数量不能确定，所以用单链表存储，再把单链表链接到该数组元素上。因此，只要这个顶点有一条单链表，那么该顶点的度不为 0。邻接表中只存储图中存在的边信息，与邻接矩阵相比，邻接表可以节省很多存储空间。

2.3 PageRank 算法介绍

PageRank(网页排序)又称网页级别、Google 左侧排名或佩奇排名[37]。由 Stanford University 的 SergeyBrin 和 Lawrence Page 工程师发明。其理论是把节点间链接权重作为网页排名的因素。每个网页都会有一个 PageRank 值来表示网页的等级，其级别从 1 到 10，网页的搜索结果会按照等级由高到低显示，即 PageRank 值越高表示网页越受欢迎。在网络中，如下图每个球代表一个站点，球的大小反应了站点的 PageRank 值的大小。指向站点 B 和站点 E 的链接很多，所以 B 和 E 的 PageRank 值较高，另外，虽然很少有站点指向 C，但是最重要的站点 B 指向了 C，所以 C 的 PageRank 值比 E 还要大。通过计算各节点的 PageRank 值，PageRank 值越高的节点越重要，最后对节点按照 PageRank 值由大到小排序，按需取排序靠前的几个节点为重要节点。

因此，网页的 PageRank 值由三个因素决定：一是网页链入当前网页的数量；二是链入网页本身的重要度，表示了链接源是否为质量高的网页；三是链入网页本身的链出数量，即该链入网页给其它网页的投票数目。

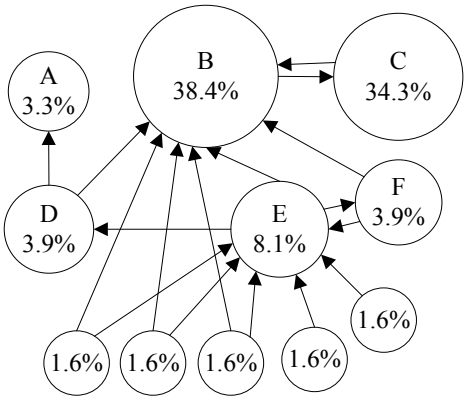

图 2-1 PageRank 示意图

设p_i表示一个页面，$F(p_i)$为页面p_i指向的页面集合，$M(p_i)$为指向页面p_i的页面的集合，$L(p_i)=|F(p_i)|$为页面p_i指向的页面数，N为页面总数，α为阻尼系数。PageRank 值用公式表示为：

$$\text{PageRank}(p_i)=\frac{1-\alpha}{N}+\alpha\sum_{p_j\in M(p_j)}\frac{\text{PageRank}(p_j)}{L(p_j)} \tag{2-10}$$

α为向后搜索其它页面的概率，1-α为用户随机浏览到新 URL 的概率。根据公式可以看出每个页面的 PageRank 值即为多个连接它的页面的投票值。PageRank 算法步骤如下：

(1) 最初阶段：首先建立站点节点间网络关系图，其次初始化每个站点一样的 PageRank 值，通过在每次迭代过程中，不断更新各个站点的 PageRank 值，当前后两次迭代的 PagePank 差值达到一定阈值，迭代终止，得到每个站点的最终 PageRank 值。

(2) 迭代计算站点 PageRank 得分方法：每轮迭代中站点的 PageRank 值会赋值到本站点的出边上，使每条边都有相同的值。站点对链入自己的边权值求和，即获得新的 PageRank 得分。一次次迭代直到计算完成。

2.4 本章小节

本章主要介绍了本文用到的相关理论和基础知识。首先介绍了空气质量指数的

基本概念以及计算公式、影响污染物传播的相关因素。其次，介绍了复杂网络的相关特性。最后，对 PageRank 算法做了简单的介绍。

第 3 章 基于复杂网络构建大气污染物传播网络模型

3.1 引言

目前的大气污染物扩散模型大部分是基于详细的污染源清单以及污染物之间的相互转化所构建的，但是由于获取所有详尽数据的难度很大，因此探究大气污染物传输规律的难度也很大[38]。然而，我国京津冀和长江三角洲等区域已经建设了分布广泛、多层次空气质量监测系统，形成了具有一定规模的常规监测网络，空气质量监测站采集到的数据具有规模庞大，结构复杂的特点。因此本章将以空气质量监测站和气象监测站历史数据为基础，依据大气污染物的传播机理来构建污染物传播模型，构建大气污染物传播网络模型，并探究该网络模型具有的特性，从而深入的利用大气污染物传播网络模型来挖掘污染物的传输规律。

3.2 数据预处理

3.2.1 传播距离计算方法

地球是一个近乎标准的椭球体，平均半径为 6356.755km。假设监测站 A 的坐标为(*lonA,latA*)，监测站 B 的坐标为(*lonB,latB*)。并根据三角定理，可以得到计算两监测站距离的公式如下：

$$C=sin(lat_A)*sin(lat_B)*cos(lon_A\text{-}lon_B)+cos(lat_A)*\cos(lat_B) \tag{3-1}$$

$$\Delta Dis_{AB}=R*arccos(C)*pi/180 \tag{3-2}$$

其中，*R* 为平均半径 6356.755km，*pi* 为 3.1415。

3.2.2 风向夹角计算方法

根据向量乘积的原理，假设存在向量$\vec{a}$和$\vec{b}$，因此存在如下等式：

$$\vec{a}\cdot\vec{b}=x_a*x_b+y_a*y_b=|\vec{a}||\vec{b}|cos(\theta) \tag{3-3}$$

其中x_a和y_a代表向量$\vec{a}$的横纵坐标，y_a和y_b代表向量$\vec{b}$的横纵坐标，$|\vec{a}|$和$|\vec{b}|$表示向量$\vec{a}$和$\vec{b}$的模，θ为两个向量之间的夹角。

假设风向的单位向量$\vec{F}$为$(cos\,\omega, sin\,\omega)$，从监测站 A 到监测站 B 方向的向量为$\left(Lon_B\text{-}Lon_A, Lat_B\text{-}Lat_A\right)$，如图 3-1 所示。

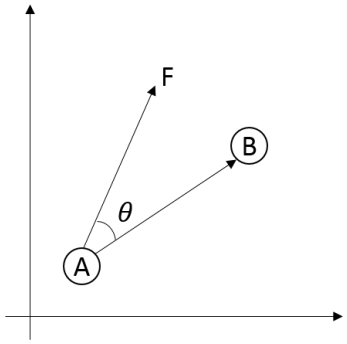

图 3-1 风向夹角示意图

因此可将上述向量乘积的原理应用到计算两个监测站方向与风向之间夹角的计算中，公式如下：

$$(cos\,\omega\,,sin\,\omega)\cdot(Lon_B\text{-}Lon_A\,,\,Lat_B\text{-}Lat_A)=\left|\vec{F}\right|\left|\overrightarrow{AB}\right|\,cos(\theta) \tag{3-4}$$

于是可以求出夹角θ为：

$$\theta=arccos\left(\frac{(cos\,\omega\,,sin\,\omega)\cdot(Lon_B\text{-}Lon_A\,,\,Lat_B\text{-}Lat_A)}{|\Delta Dis_{AB}|}\right) \tag{3-5}$$

其中ω为风向的弧度，$|\Delta Dis_{AB}|$为向量$\overrightarrow{AB}$的模。

3.2.3 污染物相关性计算方法

假设监测站 A 在 t 时刻的污染物浓度为$X_A(t)$，监测站 B 在 t 时刻的污染物浓度为$X_B(t)$，监测站 A 上的污染物浓度均值为$\overline{X_A}$，监测站 B 上污染物浓度均值为$\overline{X_B}$。因此，A、B 监测站之间污染物的相关性为可由公式(3-6)求得：

$$\rho_{AB}=\frac{\sum(X_A(t)\text{-}\overline{X_A})(X_B(t)\text{-}\overline{X_B})}{\left(\sqrt{\sum_{i=1}^{n}(X_A(t)\text{-}\overline{X_A})^2}\right)\left(\sqrt{\sum_{i=1}^{n}(X_B(t)\text{-}\overline{X_B})^2}\right)} \tag{3-6}$$

3.2.4 数据标准化方法

本文中涉及到的气象数据、污染物数据、地理数据的单位是各不相同，由于需要对这些数据进行交叉计算，因此需要对每个因素进行无量纲转换—数据标准化。

本文采用min-max标准化方法，对原始数据进行线性转换。公式如下：

$$x^*=\frac{x\text{-}min}{max\text{-}min} \tag{3-7}$$

其中max为各因素样本数据的最大值，min为各因素样本数据的最小值。x^*为各

因素数据标准化后的值，x为各因素原始数据。

3.3 污染物传播关系分析

空气污染物传播起点的气象因素，如：温度(T)、湿度(H)、风向(WD)、风速(WS)等都会影响空气污染物的向外扩散，并且空气污染物的传播难易程度不仅受本地气象因素的影响，也会受到传播目的地的影响。其中，两地之间的距离、起始区域的湿度、两地之间的海拔高度差和两地之间的山脉高度与污染物传输呈负相关系，两地之间的污染物相关性、风力系数差值、起始区域的温度以及两地之间的污染物浓度差值呈正相关关系。因此，结合第 2.1.1 节提出的大气污染物传播机理，t 时刻大气污染物从站点 A 传输至站点 B 的所需代价$Cost_{AB}(t)$为：

$$Cost_{AB}(t)=\alpha\frac{|\Delta Dis_{AB}|*Hum_A(t)*H*\Delta H}{C_{AB}*\Delta V_{AB}(t)*Tum_A(t)}-\beta*\Delta AQI_{AB}(t)+\varepsilon(t) \tag{3-8}$$

如式(3-8)中$Hum_A(t)$表示起始点 A 在 t 时刻的湿度；H表示两地之间山脉的高度；ΔH表示两地之间的海拔高度差；$Tum_B(t)$表示起始点 B 在 t 时刻的温度；$|\Delta Dis_{AB}|$表示站点 A 与站点 B 之间的水平距离；$\Delta AQI_{AB}(t)$表示站点 A 与站点 B 之间t时刻的AQI差值；C_{AB}表示站点 A 与站点 B 之间的相关性；$\Delta V_{AB}(t)$表示站点 A 与站点 B 之间 t 时刻的风力系数差值；α,β表示校正系数，$\alpha+\beta=1$，取值范围均为$[0,1]$；$\varepsilon(t)$表示传播代价$Cost_{AB}(t)$的波动值。

通过污染物在站点 A 与站点 B 之间的传播代价$Cost_{AB}(t)$可以衡量污染物传播的难易程度，污染物传播代价越大，代表污染物越不易传播，反之，越容易传播。该公式不仅考虑了污染物起始地的气象因素，同时也考虑到了污染物目的地的气象因素，以及两地之间的地形因素。

3.4 污染物实时传播路径的获取

由于地形、地貌等一些自然条件的限制，和气象因素的影响，导致在一些地区出现一些污染物的传播通道，例如：以京津冀为例，其西临太行山、北临燕山、东临渤海，导致在偏西南风气流的影响下，污染物会沿着太行山系中的洋河河谷和燕山山脉向京津冀地区传播，导致大面积、区域性的污染。因此研究污染物的关键传播路径是十分重要的，不仅可以有效的对污染进行区域联防，而且可以及时阻断污

染物的传播。并且由于京津冀地区十分广阔，不能实现监测站的全方位的覆盖，因此寻找污染物的关键路径，并合理的在关键路径上设置空气质量监测站，不仅减少了监测站的建设成本，也对整个地区起到了有效的防控作用。

3.4.1 传播约束条件分析

水平距离(D)：根据数据集 station.csv 中的经纬度以及公式(3-2)计算出站点之间的距离。如果监测站A和监测站B的实际水平距离小于*minDis*或者大于*maxDis*，则断开监测站A和监测站B之间的路径。因为当距离小于*minDis*时，基本上两个站点所有状态以及数据都是相同以及同步的，导致在一个局地地区存在大量的路径，使网络中存在大量的环，这样对研究污染物的区域性传播十分不利。从数据可知，大多数县级区域有且至少有一个监测站点，并且距离都在*maxDis*以内，又因为空气污染物的传播具有远距离传输特性，并且传输距离远大于*maxDis*，因此如果两监测站距离如果大于*maxDis*，就断开连接。

风力系数差($\Delta F_{AB}(t)$)：设站点 A 到站点 B 的方向为$\overrightarrow{\mathrm{D_{AB}}}$，站点 A 和站点 B 在 t 时刻时的风速风向分别为$\overrightarrow{F_A(t)}$、$\overrightarrow{F_B(t)}$，则$\overrightarrow{\mathrm{D_{AB}}}$分别与站点 A 和站点 B 风向的夹角为$\theta_A$,$\theta_B$，根据三角形勾股定理可知，风速$\overrightarrow{F_A(t)}$、$\overrightarrow{F_B(t)}$在$\overrightarrow{\mathrm{D_{AB}}}$方向上的分量分别为$V_A(t)$,$V_B(t)$，如图 3-2 所示。

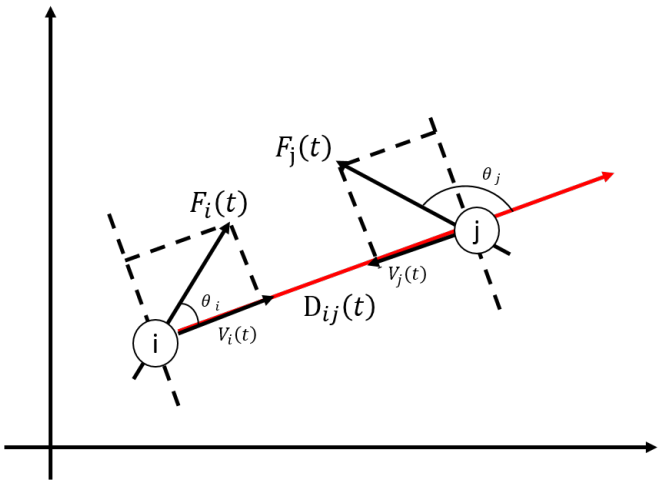

图 3-2 监测站上风力系数示意图

因此可以求出站点 A 与站点 B 之间 t 时刻的风力系数差值$\Delta F_{ij}(t)$为：

$$V_A(t)=\begin{cases}0, & \pi/2\le\theta<\pi\\ \left|\overrightarrow{F_A(t)}\right| * cos\,\theta, & 0\le\theta<\pi/2\end{cases} \tag{3-9}$$

$$V_B(t)=\begin{cases}0, & \pi/2\le\theta<\pi\\ \left|\overrightarrow{F_B(t)}\right| * cos\,\theta, & 0\le\theta<\pi/2\end{cases} \tag{3-10}$$

$$\Delta V_{ij}(t)=V_i(t)+V_j(t) \tag{3-11}$$

$$\Delta F_{AB}(t)=\begin{cases}0, & \Delta V_{AB}(t)\le 0 \;||\; \pi/2\le\theta_A<\pi\\ \Delta V_{AB}(t), & \Delta V_{AB}(t)>0\end{cases} \tag{3-12}$$

其中θ可由公式(3-5)计算得出。

如式(3-9)，当站点 A 处的风向与 A、B 站点方向$\overrightarrow{D_{AB}}$的夹角超过90°时，则认为此时站点 A 处的空气污染物不会对站点 B 处的空气质量造成影响，或者当$\Delta V_{AB}(t)$小于等于 0 时，表示站点 A 在$\overrightarrow{D_{AB}}$方向的风力系数小于等于站点 B 在$\overrightarrow{D_{BA}}$方向上的风力系数，即站点 A 处的空气污染物是无法对站点 B 处造成影响的。

山脉(hill)：当两个监测站之间存在山脉时，则认为污染物是无法在这两个地区之间进行传播的。

海拔差(ΔH)：当海拔差超过一定数值时，则认为污染物无法从低海拔的地区传播到高海拔的地区。

相关性(ρ)：当两个在适当范围内的站点上的污染物浓度出现同增或者同减的现象时，往往是因为一个站点上的污染物浓度对另一个站点产生了影响。因此，根据公式(3-6)对网络中的任意两个站点进行污染物浓度的相关性计算。当$\rho\ge 0.5$时表示站点之间有紧密依赖关系，也就是说一个站点的污染物浓度很容易就会影响到另一个站点，也就说明了这两个站点之间存在一条污染物传播通道。反之不存在传播通道。

3.4.2 污染物实时传播路径挖掘算法

当污染物从一个地区向相邻地区进行扩散传播时，往往不只是一条路径进行传播，而是存在多条不同的传播路径。由于地形因素以及气象素的影响，会导致其中一些路径对污染物的传播有极小的抑制作用，这样就使得污染物在这条路径上很容易被传播，而有些路径则对污染物的传播有极大的抑制作用，导致污染物很难在这条路径传播，本文称将这种难易程度为传播代价(*Cost*)。因此，从传播原理可知，传播代价越小的路径越是容易形成污染物的传播通道，所以，将传播代价最小的路径定义为污染物关键传播路径。

根据污染物传播的约束条件以及关键传播路径的定义，可总结出站点 A 与站点 B 存在边的条件，如下式：

$$a_{AB}(t)=\begin{cases}0, & \text{other}\\ 1, & 10\text{km}<D_{AB}\le 50\text{km and } \Delta F_{AB}(t)>0\\ & \text{and } \rho>0.5 \text{ and } Cost_{AB}(t)<0.01\end{cases} \tag{3-13}$$

其中，D_{AB}表示监测点 A 和监测点 B 之间的距离，$\Delta F_{AB}(t)$表示监测点 A 和监测点 B 在 t 时刻的风力系数差值，ρ表示监测点 A 和监测点 B 之间污染物的相关系数，$Cost_{AB}(t)$代表监测点 A 和 B 在 t 时刻的污染物传播代价。当$a_{AB}(t)$为 0 时，表示站点 A、B 之间在 t 时刻不存在一条污染物关键传播路径，为 1 时表示存在一条污染物关键传播路径。

根据污染物传播模型以及污染物传播代价即可计算出 t 时刻的从任意一点出发的污染物关键传播路径。如算法 3.1 所示：

算法 3.1 Search Critical Path(SCP)

输入： Set $AQ(t)=\{AQ_1(t), AQ_2(t),\cdots, AQ_n(t)\}$,

Set $M(t)=\{M_1(t), M_2(t),\cdots, M_n(t)\}$,

Set $S=\{S_1, S_2,\cdots, S_n\}$, $S_m\in S$

输出：$S_{min}\in S$

执行过程：

BEGIN

1. **Initialize** zero *List*, *nodeList*, $S_{min}\in S$
2. **for**(each<$S_1, S_2,\cdots, S_n$>$\in S$)
3. D=calDistance(S_i, S_m);
4. **if**($10\le D\le 50$)
5. $angle_i(t)$ = switchWindVecoter($M_i(t).wd$);
6. $includAngle_{im}(t)$=calWindAngle($M_i(t).wd, M_m(t).wd$)
7. **if**$\left(0\le includAngle_{im}(t)\le 90\right)$
8. $cost_{im}(t)$=calCost$\left(AQ_i(t), M_i(t), AQ_m(t), M_m(t)\right)$;
9. **if**($cost_{im}\le 0.01$)
10. add S_i into *nodeList*
11. **end if**

12. **end if**
13. **end if**
14. **end for**
15. S_{min}=*nodeList*[0]
16. **for**(each $node_i$ in *nodeList*)
17. **if**($S_{min} \geq node_i$) $S_{min}=node_i$
18. **end for**

END

如算法 3.1 所示，首先任意选择一个污染物传播的起始点S_k，方法calDistance根据公式(3-2)计算起始点S_k与每个监测站的距离D。并选择出距离在 10km 到 50km 范围内的站点。将这些站点的集合记作$DisList$，然后接下来的计算将基于集合$DisList$中的站点进行。由于数据集中风向不是用角度进行表示的，因此需要通过方法switchWindVecoter进行角度转换。通过方法calWindAngle计算时刻 t 时$DisList$中每个站点与S_k的风向夹角$includAngle_{im}(t)$，并筛选出夹角在 0 度到 90 度的站点，并将这些站点的集合记作 AngleList，然后在 AngleList 的基础上通过方法calCost计算每个站点与S_k的在 t 时刻时的传播代价$cost_{im}(t)$，并将符合传播代价小于 0.01 的站点的集合记作 $CostList$，最后在 $CostList$ 中找到传播代价最小的站点S_{min}，因此路径$p_{m,min}$即为在 t 时刻以S_m为起点的一条最容易传播的路径。

3.5 构建大气污染物传播复杂网络

根据第 3.2 节提出的大气污染物传播代价公式以及大气污染物传播模型，计算出每个时刻(小时)的大气污染物传播路径，由于本文是以年为周期进行的统计分析，因此在一个周期中会产生大量的污染物传播路径，然后将每个时刻的路径存储到一个矩阵当中，并将该位置的值置为 1，若在另一个时刻该条路径再次出现时，则在之前的基础上进行加 1，以此类推，最终会得到大气污染物在一个周期内传播的邻接矩阵，其中的矩阵中的值表示了该条路径在一个周期内出现的次数。

在大气污染物传播网络中，节点代表空气污染物监测站点，边代表污染物从一个区域传播到另一个区域的路径，当在一个周期(年)中，大气污染物多次从一个地区传播到另一个地区时，就表明了这两个地区之间很可能存在一条有利于污染物传播的通道。因此，考虑节点之间的依赖关系主要是该路径在周期中出现的次数，即边

上的权重代表可以用路径在一个周期中出现的次数来表示。

因此，大气污染物传播网络图可表示为三元组$G=\{V, A, W\}$，其中，$V=\{v_1,v_2,\cdots,v_n\}$表示网络中点的集合，$A=\{a_{ij};i,j=1,2,\cdots,n\}$表示网络中边的集合，$W=\{w_{ij};i,j=1,2,\cdots,n\}$表示网络中边上的权重的集合。图 3-3 为大气污染传播网络示意图。其中计算权重的算法具体过程如算法 3.2 所示。

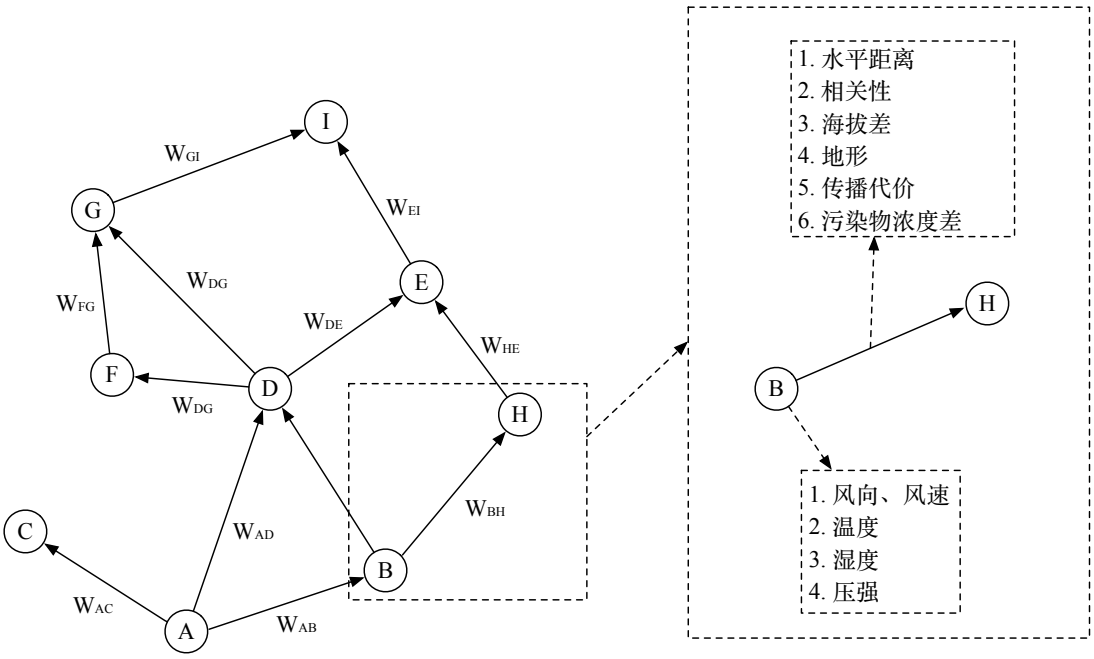

图 3-3 污染物传播网络示意图

算法 3.2 统计周期内大气污染物传播网络边上的权值

输入：站点的集合 *stationList* 和周期内传播路径集合 *pathList*

输出：边出现次数集合 *martixNum*

BEGIN

1. **Initialize** *matrixNum* //用于存储每条边的出现次数
2. **for** (each *stationStart* in *stationList*)
3. **Initialize** *map*
4. **for**(each *stationEnd* in *stationList*)
5. *map*.put(*stationEnd* , 0) //存储终点，并初始化为 0
6. **end for**
7. *matrixNum*.put(*stationStart* , *map*) //存储起点和终点
8. **end for**
9. **for** (each *path* in *pathList*)

10. **for** (each *p* in *path*)
11. *matrixNum*.get(*path*).put(*p*, *matrixNum*.get(*path*).get(*p*)+1)
12. **end for**
13. **end for**

END

如算法 3.2 所示，首先对京津冀内所有的监测站集合进行遍历，并根据这些监测站初始化矩阵 *martixNum*，矩阵中的行表示大气污染物传播的起点，列表示大气污染物传播的终点，矩阵的初始值均设置为 0。然后对周期内大气污染物传播路径集合进行遍历，并将每条大气污染物传播路径中的路径片段存储到 *martixNum* 矩阵中，并在之前的数值基础上进行累加。最终会得到一个周期内的大气污染物传播网络的邻接矩阵。

3.6 算法实例

下面将通过一个实例展示出算法 3.1 的过程。该实例选取井陉县气象局 2014 年 4 月 26 日 0 时的气象数据以及污染物浓度数据来寻找以井陉县气象局为起点的污染物关键传播路径。井陉县气象局的经纬度为(114.152, 38.037)。根据calDistance方法计算出满足与井陉县气象局距离在 10km 到 50km 范围内的监测站，如表 3-1 所示。

表 3-1 距离井陉县气象局10km~50km的监测站

监测站	监测站	监测站	监测站	监测站
市区世纪公园	市区西南高教	市区化工学校	市区职工医院	市区人民会堂
市区西北水源	市区高新区	市区封龙山	鹿泉一中	正定联通公司
栾城通讯公司	元氏住建局	赞皇县政府	平山冶河	灵寿供水

通过方法calWindAngle计算出表 3-1 中监测站与井陉县气象局风向的夹角，如表 3-2 所示。

表 3-2 各监测站与井陉县气象局风向夹角

角度(°)	角度(°)	角度(°)	角度(°)	角度(°)
63	135	81	150	165
45	143	35	132	145
87	56	156	180	90

由于当夹角为钝角时，视为污染物不能传播到下一个污染物监测站，因此去除夹角为钝角的点，则剩下的监测站为污染物可以影响到的最大范围，其中符合要求的监测站如表 3-3 所示。

表 3-3 符合夹角为锐角的监测站

监测站	夹角(°)
市区世纪公园	63
市区西北水源	45
栾城通讯公司	87
元氏住建局	56
市区化工学校	81
市区封龙山	35

最后通过方法calCost计算出监测站井陉县气象局到表 3-3 各点的传播代价，计算结果如表 3-4 所示。

表 3-4 井陉县气象局到各个监测站的传播代价

监测站	夹角(°)	传播代价
市区世纪公园	63	0.002
市区西北水源	45	0.0015
栾城通讯公司	87	0.04
元氏住建局	56	0.025
市区化工学校	81	0.035
市区封龙山	35	0.015

根据表 3-4 的传播代价就可找出污染物的关键传播路径，即传播代价最小的站点即为下一个传播起点。其中由井陉县气象局为起点的关键传播路径为：

井陉县气象局 → 市区西北水源

由表 3-4 中的风向夹角和传播代价可以看出，并不是所有的风向夹角的越小的监测站上的传播代价越小。这是因为风向只是决定了污染物的主要的传播风向，而无法反映出污染物对下一个站点的影响程度，其中地形因素、气象因素，都很大程度上决定了污染物是否能影响到其他站点。因此就需要通过污染物传播代价来衡量污染物的传播难易程度，并以此来找出实时污染物传播路径。

3.7 实验结果与分析

本文的最终目的是通过实时污染物传播路径来构建污染物传播网络。通过第 2.1.2 节中污染物传播机理的详细分析，找出了影响污染物传播的关键因素，并基于污染物传播机理建立了污染物传播模型。第 3.3 节提出了污染物传播的约束条件和获取实时污染物传播路径的详细步骤以及方法，接下来将对构建出的大气污染物传播网络进行分析。

3.7.1 实验环境配置

实验的系统配置为 macOS 10.13.2 操作系统，硬件配置为 Intel Core i5，主频为 3.10GHz，内存为 16 GB 2133 MHz LPDDR3。算法采用 Java，Matlab 语言编写。数据展示采用 Echarts、Excel，运行环境为 IntelliJ IDEA，JDK 版本为 1.7，采用的数据库为 MySQL。

3.7.2 实验数据

本文所需的数据集来源于微软亚洲研究院郑宇博士所提供京津冀地区 2014 年 5 月 1 日到 2015 年 4 月 30 日的数据集，数据集中包含 airquality.csv、city.csv、district.csv、meteorology.csv、station.csv 五个文件。其中数据集详细解析如下：

表 3-5 station.csv

站点编号	名称	纬度(°)	经度(°)	区域编号
1001	海淀北部新区	40.090679	116.173553	101
6010	天津前进道	39.092699	117.201676	607
14004	建设大厦	39.942	119.537	1401

station.csv文件中展示了京津冀范围内所有空气质量监测站的详细信息：站点编号、站点名称、站点所在位置的经纬度，站点所属区域编号。

表 3-6 city.csv

城市编号	名称	纬度(°)	经度(°)	类别
001	北京	39.904210	116.407394	1
006	天津	39.084158	117.200982	1

表3-6（续表）

城市编号	名称	纬度(°)	经度(°)	类别
011	石家庄	38.042307	114.514860	1

city.csv文件中展示了京津冀范围内所有的市级城市的详细信息：城市编号、城市名称、城市所在位置的经纬度、城市类别(1：北方；2：南方)。

表 3-7 district.csv

区域编号	名称	城市编号
00101	海淀区	001
00102	石景山区	001
00103	丰台区	001

district.csv文件中展示了京津冀区域内所划分的区域(区、县)的详细信息：区域编号、区域名称、区域所属城市。

表 3-7 airquality.csv

站点编号	日期	PM2.5(μg/m^3)	PM10(μg/m^3)	NO2(μg/m^3)	CO(mg/m^3)	O3(μg/m^3)	SO2(μg/m^3)
1001	2014/5/1 00	138	159.4	56.3	0.9	50.8	17.2
1002	2015/4/20 10	45	41.6	56.7	0.7	39.8	2.9
1003	2015/3/10 02	11	30	34	0.6	37	5

airquality.csv文件展示了京津冀区域内每个空气质量监测站每个时刻(小时)采集到的污染物浓度信息，其中需要采集的污染物为：PM25、PM10、NO2、CO、O3、SO2。它们的单位均除了污染物CO为mg/m^{3}以外，其余均为μg/m^3。由于本文中需要用到的是污染物的日均AQI值进行计算，因此需要将时均浓度值转化为日均值，并按照AQI的计算公式将浓度转化为AQI。

表 3-8 meteorology.csv

区域编号	日期	温度(℃)	压强(Pa)	湿度(%rh)	风速(m/s)	风向
616	2015/4/18 03	15	1009	65	1.8	1
1108	2015/4/6 10	7	1027	29	6.12	13
1405	2015/3/30 05	5	1010	85	3	24

meteorology.csv展示了每个区域在每一个时刻(小时)的气象状况。其中包含温度、

压强、湿度、风速、风向五个数据。由于本文中风向需要用角度表示，因此需要根据十六风向图将该数据集中的风向转换为角度，其中风向标识与风向角度的对应关系如下表所示：

表 3-9 风向标识与风向角度对应关系

风向标识	描述	角度	风向标识	描述	角度
0	无风	Null	9	不稳定	Null
1	东风	180°	13	东南风	135°
2	西风	0°	14	东北风	225°
3	南风	90°	23	西南风	45°
4	北风	270°	24	西北风	315°

3.7.3 实验结果与分析

根据空气质量监测站以及气象监测站的历史数据，通过 3.3 节提出的污染物传播代价公式，计算出一个周期内所有时刻的污染物传播路径，并根据 3.4 节提出的构建大气污染物传播网络的方法构建污染物传播网络。最后将污染物传播网络映射至邻接矩阵中，矩阵中每一项表示对应路径在周期中出现的次数，将该矩阵称之为污染物传播网络邻接矩阵。由于站点数量较多，无法展示全部数据，因此，本文只展示了污染物传播网络邻接矩阵部分数据，如下表 3-10 所示。

表 3-10 污染物传播网络邻接矩阵部分数据

	1001	1002	1003	1004	1005	1006
1001	0	0	0	0	0	0
1002	581	0	0	0	0	0
1003	129	663	0	0	0	331
1004	40	146	507	0	0	143
1005	6	29	131	0	0	28
1006	203	0	0	0	0	0

对大气污染物传播网络中站点的度以及度分布进行统计分析，污染物传播网络中站点的度表示该站点受其他站点污染物传播的影响程度以及该站点对其他站点的影响程度，即站点的度越大，表明该站点在网络中的影响力就越大，反之越小。图

3-4 为污染物传播网络中各站点度的分布情况，其中横坐标为站点编号，纵坐标为对应站点的度。

由图 3-4 可以看出，大气污染物传播网络中各站点度的分布呈现离散分布，而且每个度出现的概率也是不均匀的。

图 3-5 为大气污染物传播网络中的度概率分布，其中横坐标为网络中的度值，纵坐标为对应度出现的概率。由图 3-5 可以看出，大部分的站点具有很小的度，只有一小部分站点拥有很大的度，这正是符合了复杂网络中无标的特征。因此，通过上述试验结果分析式可知，大气污染物传播网络具有明显的无标度的特性，因此大气污染物传播网络是一个复杂网络。因此，可以采用复杂网络的研究方法对大气污染物传播网络进行深入研究。

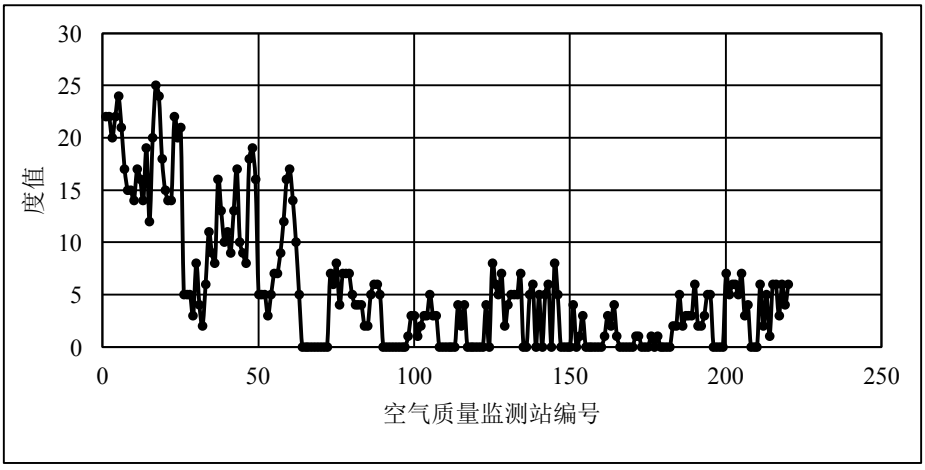

图 3-4 京津冀监测站度分布情况

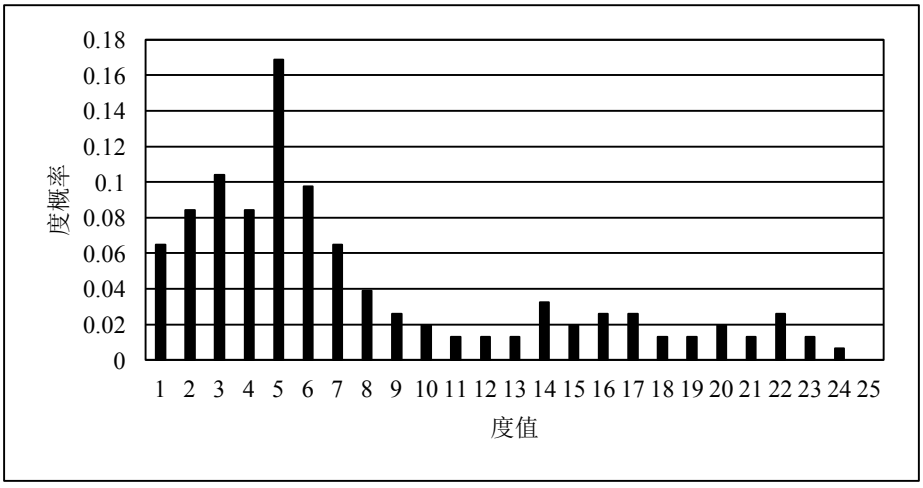

图 3-5 度概率分布

3.8 本章小结

为了更加全面的研究大气污染物的传播规律，本章对大气污染物的传播机理进行了深入的研究，分析出了影响大气污染物传播的相关因素。一个地区的大气污染不仅仅与本地的污染物排放有关系，而且与该地区的相邻地区的大气污染物的输送有着紧密的联系。影响大气污染物传输的因素有许多，其中风速、风向、降雨、湿度、压强、地面粗糙度、地形等都对大气污染物的传输起着至关重要的作用。因此本文依据大气污染物的传播机理，推导出大气污染物的传播代价，以此来衡量大气污染物传播的难易程度。并以此为依据，构建出周期内大气污染物传播网络图，通过对该网络图的特征分析发现大气污染物传播网络图符合复杂网络的特征。因此，可应用复杂网络的方法对大气污染物传播网络进行深入的分析研究。

第 4 章 基于图矩阵的关键传播路径挖掘

4.1 引言

由于大气污染物具有区域性传输的特点，往往会因为一个地区发生污染而导致相邻区域也遭受不同程度的污染，当两个地区之间存在一条在某种气象条件下对污染物传输十分有利的路径时，这条路径在这个时刻就会成为这两个地区的污染物关键传播路径。因此，为了全面掌握一个区域的污染状况，并以此制定出有效的防治策略，就需要挖掘在一个周期内易于污染物传播且频繁出现的路径，并将这些路径视为污染物关键传播路径，并针对这些路径经过的区域制定相应的污染物防治的策略，才能在整体上对污染物的起到防治的作用。

4.2 基本概念与定义

根据 3.4 节获得的大气污染物传播网络有向图$G=\{V, A\}$，为了获得大气污染物的关键传播路径，本文提出了基于图序列的挖掘方法，此方法在挖掘的过程中需要对访问过的边和节点进行标记。因此，为了记录网络中节点或者边的访问情况，接下来定义了大气污染物传播网络有向标识图。

定义 4.1 染物传播网络有向标识图：图$G=\{\langle V, A\rangle, FlagV, fn\}$是一个有向标识图，其中：$\langle V, A\rangle$是一个有向图，其符合大气污染物传播网络有向加权图的定义，$FlagV$是图G的节点标识集合，其作用是表示监测站v_i被访问情况。$FlagV$的值只有 true 和 false 两个。fn是一个映射函数，其作用计算有向图G的节点被访问的情况，并赋予相应的标识，$fn:V\rightarrow FlagV$。

定义 4.2 子图：存在大气污染物传播网络有向标识图$G_1=\{\langle V_1, A_1\rangle, FlagV_1, fn_1\}$和$G_2=\{\langle V_2, A_2\rangle, FlagV_2, fn_2\}$，若图$G_1$和$G_2$的节点、边以及映射函数满足条件$A_1\subseteq A_2$且$V_1\subseteq V_2$且$fn_1\subseteq fn_2$,则将图$G_1$称作图$G_2$的子图，记作$G_1\subseteq G_2$。

定义 4.3 有向图矩阵 *APTM*(Air Pollutant Transport Matrix)：如果已经知道图 G 中的边与节点的出现情况，则创建一个矩阵，矩阵中每个位置的值由 0/1 字符串构成，表示该对站点对应的边的出现情况，将该图序列矩阵称之为大气污染物传播网络有向图矩阵 *APTM*。

定义 4.4 污染物关键传播路径序列：在大气污染物传播网络有向图 G 中，若存在一个频繁连通子图可以通过序列表示，并且由满足频繁阈值和对应 Top-K 子图序列的路径片段组成的路径序列称之为污染物关键传播路径序列。

定义 4.5 站点覆盖率：污染物传播路径 A 与污染物传播路径 B 相同站点的个数与污染物传播路径 B 站点总数只比，即为路径 A 对路径 B 的站点覆盖率。

定义 4.6 污染物关键传播路径覆盖率：在重污染过程中，污染物在关键路径上传播的次数与污染物在所有路径上的传播次数之比，即为污染物关键传播路径覆盖率。

4.3 基于图矩阵的污染物关键传播路径挖掘

本文通过图矩阵的数据结构存储大气污染物在两个地区之间的传播关系，并在此基础上运用图序列的挖掘方法挖掘出大气污染物的关键传播路径。此关键传播路径挖掘过程的技术路线图如图 4-1 所示。

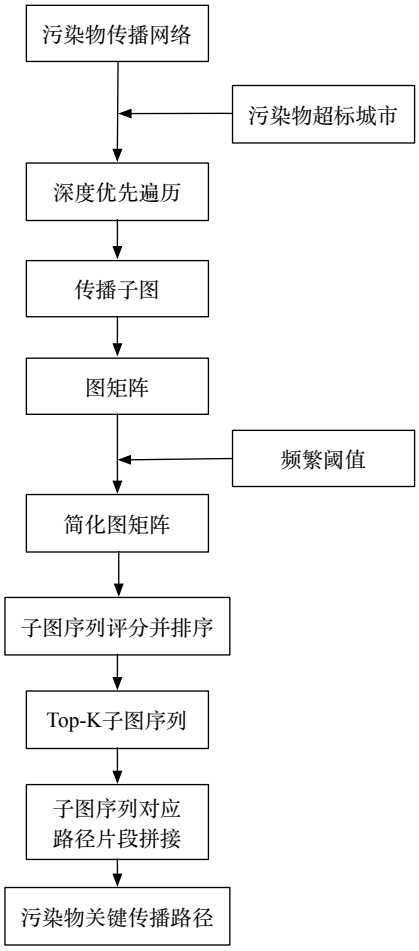

图 4-1 挖掘关键传播技术路线图

首先，根据周期内的空气质量监测站的污染物浓度的数据信息，统计出京津冀地区各个城市污染物的超标率，并将这些超标的城市分别作为大气污染物传播的起点，获取若干个污染物传播的路径，可将这些污染物传播路径看成是大气污染物传播网络有向图的一个子图。并根据定义 4.3，将路径中边的出现情况按照子图的顺序依次用 0/1 表示，构建有向图矩阵 *APTM*。

其次，设置频繁阈值，筛选出满足频繁阈值要求的序列，并将这些满足要求的序列转换成十进制数，便于后续统计操作。

最后，统计出每个子图序列中 1 的个数和每个子图序列下的路径片段，并计算出每个子图序列的评分并排序，取 Top-K 子图序列，将所有序列中的路径片段拼接成路径序列，即大气污染物传播网络的关键传播路径。

4.3.1 获取污染物传播网络子图

为了生成大气污染物传播网络子图，则需要获取大气污染物传播的起始点，本文采用一个周期(年)内污染物超标率比较严重的地区作为起始点，具体算法过程如算法 4.1 所示。然后分别以这些地区为起点对污染物传播网络进行深度优先遍历，从获得若干条大气污染物传播路径序列，这些序列即为大气污染物传播网络子图，具体算法过程如算法 4.1 所示。

算法 4.1 统计一个周期内京津冀各城市的污染物超标率

输入：*HebeiInfo*

输出：*ExcStandarRate*, *AverageCon*

执行过程：

```
BEGIN
1.  Initialize dataMap
2.  for(each info in HebeiInfo)
3.      //对每行数据按逗号进行分割,获取每个时刻的污染物浓度信息
4.      curLineArray = info.split(",")
5.      key = curLineArray[1]
6.      value = Integer.parseInt(curLineArray[6])
7.      if(dataMap.containsKey(key))
8.          dataMap.get(key).set(0, dataMap.get(key).get(0)+value)
```

9. *dataMap*.get(*key*).set(1, *dataMap*.get(*key*).get(0)+1)
10. **if**(*value* ≥50)
11. *dataMap*.get(*key*).set(2, *dataMap*.get(*key*).get(2)+1)
12. **end if**
13. **end if**
14. **else** 将*dataMap*中的值都设置为初始值 0
15. **end for**
16. **for**(each *key* in *dataMap*)
17. *num*=*dataMap*.get(*key*).get(1)
18. *AverageCon* = *dataMap*.get(*key*).get(0)/*num*
19. *ExcStandarRate*= *dataMap*.get(*key*).get(0)/*num*
20. **end for**

END

算法 4.1 中首先初始化*dataMap*用于存储一个地区的污染物浓度总和、污染天数、污染物超标天数，然后对一个周期(年)内的京津冀大气污染物浓度的信息进行遍历，由于每一条信息是按逗号分隔的，因此将其分割，并取得对应城市的编号以及该地区的污染物浓度值(5-7 行)，若*dataMap*中不存在对应的地区的*key*值，则初始化该地区的污染物浓度总和、污染天数、污染物超标天数均为 0，否则将该地区每一时刻(小时)的污染物浓度进行累加，污染天数进行累加，如果该时刻的污染物浓度超出了阈值，则将污染物超标天数进行累加(9-12 行)。最后，按*key*对*dataMap*进行遍历，计算出每个城市一个周期内平均污染物浓度*AverageCon*以及污染物超标率*ExcStandarRate*。算法结束。

算法 4.2 根据传播网络计算传播路径

输入：传播网络邻接矩阵 *M*、城市编号 *cityId*

输出：传播路径集合 *pathList*

BEGIN

1. **Initialize** *pathList*, *list*
2. **for**(each *key* in *cityId*)
3. int[] *len* = {0}
4. *list*.add(DFS(key, map, new Array<String>(), new Array<String>(), len))

```
5.  end for
6.  function DFS(String key, Set map, List marked, List sList, int[] len)
7.      if(marked.contains(key))
8.          if(sList.size > len[0])
9.              maxList = new ArrayList(sList)
10.             len[0] = sList.size()
11.     end if
12.   else
13.       marked.add(key)
14.       sList.add(key)
15.       Set values = map.get(key)
16.       for(each value : values)
17.           DFS(value, map, marked, sList, len)
18.       end for
19.     slist.remove(key)
20.   end else
21.   return maxList
22.  end DFS
END
```

由算法 4.2 可知，以每个城市为起点对污染物传播网络进行深度优先遍历，其中在深度优先遍历时，用*marked*表示站点是否被访问过，如果被未访问过，则将该节点存储至序列*sList*中，并将该节点标示为已被访问状态(13-14 行)，并继续进行深度优先遍历(18 行)。否则，将序列*sList*输出。最终，将可获取以每个城市为起点的所有大气污染物传播路径。

4.3.2 构建有向图矩阵

由算法 4.3 获取的大气污染物传播路径即为大气污染物传播网络有向图的子图，然后根据子图中的信息来构建定义 4.4 中的有向图矩阵 *APTM*。首先，获取子图中所有的节点。其次，顺序遍历所有子图，将站点编号看作矩阵的行列号，如果子图中有对应边的存在，则在对应子图的位置上置为 1，否则置为 0。再次，在所有的子图

中，可能某条边在多个子图中出现过，所以将所有子图对应边位置上的 0/1 值拼接起来形成新的字符串。最后，将刚才得到的字符串序列存储到矩阵的对应位置中，即形成了有向图矩阵 *APTM*。其具体的算法过程如下所示。

算法 4.3 构建有向图矩阵

输入：$L=\langle L_1, L_2, \cdots, L_n \rangle$

输出：*APTM*

```
BEGIN
1.  Initialize  nodeList
2.  for(each l in L )
3.      nodes = l.split(",")
4.      for(each node in nodes)
5.        if(!nodeList[node]) nodeList.add(node)
6.         else continue;
7.      end for
8.  end for
9.  LineLenght = L.length
10. nodeLength = nodeList.length
11. LineMatrix=zeros[nodeLength ,nodeLength ]
12. for(i=0;i≤LineLenght ;i++)
13.     nrow = LineLenght [i,1]
14.     ncol = LineLenght [i,2]
15.     LineMatrix_i[nrow,ncol]=1
16. end for
17. for(i=0;i≤LineLenght ;i++)
18.    e_i=strcat(LineMatrix_i)
19.    put all e_i to APTM
20. end for
END
```

在算法 4.3 中，L是所有大气污染物传播网络的子图的集合，$LineMatrix_i$是根据L_i中的三列构造的邻接矩阵。首先，初始化矩阵*nodeList*，用来存储所有子图中涉及到

的节点(2-8 行)，根据*nodeList*中的节点来初始化矩阵，节点编号作为行列号。其次，将每个大气污染物传播网络子图转换成邻接矩阵$LineMatrix_i$(12-16 行)，并将每个$LineMatrix_i$矩阵中的元素转换成字符，把每个$LineMatrix_i$矩阵中对应行列位置的字符拼接起来，形成字符串序列e_i(17-18 行)。最后，将所有的e_i序列按照对应的行列存储到*APTM*矩阵中(19 行)。算法结束。

4.3.3 构建简化图矩阵

对算法 4.3 构造出的大气污染物传播网络有向图矩阵*APTM*按照规则依次对每一项进行判断筛选，从而获得简化图矩阵*SimpleM*。首先，统计出矩阵中每一项子图序列中 1 出现的次数，通过预先设置的频繁阈值对每条边进行判断，如果超过了频繁阈值，则保留，否则，将该边对应的子图序列中的每个位置置为 0。最后，将矩阵中的所有二进制数转换为十进制，为后续挖掘污染物关键传播路径提升效率做准备。

算法 4.4 构建简化图矩阵*SimpleM*

输入：矩阵*APTM*和频繁阈值*s*

输出：*SimpleM*

BEGIN

1. **for**(each e_i in *APTM*)
2. **if**(computed(1) in e_i >*s*){ // computed(1)表示计算e_i中 1 的个数
3. add bin2dec(e_i) to *SimpleM*
4. **end if**
5. **else** e_i置 0
6. **end for**

END

在算法 4.4 中，computed(1)是对e_i序列中 1 的个数的统计方法。首先，对有向图矩阵*APTM*进行遍历搜索判断，对每个e_i序列中 1 的个数进行统计判断(2 行)，若computed(1)的值小于预先设置的频繁阈值*s*，则认为该边不是频繁出现的，则将e_i置为 0，并将该序列从向图矩阵*APTM*中移除。若computed(1)大于*s*，将e_i序列转换成十进制数并根据对应的行列号存入简化图矩阵*SimpleM*中(4 行)。否则，将e_i置为 0(3 行)，最后得到简化图矩阵*SimpleM*。算法结束。

4.3.4 挖掘污染物关键传播路径

大气污染物的关键传播路径是由若干符合频繁阈值要求具有周期规律的路径片段组成。通过算法 4.4 获得污染物传播网络G中的频繁出现的边，并根据这些边生成了简化图矩阵*SimpleM*。然后采用深度优先遍历的方法对矩阵*SimpleM*进行遍历，搜索数值相同的路径，并统计出相同数值路径的个数m以及数值对应二进制数中 1 的个数n。在简化图矩阵*SimpleM*中，会存在两种子图序列：一种是该子图序列下的路径片段很少，但是子图序列中涉及到的子图很多；另一种是该子图序列下的路径片段很多，但是子图序列中涉及到的子图却很少。因此，为了更有效的评价一个子图序列下的路径片段是否可作为污染物关键传播路径里的一部分，本文提出了一种衡量标准，如公式(4-1)所示。

$$S=\alpha\cdot m+(1-\alpha)\cdot n \tag{4-1}$$

公式(4-1)中，S为子图序列下的路径片段的评分。n表示了路径片段在所有子图中的出现次数；m表示了在同一个子图序列下的路径片段个数。α表示调节系数，用来调整m、n所占的比重，本文中α取值为 0.5。

得到评分之后，选取Top-K子图序列下的路径片段，其中，K 取值所有传播子图个数的一半。并将这些路径片段拼接成序列，此序列即为大气污染物关键传播路径。具体的算法实现如下所示。

算法 4.5 污染物关键传播路径挖掘算法

输入：*SimpleM*

输出：*APCP*

BEGIN

1. **Initialize** *mapPath*，*mapNum*，*score*
2. **for**(each e_i in *SimpleM*)
3. **if**($e_i \neq 0$) continue
4. **else**
5. *mapPath*.put(e_i, compute(e_i)) // compute(e_i)计算数值e_i出现次数
6. // computeNum(e_i)计算e_i中 1 出现的次数
7. *mapNum*.put(e_i, computeNum(e_i))
8. **end for**

9. **for**(each e_i in *mapPath*)
10. *score*.put(e_i, 0.5**mapNum*.get(e_i)+0.5**mapNum*.get(e_i))
11. **end for**
12. Sort(*score*) //对*score*进行排序
13. *NewS* = GetTopK(*score*) //取 top-k 传播模式
14. **for**(each S_i in *NewS*)
15. 将每个传播模式下的路径片段拼接
16. **end for**

END

算法 4.5 中首先对*SimpleM*矩阵进行遍历，判断矩阵中每一项的值是否为 0，如果为 0，则判断下一条路径片段。否则将通过computeNum()方法计算每一项中 1 的个数，并存储到*mapNum*(7 行)；通过compute()方法统计具有相同值的路径个数，并存储到*mapPath*(5 行)。其次，计算每个子图序列的评分，并将结果存储到*score*中(9-11 行)。再次，对所有子图序列的评分进行排序(12 行)，取 Top-K 子图序列(13 行)。最后，将每个子图序列下的路径片段拼接成序列(14-16 行)，此序列即为大气污染物关键传播路径。算法结束。

4.3.5 算法实例

由于污染物传播网络节点和边数目十分多，因此本章以图 4-1 为例，来对大气污染物关键传播路径算法展开分析。

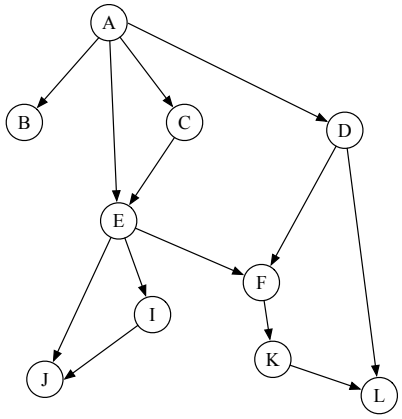

图 4-1 大气污染物传播网络图

如图所示 4-1 所示，以站点A为起点，对整个污染物传播网络图进行深度优先遍

历，共得到 9 条污染物传播路径。其具体内容如表 4-1 所示。

表 4-1 传播路径节点序列

序号	路径
1	A B
2	A E J
3	A E I J
4	A E F K L
5	A C E J
6	A C E I J
7	A C E F K L
8	A D F K L
9	A D L

如表 4-1 所示，由于每条传播路径的长度各不相同，其中包含的节点也不同，因此需要将每条传播路径转换成邻接矩阵，矩阵的大小应该由节点的个数来决定。因此，该邻接矩阵是一个10×10大小的矩阵，然后根据每条路径片段在每个子图中的出现情况，进一步生成以A为起点的污染物传播网络有向矩阵 *APTM*，矩阵中的每一项表示该路径片段在每个子图中出现的情况，如表 4-2 所示。

表 4-2 大气污染物传播网络有向图矩阵 *APTM*

	A	B	C	D	E	F	J	I	K	L
A	00000	10000	00001	00000	01110	00000	00000	00000	00000	00000
	0000	0000	1100	0011	0000	0000	0000	0000	0000	0000
B	00000	00000	00000	00000	00000	00000	00000	00000	00000	00000
	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000
C	00000	00000	00000	00000	00001	00000	00000	00000	00000	00000
	0000	0000	0000	0000	1100	0000	0000	0000	0000	0000
D	00000	00000	00000	00000	00000	00000	00000	00000	00000	00000
	0000	0000	0000	0000	0000	0010	0000	0000	0000	0001
E	00000	00000	00000	00000	00000	00010	01000	00100	00000	00000
	0000	0000	0000	0000	0000	0110	1000	1000	0000	0000

表 4-2（续表）

	A	B	C	D	E	F	J	I	K	L
F	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000	00010 0110	00000 0000
J	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000
I	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000	00100 1000	00000 0000	00000 0000	00000 0000
K	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000	00010 0110
L	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000	00000 0000

通过算法 4.6 来统计有向图矩阵 *APTM* 中每一项里 1 的个数，并设置频繁阈值为 2，通过统计可知，边<AB>、<DF>、<DL>中出现了 1 次 1，因此将这些边所在位置置为 0，然后将剩余的子图序列由二进制转换为十进制数，所生成的矩阵即为简化图矩阵*SimpleM*。具体结果如下表所示。

表 4-3 简化图矩阵*SimpleM*

	A	B	C	D	E	F	J	I	K	L
A	0	0	28	3	224	0	0	0	0	0
B	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	28	0	0	0	0	0
D	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	38	136	72	0	0
F	0	0	0	0	0	0	0	0	38	0
J	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	72	0	0	0
K	0	0	0	0	0	0	0	0	0	38
L	0	0	0	0	0	0	0	0	0	0

在表 4-3 中，经过频繁阈值的筛选和十进制数转换之后，数据的分布情况可清晰

的呈现出来。在简化图矩阵*SimpleM*中，经过算法 4.6 的筛选之后，矩阵中剩余的二进制数分别为 3，28，224，72，38，136。其中各个值对应的边的情况如下表所示。

表 4-4 各值对应路径片段

值	n(次)	m(次)	路径片段
3	2	1	AD
28	3	2	AC、CE
224	3	1	AE
72	2	2	EI、IJ
38	3	3	EF、FK、KL
136	2	1	EJ

根据公式(4-1)对表 4-4 中个值进行评分计算，表中 n 表示子图序列中 1 的个数，m 表示子图序列对应的路径片段个数。各值对应的评分如表 4-5 所示。

表 4-5 各值评分及名次

值	评分	名次
3	1.5	4
28	2.5	2
224	2	3
72	2	3
38	3	1
136	1.5	4

由表 4-4 可知，共有 6 个不同的值，因此取 Top-3 个值对应的所有路径片段，即 38、28、72、224 对应的路径片段。最后将所有路径片段拼接起来，可得以 A 为起点的污染物关键传播路径为 ACEFKL、AEFKL、ACEIJ、AEIJ。该污染物的关键传播路径图如图 4-2 所示。

则该污染物关键传播路径的平均节点覆盖率和路径覆盖率为：

$$\left(\frac{0}{2}+\frac{3}{3}+\frac{4}{4}+\frac{5}{5}+\frac{4}{4}+\frac{5}{5}+\frac{6}{6}+\frac{4}{5}+\frac{2}{3}\right)/9=83\%$$

由此可见，该污染物关键传播路径上的节点覆盖了污染物传播网络中 83%的节点，因此证明了该方法的合理性。

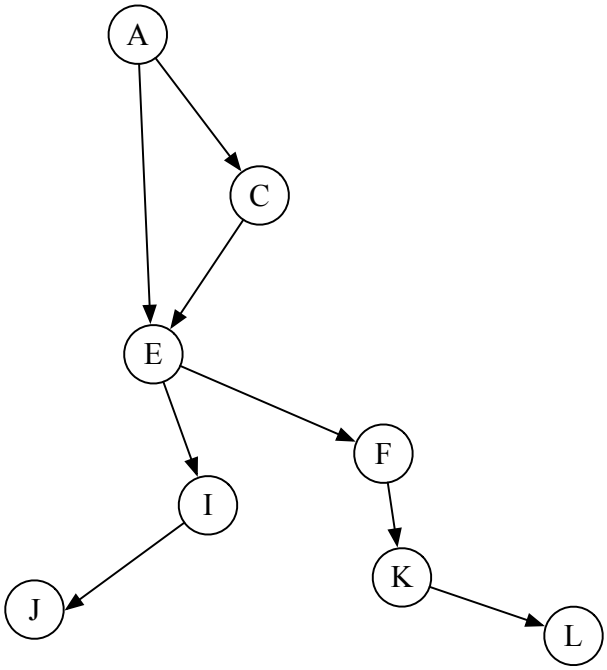

图 4-2 以 A 为起点的污染物关键传播路径

4.4 实验结果与分析

4.4.1 实验环境配置

实验的系统配置为 macOS 10.13.2 操作系统，硬件配置为 Intel Core i5，主频为 3.10GHz，内存为 16 GB 2133 MHz LPDDR3。算法采用 Java，Matlab 编写。数据展示采用 Echarts 以及 Excel，运行环境为 IntelliJ IDEA，JDK 版本为 1.7，数据库采用 MySQL 数据库。

4.4.2 实验数据

本文采用的实验数据为第 4.2 节中产生的一个周期内的大气污染物传播网络的邻接矩阵以及京津冀地区 2014 年 5 月 1 日到 2015 年 4 月 30 日的空气质量数据集 airquality.csv 以及气象数据集 meteorology.csv。

4.4.3 实验结果分析

通过算法 4.2 统计出一个周期内京津冀各个城市的污染物超标率，如图 4-3 所示。由图可见廊坊、衡水、石家庄、唐山、沧州、保定六个城市的超标率超过了 65%。

由图 4-3 可知廊坊、衡水、石家庄、唐山、沧州、保定这 6 个地区的空气污染比较严重，而秦皇岛、北京、承德、天津空气污染情况相对较轻，只有张家口污染物超标率低于 30%，是空气质量最好的城市。这些数据的统计结果与近几年京津冀各城市的空气质量状况基本吻合，由此证明统计结果的准确性。因此，选取廊坊、衡水、石家庄、唐山、沧州、保定六个地区中的监测站为起点，对污染物关键传播路径进行挖掘。

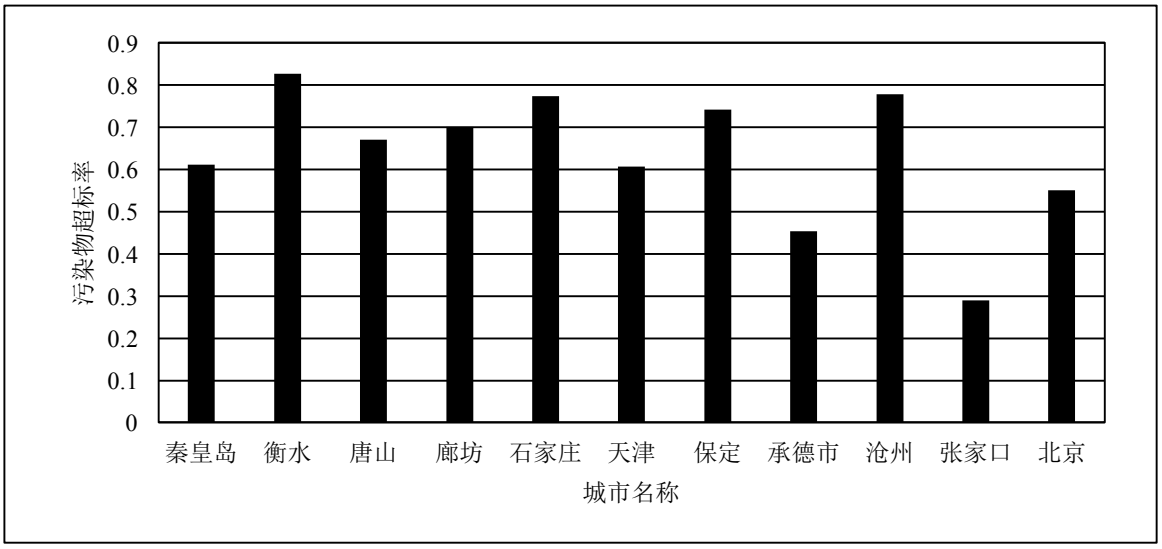

图 4-3 京津冀各城市污染物超标率

首先，根据算法 4.2 计算出以每个地区为起点的污染物传播子图个数以及节点个数。经计算，各地区的污染物传播子图个数以及节点个数情况如表 4-5 所示：

表 4-5 各地区污染物传播子图及节点个数表

地区	传播子图个数	节点个数
廊坊	8	26
衡水	10	30
石家庄	16	32
唐山	9	19
沧州	11	38
保定	15	23

其次，根据算法 4.4 和算法 4.5 计算出每个地区污染物传播网络的有向图矩阵以及简化矩阵*SimpleM*。但是，由于每个矩阵中的数据量十分巨大，因此本章不再对污

染物传播子图、构建有向图矩阵以及构建简化矩阵*SimpleM*的数据进行展示，具体的计算过程可见 4.3.5 节的算法实例说明，本章只对符合频繁阈值的边进行展示，如表 4-6~4-11 所示。

表 4-6 以廊坊地区监测站为起点的路径片段表

值	n(次)	m(次)	路径片段	评分	名次
184	4	4	22010，1023、1023，1024、1024，1031、1031，19009	4.0	1
71	4	4	1024，1027、1007，1024、1027，1028、1028，1032	4.0	1
69	3	3	22005，17011、17011，1005、1005，1007	3.0	2

表 4-7 以衡水地区监测站为起点的路径片段表

值	n(次)	m(次)	路径片段	评分	名次
957	8	5	23009，21011、21011，21009、21009，21006、21006，22007、22007，22009	6.5	2
1023	10	8	22009，22005、22005，17011、17011，1005、1005，1007、1007，1024、1024，1027、1027，1028、1028，1032	9.0	1
956	7	1	23006，23009	4.0	3
432	4	1	23013，23006	2.5	4

表 4-8 以石家庄地区监测站为起点的路径片段表

值	n(次)	m(次)	路径片段	评分	名次
65535	16	13	17027，17007、17007，17009、17009，17015、17015，17022、17022，22009、22009，22005、22005，17011、17011，1005、1005，1007、1007，1024、1024，1027、1027，1028、1028，1032	14.5	1
65135	13	1	18001，17027	7.0	2
65134	12	1	11025，18001	6.5	3
65130	11	1	11024，11025	6.0	4
60000	7	1	11010，11024	4.0	5

表 4-9 以唐山地区监测站为起点的路径片段表

值	n(次)	m(次)	路径片段	评分	名次
168	3	6	13018，6015、6015，1033、6015，1033、6015，1033、6015，1033、6015，1033	4.5	1
343	6	1	13011，13013	3.5	2
327	5	1	13014，13011	3.0	3
321	3	1	13016，13014	2.0	4

表 4-10 以沧州地区监测站为起点的路径片段表

值	n(次)	m(次)	路径片段	评分	名次
1729	5	10	21006，22007、22007，22009、22009，22005、22005，17011、17011，1005、1005，1007、1007，1024、1024，1027、1027，1028、1028，1032	7.5	1
705	4	1	21009，21006	2.5	3
314	5	1	21010，21007	3.0	2

表 4-11 以保定地区监测站为起点的路径片段表

值	n(次)	m(次)	路径片段	评分	名次
30207	13	2	22009，22005、22005，17011	7.5	2
25892	6	1	17007，17009	3.5	6
25900	7	1	17009，17015	4.0	5
30077	11	1	17015，17022	6.0	4
30079	12	1	17022，22009	6.5	3
32767	15	6	17011，1005、1005，1007、1007，1024、1024，1027、1027，1028、1028，1032	10.5	1

表 4-6~4-11 中 *n* 表示子图序列中数字 1 的个数，*m* 表示相同子图序列对应边的个数。根据表 4-6~4-11 中的数据，取 Top-K 子图序列，将这些子图序列对应的路径片段拼接成路径序列。如表 4-12 所示。

表 4-12 各城市污染物关键传播路径

地区	关键传播路径
廊坊	22010，1023，1024，1031，19009 1024，1027，1028，1032 1007，1024，1031，19009
衡水	23009，21011，21009，21006，22007，22009，22005，17011，1005，1007，1024，1027，1028，1032
石家庄	11025，18001，17027，17007，17009，17015，17022，22009，22005，17011，1005，1007，1024，1027，1028，1032
唐山	13018，6015，1033，1023，1024，1031，19009 13011，13013
沧州	21006，22007，22009，22005，17011，1005，1007，1024，1027，1028，1032 21010，21007
保定	17022，22009，22005，17011，1005，1007，1024，1027，1028，1032

根据表 4-12 中的大气污染物关键传播路径计算出每个地区关键路径上站点的覆盖率，如表 4-13 所示。

表 4-13 各地区污染物关键传播路径站点和路径覆盖率

地区	节点覆盖率	路径覆盖率	地区	节点覆盖率	路径覆盖率
廊坊	66.69%	65.79%	唐山	64.87%	65.38%
衡水	79.37%	64.44%	沧州	64.92%	69.23%
石家庄	89.43%	72.09%	保定	80.53%	53.66%

由表 4-13 可知，各地区的污染物关键传播路径上的站点覆盖率基本都超过了 60%，而且在重污染过程中在各地区的污染物关键传播路径上的传播次数均达到了半数以上，甚至有些在 70%左右，因此可认为这些污染物关键传播路径的有效性以及可行性。最后，通过 Echarts 将各地区的污染物关键传播路径绘制在京津冀区域的地图上，即可获得京津冀地区污染物关键传播路径图。如下图 4-4 所示。

其中，红色路线为各城市的污染物关键传播路径，其中部分路径存在重叠。由图 4-4 可以看出，在京津冀地区的西南部，污染物沿着太行山脉的走势由石家庄地区将污染物输送至北京地区，途中与保定、廊坊地区的污染物传播路径相汇合；东南

方向污染物由沧州和天津输送至北京地区，途中与廊坊地区的污染物传播路径相汇合。东北方向由唐山地区沿燕山山脉传播至北京地区，途中与天津地区的污染物传播路径相汇合。由此可见，此污染物的关键传播路径与实际中的地理特征以及气象特征基本吻合，即证明了此挖掘关键路径方法的可行性。

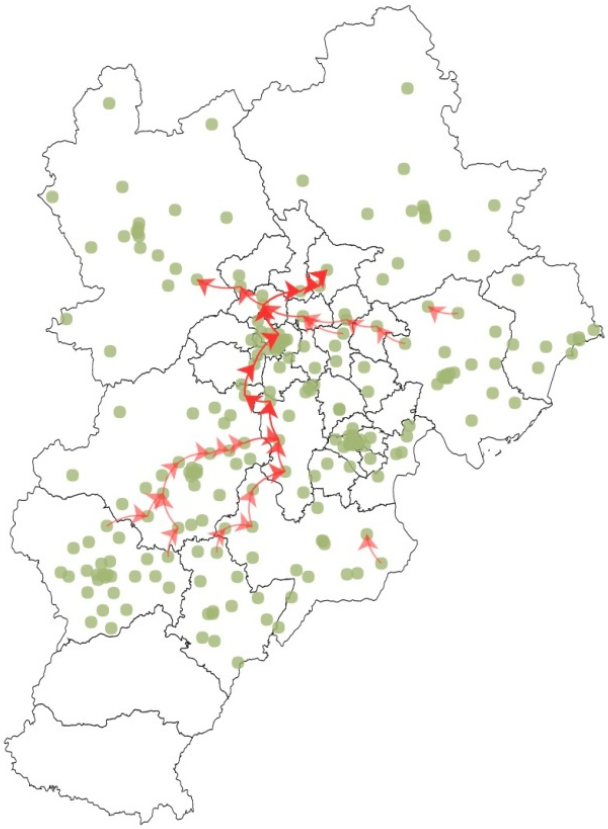

图 4-4 京津冀周期内污染物关键传播路径

4.5 本章小结

本章在大气污染物传播复杂网络的基础上进行对大气污染物关键传播路径的研究。首先，统计京津冀地区污染物严重的地区，在污染物传播网络的基础上，挖掘出各个地区的污染物传播子图。然后，运用图矩阵的方法，存储每条路径片段的子图序列。最后，计算出各个子图序列的评分并排序，取 Top-K 子图序列，将其中的路径片段拼接成路径序列，从而挖掘出污染物关键传播路径，并对挖掘出的污染物关键传播路径进行了验证分析。

第 5 章 基于 StationRank 污染物传播网络重要节点挖掘

5.1 引言

根据第三章提出的大气污染物传播机理模型，构建出了一个周期内的京津冀地区的大气污染物关键传播路径网络，并证实了该网络符合复杂网络的特性。在大气污染物关键传播路径网络中，挖掘污染物传播网络中重要节点，即挖掘京津冀地区内最易受到大气污染物污染的地区。挖掘京津冀地区内最易受到大气污染物污染的地区的意义在于通过挖掘网络中的重要节点，在该区域建立更加完善的大气污染物监测体系，有效的治理该区域的污染状况，实现有效的大气污染区域联防；也有利于在准确的位置设置空气质量监测站，从而大大减少了由于建设监测站带来的成本。

5.2 重要节点评估方法

实际上，几乎所有的复杂系统，比如社会、生物、信息、交通运输系统都可以用复杂网络表示。其中，复杂网络中的关键节点就是相比网络中其他节点而言，可以更大程度上影响网络的功能与结构的一些特殊节点。通常来说，关键节点可以非常快的影响到网络中的大部分节点。如下列举了几种基于复杂网络基础的方法：

度中心性：认为网络中节点的邻居节点越多，其影响力就越大。即通过计算网络中节点的度(入度+出度)来表征网络中节点的中心度，如果网络为加权网络，则通过计算节点的度强度来衡量节点的中心度。

特征向量中心法：认为节点的不仅仅取决于自身邻居节点的数目，也取决与每个邻居节点的重要程度，即节点上的权重。因此通过节点邻居节点权重之和来表征节点的重要程度。

节点删除法：破坏性反映重要性。通过去除网络中的某个节点，来计算该节点的失去对网络的破坏程度，从而表征该节点在网络中的重要程度。

节点收缩法：将一个节点和它相邻节点收缩成一个新的节点，如果该节点十分重要，收缩后整个网络将会更好的聚集在一起，从而以此标准来衡量节点的重要性。

以上列举的网络关键节点的识别办法主要是应用于无向网络中，而本文所要研究的大气污染物传播网络是有向加权网络，因此不是十分适用于大气污染物传播网

络网络。而 PageRank 算法应用于有向网络挖掘关键节点是十分有优势的，但是由于大气污染物传播网络是一个有向加权网络，而且每个站点的重要性受周围各站点的影响很大但不是等概率分配的，因此需要对原始的 PageRank 算法进行改进。最终，本文在原始 PageRank 算法的基础上提出了 StationRank 算法，该算法结合了相邻边的权重以及邻居节点的重要性。

5.3 基于 StationRank 的污染物传播网络重要节点挖掘

5.3.1 PageRank 算法理论

PageRank 简化计算模型[39]公式(5-1)如下所示：

$$\mathrm{PageRank}(X)=\sum_{i\epsilon B}\frac{\mathrm{PageRank}(Y_i)}{C_{out}(Y_i)} \tag{5-1}$$

公式(5-1)中，PageRank(X)表示页面X的重要度，即PageRank值；PageRank(Y_i)表示每个链接到页面X的页面Y_i的PageRank值；$C_{out}(Y_i)$表示页面Y_i中所有页面链接的数量；B表示所有链接到页面X的页面集合。

根据 PageRank 理论[40]可知，PageRank 的计算公式(5-2)如下：

$$\mathrm{PageRank}(X)=\frac{1-\alpha}{N}+\alpha\sum_{i\epsilon B}\frac{\mathrm{PageRank}(Y_i)}{C_{out}(Y_i)} \tag{5-2}$$

公式(5-2)中，PageRank(X)表示页面X的PageRank值；PageRank($\mathrm{Y_i}$)表示每个链接到页面X的页面Y_i的PageRank值；$C_{out}(Y_i)$表示页面Y_i中所有页面链接的数量；B表示所有链接到页面X的页面集合；α表示阻尼系数，取值为[0~1]；N表示网页总数。

阻尼系数α表示用户浏览某个网页时继续浏览其他网页的概率。页面X中的PageRank值中来自其可连接到的的网页的贡献占页面X所有的PageRank值α，其余的1-α部分分配在在整个网络的网页中，通常α取值为 0.85[41,42]。也就是一个网页X的85%的PageRank值来自于能够链接到页面X的网页，剩下的15%则可能来自用户通过从浏览器地址栏输入 URL 访问网页X。

5.3.2 PageRank 算法改进

在现实世界中，社会关系网络、飞机航线网络、流行病毒网络等都为加权有向网络，PageRank 算法在这些网络中都有这重要的应用。在本文构建的大气污染物传

播网络中，空气质量监测站为网络中的节点，两个站点之间的污染物传播路径为网络中的边，边上的权重为一个周期内站点之间大气污染物传播路径出现的次数。由此可见，大气污染物传播网络也是一个有向加权网络。但是，由于原始 PageRank 算法无法满足于本文构建的大气污染物传播网络。因此，本文提出一种应用于大气污染物传播网中的基于 PageRank 算法改进后的 StationRank 算法。此算法不仅考虑了大气污染物传播的方向性、站点之间的路径在周期中出现的次数，并且对 PageRank 算法中的转移概率做出了适用于大气污染物传播网络的调整。

PageRank 算法中的从页面i到页面j转移概率是按照公式(5-3)进行的平均分配。

$$p_{ij}=\frac{1}{N(i)} \tag{5-3}$$

公式(5-3)中，p_{ij}表示从污染物从站点i到站点j转移概率，$N(i)$表示污染物可由站点i传播到的所有站点的数量。

但是，由于大气污染物传播网络的复杂性，站点之间污染物的转移概率被站点所有链接均分是不科学的。因为，有的路径在周期中出现的次数很多，也就代表了该条路径是一条有利于污染物传输的通道，因此，这条路径上污染物的转移概率就应该比出现次数少的路径转移概率大。由此，可得出改进后的大气污染物传播网络的概率转移公式为：

$$p_{ij}=\frac{W_{ij}}{\sum_{j\in G(i)} W_{ij}} \tag{5-4}$$

公式(5-4)中，W_{ij}表示从站点i到站点j边上的权值，即该边出现的次数；$G(i)$表示污染物可由站点i传播到的所有站点的集合。因此最终得出了评价节点重要性的指标 StationRank 值的计算公式为：

$$SR(i)=\alpha\sum_{i=1}^{n} p_{ji}\,SR(j)+(1-\alpha)\cdot\beta \tag{5-5}$$

公式(5-5)中，$SR(i)$表示监测站i的 StationRank 值；$SR(j)$表示大气污染物输送路径$j\rightarrow i$ 中监测站j的 StationRank 值；β表示监测站当地发生污染的概率；p_{ij}表示转移概率；N表示污染物传播网络中的所有站点。

由公式(5-5)可知，$SR(i)$由两部分构成：一部分是由于自身发生污染导致自身 StationRank 值增加；另一部分是监测站i周边其他地区对监测站i所在地区污染物状况的贡献值。由公式(5-5)可得出，一个站点的入度强度越大，该站点就越容易受到其

他地区污染物的影响，其在网络中就越重要，其 StationRank 值就越大。

因此，公式(5-5)不仅考虑了监测站周边地区的污染状况，也考虑了其自身的污染状况，因此该公式更加适合本文提出的大气污染物传播复杂网络。

5.3.3 StationRank 算法实现

StationRank 算法以大气污染物传播网络为基础，对网络中的所有监测站进行打分排序。StationRank 算法的主要步骤如下：

算法 5.1 StationRank 算法

输入：大气污染物传播网络$G=\{V, A, W\}$

输出：站点评分排序

BEGIN

(1) 输入大气污染物传播网络G

(2) 根据公式(5-4)计算站点i到站点j的转移概率，并将每个值存储至矩阵P

(3) 根据公式(5-5)计算 StationRank 矩阵SR

(4) 使用幂法迭代求解 StationRank 矩阵SR中特征值为 1 对应的特征向量，即平稳分布

(5) 对结果进行节点评分排序

END

其中，第 4 步中求解 StationRank 矩阵SR平稳分布的幂法迭代的算法过程如下：

算法 5.2 StationRank-Interate

输入：StationRank 矩阵SR，转移概率矩阵P

输出：NR

BEFIN

1. $SR_0 \leftarrow e/n$
2. $k \leftarrow 1$
3. **repeat**
4. $SR_{k+1}=P \cdot SR_k$
5. $k \leftarrow k+1$
6. **until** $\left\| SR_k - SR_{k+1} \right\| < \varepsilon$
7. **return** SR_k

END

算法 5.2 中初始每个站点的 StationRank 依旧按照原始 PageRank 进行计算，若两次迭代后差值小于临界值ε，则停止继续迭代。

5.4 实验结果与分析

根据算法 5.1 的步骤，对京津冀地区空气污染物传播网络图采用 StationRank 算法，对其中的站点进行评分排序，并挖掘出网络中的重要节点。表 5-1 为算法 StationRank 评分前 20 的空气质量监测站，即重要节点。

表 5-1 排名前 20 空气质量监测站 StationRank 值

排名	站名	StationRank	排名	站名	StationRank
1	京西南琉璃河	0.10534	11	安平县环保局	0.02253
2	无极环保局	0.10475	12	大兴黄村镇	0.02166
3	满城税务局	0.05558	13	徐水环保局	0.01976
4	新乐市委东楼	0.05090	14	栾城通讯公司	0.01940
5	深泽供电局	0.03348	15	晋州博纳德	0.01847
6	高邑县政府	0.03172	16	赵县环保局	0.01835
7	正定联通公司	0.02789	17	房山良乡	0.01814
8	行唐县委办公楼	0.02461	18	任丘华油八处	0.01782
9	元氏住建局	0.02407	19	天津团泊洼	0.01738
10	亦庄开发区	0.02329	20	藁城实验学校	0.01635

根据表 5-1 中的排名前 20 空气质量监测站，统计出每个监测站一年的污染物超标率如下表所示

表 5-2 重要节点污染物超标率

排名	站名	污染物超标率	排名	站名	污染物超标率
1	京西南琉璃河	33.05%	11	安平县环保局	28.87%
2	无极环保局	32.28%	12	大兴黄村镇	28.77%
3	满城税务局	30.71%	13	徐水环保局	28.72%
4	新乐市委东楼	30.63%	14	栾城通讯公司	28.39%
5	深泽供电局	30.01%	15	晋州博纳德	27.75%

表 5-2（续表）

排名	站名	污染物超标率	排名	站名	污染物超标率
6	高邑县政府	29.42%	16	赵县环保局	27.56%
7	正定联通公司	29.32%	17	房山良乡	27.55%
8	行唐县委办公楼	29.2%	18	任丘华油八处	27.47%
9	元氏住建局	29.09%	19	天津团泊洼	27.44%
10	亦庄开发区	29.08%	20	藁城实验学校	27.25%

由表 5-2 可知，空气质量监测站的污染物超标率与 StationRank 值的趋势基本相符，随着 StationRank 值的降低，监测站的污染物超标率也随之降低。由此可证明 StationRank 算法的可行性以及正确性。

图 5-1 为通过 StationRank 算法计算出的空气质量监测站 StationRank 值排序的曲线图，横坐标为站点关键度排序编号，纵坐标为各站点的 StationRank 值。然后，根据 StationRank 值排序后的站点编号，对每个站点的入度强度进行统计，如图 5-2 所示，横坐标为站点关键度排序编号，纵坐标为各站点的入度强度。

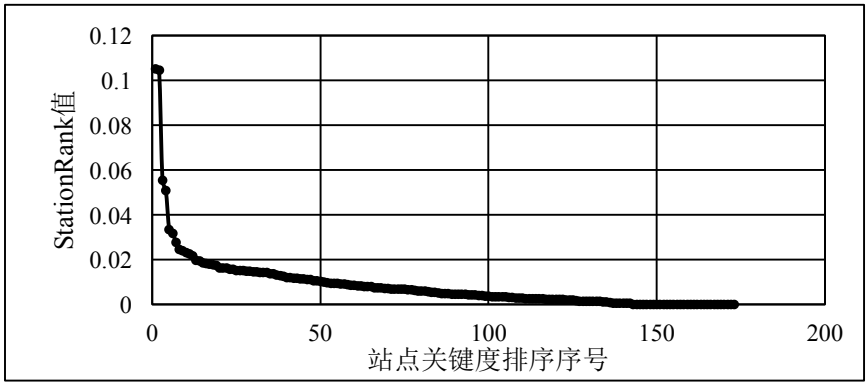

图 5-1 空气质量监测站 StationRank 值排序

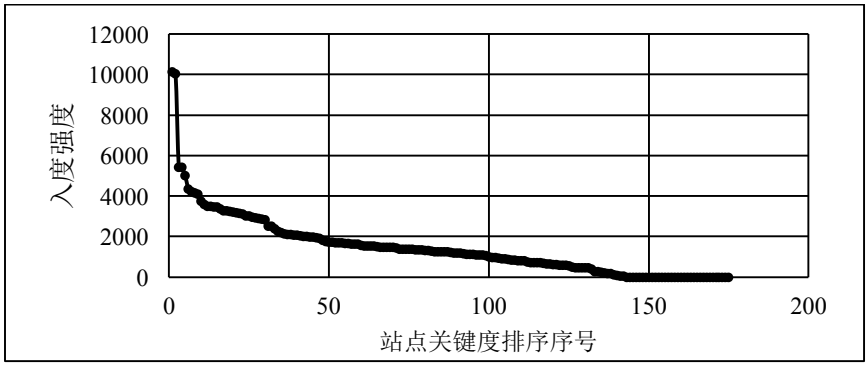

图 5-2 空气质量监测站入度强度

由图 5-1 和图 5-2 可知，随着站点关键度排序序号的增大，站点的入度强度和 StationRank 值下降的趋势基本吻合，两者基本符合正相关系，因此证明了 StationRank 算法的正确性。

5.5 本章小结

本章主要是在大气污染物传播网络图的基础上，对网络中的重要节点进行了挖掘。首先，对大气污染物传播网络中的节点进行了分析，并列举了常见的几种重要节点评价指标；其次，简述了传统 PageRank 算法，并根据构建的大气污染物传播网络，对传统 PageRank 算法中的节点数和转移概率两方面进行了改进，最终提出了算法 StationRank，并详细描述了算法 StationRank 的实现过程；最后，给出了试验结果并对试验结果进行了分析，分别与实际各站点污染物超标率、各站点的入度强度分布进行了对比，发现它们分布情况十分接近，因此证明了挖掘出的重要节点的有效性。

结 论

随着我国经济的飞速发展，大气污染物问题越来越引起了人们的重视，而大气污染物传播有着明显的区域性特点。因此，有效的防治大气污染物传播，改善区域内的空气质量就显得尤为重要。本文在京津冀区域内空气质量监测站以及气象监测站历史数据的基础上，做出了以下研究：

(1) 提出了一种基于大气污染物传播机理的污染物传播模型。分析大气污染物的传播机理，筛选出影响污染物传播的主要因素，如风、降雨量、海拔、水平距离等因素。并以此为依据，构建大气污染物传播模型，其值为大气污染物传播代价。

(2) 提出了一种基于图矩阵的大气污染物关键传播路径的方法。首先，并在大气污染物传播代价模型的基础上构建大气污染传播网络，经分析发现大气污染传播网络符合复杂网络的特征。然后，统计出周期内污染物超标的城市，以这些地区作为起点挖掘出污染物传播子图，并将每个子图序列映射到图矩阵中。最后在污染物传播图矩阵的基础上，计算出各个子图序列的评分并排序，取 Top-K 子图序列，将其中的路径片段拼接成路径序列，从而挖掘出污染物关键传播路径。此方法具有较低的时间复杂度，而且挖掘出的路径具有代表性，也基本符合京津冀地区的地理特征和气象规律。

(3) 提出了一种基于 PageRank 算法改进的适用于空气污染物传播网络的 StationRank 算法。StationRank 算法不仅考虑了站点之间的权重，而且对 PageRank 算法中的转移概率做出了改进，充分考虑了邻居站点的重要性。最后通过对每个站点进行评分排序，最终得出排名靠前的站点即为污染物传播网络中的重要节点。并将站点排序结果与实际各站点的污染物超标率和入度强度进行对比，发现有很好的同步性。

本文虽然取得了一些研究成果，但是仍然存在值得改进的地方，因此下一步的工作将主要从以下方面进行研究：

(1) 由于本文只是研究了京津冀地区的污染物关键传输路径以及重要节点，没有对其他区域进行研究，因此，为了进一步验证本文提出的污染物传播模型的广泛性以及适应性，接下来将着重研究其他区域、特殊地形下的污染物传输规律，来进一步完善污染物传播模型。

(2) 大气污染是一个极其复杂的过程，不仅仅与污染源有关，大部分的来源污染物的二次污染，也就是污染物之间发生化学反应，再次生成大气污染物。因此，还需要从化学原理上对大气污染进行分析，才能更加准确的找到防治大气污染的有效方法。

参考文献

[1] Kampa M, Castanas E. Human health effects of air pollution[J]. Environmental Pollution, 2008, 151(2):362-367.

[2] 张志刚, 高庆先, 韩雪琴,等. 中国华北区域城市间污染物输送研究[J]. 环境科学研究, 2004, 17(1):14-20.

[3] Bove M C, Brotto P, Cassola F, et al. An integrated PM 2.5, source apportionment study: Positive Matrix Factorisation vs. the chemical transport model CAMx[J]. Atmospheric Environment, 2014, 94(94):274-286.

[4] Zheng Y, Liu F, Hsieh H P. U-Air: when urban air quality inference meets big data[C]. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2013:1436-1444.

[5] Zheng, Yu, Furui Liu, and Hsun-Ping Hsieh. U-Air: when urban air quality inference meets big data. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013:1436-1444.

[6] 徐祥德, 周丽, 周秀骥,等. 城市大气环境重污染过程周边源影响域[C]// 中国气象学会 2004 年年会. 2004, 34(10):958-966.

[7] Li J, Wang Z, Huang H, et al. Assessing the effects of trans-boundary aerosol transport between various city clusters on regional haze episodes in spring over East China[J]. Tellus Series B-chemical & Physical Meteorology, 2013, 65(1):60-73.

[8] 周曙东, 欧阳纬清, 葛继红. 京津冀 PM2.5 的主要影响因素及内在关系研究[J]. 中国人口·资源与环境, 2017, 27(4):102-109.

[9] M. C. Bove, et al., An integrated PM2.5 source apportionment study: Positive Matrix Factorisation vs. the chemical transport model CAMx, Atmospheric Environment, vol. 94, Sep 2014, pp. 274-286.

[10] 王扬锋, 左洪超, 马雁军,等. Models-3 在沈阳市空气质量的数值模拟研究. 中国气象学会年会. 2007, 35(10):2908-2916.

[11] 薛文博, 王金南, 付飞,等. CMAQ 与 CAMx 模型对中国 PM2.5 数值模拟对比分析[C]// 全国环境规划院. 2013, 34(6):1361-1368.

[12] 张稳定, 李杰, 向伟玲,等. 区域输送对河南郑州空气质量影响的数值模拟研究[J]. Advances in Environmental Protection, 2013, 03(2A):18-24.

[13] 吕炜, 李金凤, 王雪松,等. 长距离污染传输对珠江三角洲区域空气质量影响的数值模拟研究[J]. 环境科学学报, 2015, 35(1):30-41.

[14] 赵宏, 刘爱霞, 王恺,等. 基于 GA_ANN 改进的空气质量预测模型[J]. 环境科学研究, 2009, 22(11):1276-1281.

[15] Perez P. Combined model for PM10 forecasting in a large city[J]. Atmospheric Environment, 2012, 60(60):271-276.

[16] Russo A, Raischel F, Lind P G. Air quality prediction using optimal neural networks with stochastic variables[J]. Atmospheric Environment, 2013, 79(7):822-830.

[17] Pai T Y, Ho C L, Chen S W, et al. Using Seven Types of GM (1, 1) Model to Forecast Hourly Particulate Matter Concentration in Banciao City of Taiwan[J]. Water Air & Soil Pollution, 2011, 217(1-4):25-33.

[18] 丁卉, 刘永红, 曹生现. 基于模糊-灰色聚类方法的城市空气质量评价研究[J]. 环境科学与技术, 2013(s2):374-379.

[19] 司志娟, 孙宝盛, 李小芳. 基于改进型灰色神经网络组合模型的空气质量预测[J]. 环境工程学报, 2013, 7(9):3543-3547.

[20] Shen J, Zhong L, Fangfang H E, et al. Development of Air Quality Forecast Model Based on Clustering and Multiple Regression[J]. Environmental Science & Technology, 2015, 341:75-82.

[21] Chu, P.C., Chen, Y.C., Lu, S.H., 2008. Atmospheric effects on winter SO2 pollution in Lanzhou, China. Atmos. Res. 89, 356e373.

[22] Xu, W.Y., Zhao, C.S., Ran, L., Deng, Z.Z., Liu, P.E., Ma, N., Lin, W.L., Xu, X.B., Yan, R., He, X., Yu, J., Liang, W.D., Chen, L.L., 2011. Characteristics of pollutants and their correlation to meteorological conditions at a suburban site in the North China Plain. Atmos. Chem. Phys. 11, 4353-4369.

[23] Liu, Y., He, K.B., Li, S.S., Wang, Z.X., Christiani, D.C., Koutrakis, P., 2012. A statistical model to evaluate the effectiveness of PM2.5 emissions control during the Beijing 2008 Olympic Games. Environ. Int. 44, 100-105.

[24] Barmpadimos I, Hueglin C, Keller J, et al. Influence of meteorology on PM10 trends and variability in Switzerland from 1991 to 2008[J]. Atmospheric Chemistry & Physics & Discussions,

2011, 11(4):1813-1835.

[25] Maraziotis E A, Sarotis L L, Marazioti C E. "An analysis of inhalable (PM10) and fine particles (PM2.5) concentration levels in urban area of Patras, Greece",[J]. 2007, 52(6):959-966.

[26] Abdalmogith S S, Harrison R M. The use of trajectory cluster analysis to examine the long-range transport of secondary inorganic aerosol in the UK[J]. Atmospheric Environment, 2005, 39(35):6686-6695.

[27] Garcia Menendez F. High-resolution three-dimensional plume modeling with Eulerian atmospheric chemistry and transport models[J]. Georgia Institute of Technology, 2014, 329(24):1753-1759.

[28] Piñero-García F, Ferro-García M A, Chham E, et al. A cluster analysis of back trajectories to study the behaviour ofradioactive aerosols in the south-east of Spain[J]. Journal of Environmental Radioactivity, 2015, 147:142-152.

[29] Wang F, Chen D S, Cheng S Y, et al. Identification of regional atmospheric PM10, transport pathways using HYSPLIT, MM5-CMAQ and synoptic pressure pattern analysis[J]. Environmental Modelling & Software, 2010, 25(8):927-934.

[30] G. Y. Huang, et al., An algorithm to find critical execution paths of software based on complex network, International Journal of Modern Physics C, vol. 26, Sep 2015.

[31] Karagiannidis A, Poupkou A, Giannaros T, et al. The Air Quality of a Mediterranean Urban Environment Area and Its Relation to Major Meteorological Parameters[J]. Water Air & Soil Pollution, 2015, 226(1):2239.

[32] He J, Gong S, Yu Y, et al. Air pollution characteristics and their relation to meteorological conditions during 2014-2015 in major Chinese cities.[J]. Environmental Pollution, 2017, 223:484-496.

[33] Li L, Gong J, Zhou J. Spatial Interpolation of Fine Particulate Matter Concentrations Using the Shortest Wind-Field Path Distance[J]. Plos One, 2014, 9(5):e96111.

[34] 李佳霖, 樊子德, 邓敏. 顾及风向和风速的空气污染物浓度插值方法[J]. 地球信息科学学报, 2017, 19(3):382-389.

[35] Watts DJ, Strogatz SH. Collective dynamics of small-world' networks[J]. Nature, 1998, 393(4):440-442.

[36] Barabasi A-L,Albert R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439):

509-512.

[37] Gleich D F. PageRank beyond the Web[J]. Computer Science, 2014, 57(3).

[38] 李威凌, 吴怀宇, 陈洋. 基于高斯模型的武汉市区 PM2.5 扩散问题研究[J]. 高技术通讯, 2014, 24(11):1153-1159.

[39] 仇丽青, 陈卓艳. 基于 Pagerank 的社会网络关键节点发现算法[J]. 软件导刊, 2014(8):48-50.

[40] Sarma A D, Molla A R, Pandurangan G, et al. Fast distributed PageRank computation[J]. Theoretical Computer Science, 2015, 561:113-121.

[41] Arasu A, Cho J, Garcia-Molina H, et al. Searching the Web[J]. Acm Transactions on Internet Technology, 2002, 1(1):2-43.

[42] Bressan M, Peserico E, Pretto L. The power of local information in PageRank[C]. International Conference on World Wide Web Companion. 2016:179-180.

致 谢

值此论文完成之际，谨向给予我指导、关心和帮助的老师、朋友、同学及家人表示衷心的感谢。

深深感谢我的导师黄国言教授，黄老师严谨的治学态度、渊博的专业知识、执着追求的敬业精神和高尚的人格品质都使我受益匪浅。在攻读硕士学位期间，从课题的研究到论文的撰写都得到了导师全面、认真的指导。在学习遇到困难时，黄老师给予了耐心的解答和帮助，还提出了很多宝贵的意见和建议，使我渐渐的熟悉了研究内容。从导师那里我不仅学到了丰富的专业知识和科学的研究方法，更重要的是学到了一丝不苟的工作态度和谦和的为人处世原则。在此，谨向培养我的恩师致以最衷心的感谢和最诚挚的敬意。

感谢组里的郝晨谦、贾越洋、李鹏飞、田相敏、刘新倩、何洪豆、齐聪雅等同学，他们在学术研究和项目实践等诸多方面给与我了许多的支持和帮助。在大家的共同努力下，实验室不仅是共同科研、一起学习进步的场所，更是一个温暖有爱的大家庭，互帮互助，互相激励，我们在实验室度过的每一天都温馨而充实。

感谢燕山大学对我七年的培养，在这里，我从一个懵懂无知的高中毕业生，成长为乐一名遇事有担当的青年。在这里，我遇到了人生的良师，结交了人生的挚友，学习到了将来工作中所需的诸多专业技能，在燕山大学浓厚的学术氛围和优美的校园环境中积累了宝贵的知识和精神财富。

感谢我的家人，多年来对我无私帮助和鼎力支持，不管我遇到任何困难，都无条件的支持我、鼓励我、信任我。

最后，再次感谢所有关心和帮助过我的人们。