

一种基于 K-means 聚类算法的沙尘天气客观识别方法

段赛男¹, 焦瑞莉¹, 吴成来²

1 北京信息科技大学信息与通信工程学院, 北京 100101

2 中国科学院大气物理研究所国际气候与环境科学中心, 北京 100029

摘要 鉴于以往基于污染物浓度时间序列进行分析的沙尘天气识别方法在判断标准上存在一定的主观性, 本文提出一种基于 K-means 聚类算法的沙尘天气客观识别方法。本方法利用环境监测总站的 PM_{2.5} 和 PM₁₀ 小时浓度资料进行聚类, 首先选取最优的分类数目 K 进行聚类, 其次对聚类结果中离散程度较高的类别进行再次聚类, 直到无需分类。将本方法应用于西安市 2018 年 2~4 月沙尘天气的识别中, 结果表明, 本方法可有效识别主要沙尘天气。此外, 利用本方法可得到沙尘天气典型特征: PM_{2.5} 占 PM₁₀ 浓度的比例小于 43.5%、PM₁₀ 浓度高于 228 μg/m³, 符合沙尘天气期间 PM₁₀ 浓度较高且以粗颗粒物为主的物理特征。总体上看, 本方法物理基础清晰, 可操作性强, 适用于大规模数据处理, 具有较好的实用价值和应用前景。

关键词 沙尘天气识别; K-means 聚类; 客观识别; PM_{2.5}; PM₁₀

doi: 10.3878/j.issn.1006-9585.2023.23042

An Objective Identification Method for Dust Weather Based on the K-means Clustering Algorithm

DUAN Sainan¹, JIAO Ruili¹, WU Chenglai²

1 College of Information and Communication Engineering, Beijing

收稿日期 2023-05-04; 收修定稿 2023-12-04

作者简介 段赛男, 女, 2000 年出生, 硕士研究生, 主要从事数据挖掘方法在气象和环境领域的应用研究。

E-mail: sainanduan@163.com

通讯作者 吴成来, E-mail: wuchenglai@mail.iap.ac.cn

资助项目 国家重点研发计划项目 2020YFA0607801、国家自然科学基金项目 41975119、42075166、41830966, 北京信息科技大学横向项目 S2226080

Funded by National Key Research and Development Program of China (Grant 2020YFA0607801), National Natural Science Foundation of China (Grant 41975119, 42075166, 41830966), Beijing Information Science and Technology University (S2226080)

Information Science and Technology University, Beijing 100101

2 International Center for Climate and Environment Sciences, Institute of

Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029

Abstract Time series analysis methods have been developed before to identify dust weather based on pollutant concentrations, but the criteria used are subject to considerable uncertainty. Therefore, in this study, we propose an objective identification method for dust weather based on the K-means clustering method by using the hourly concentration of $PM_{2.5}$ and PM_{10} from the environmental monitoring stations. The flow path of this method is as follows: first select the optimal number of classifications K for cluster analysis; then classify the cluster groups that show large scattering in the distribution of $PM_{2.5}$ and PM_{10} concentrations until no further classification is needed. This method is applied to identify dust weather in Xi'an from February to April 2018. The results show that this method can effectively identify the main dust weather events. Based on this method, typical characteristics of dust weather can be obtained: the ratio of $PM_{2.5}$ to PM_{10} concentration is less than 43.5%, and the PM_{10} concentration is greater than $228\mu g/m^3$, which is consistent with physical characteristics that the PM_{10} concentration is high and mainly consists of coarse particles during the dust event. Overall, this method has a clear physical basis, and it is easy to operate, suitable for massive data processing, and promising for applications in relevant areas.

Keywords Identification of dust weather, K-means clustering, Objective identification, $PM_{2.5}$, PM_{10}

1 引言

沙尘天气指沙粒、尘土悬浮空中,使空气浑浊、能见度降低的一种天气现象,

是影响大气环境质量的灾害性天气。我国是沙尘天气多发的地区，尤其是北方地区，易受沙尘天气影响。沙尘天气不仅严重威胁生态环境质量、人体健康和电力设施，而且对当地生产、生活和交通都造成不同程度的损害（石广玉和赵思雄，2003；姜大海等，2011；石广玉等，2018）。影响我国的沙尘源区主要位于蒙古国南部的戈壁和我国西北、内蒙古、华北北部、东北西部等地区的沙漠和沙地（Shao and Dong, 2006；Wu et al., 2016；2018；2020；Wang et al., 2021）。沙尘从源地扬起后，输送到下游，可以影响到我国的华北、华中、华东，甚至到华南（吴成来和林朝辉，2014）。尽管近几十年来我国沙尘活动有减弱的趋势（周自江等，2002；Wu et al., 2018；2022），然而 2021 年 3 月两次强沙尘暴的发生使得沙尘暴的研究又受到重视（杨晓军等，2021；高泽田等，2022；张璐等，2022）。

沙尘天气的监测对于实时获取沙尘影响范围、研判其变化趋势并进行预警极为重要（陈思宇等，2023）。目前，沙尘天气的监测方法主要有三种，分别为卫星遥感、地面气象台站观测、地面环境监测站观测。卫星遥感利用沙尘与其他气溶胶粒子、云粒子、地表覆被等的光学特性的不同加以区分，能得到大范围区域的沙尘分布特征（张鹏等，2018；高泽田等，2022；张丽娟等，2022；邱昀等，2018；李彬等，2018）。地面气象台站分布较广，主要观测地面能见度和沙尘天气等级（周自江等，2002；钱正安等，2006；Wu et al., 2022），能够用于分析沙尘天气的演变和影响范围。尽管在沙漠及临近地区的站点分布不如靠近居民区或城市地区的站点密集，但气象台站资料具有较长的观测时间（从 20 世纪 50 年代开始），可用于研究沙尘的长期变化趋势研究（Wu et al., 2018）。这两种方法作为沙尘天气监测方法，已有广泛的应用，但是其观测量并不是直接的沙尘浓度，不利于分析沙尘污染的程度。

另外一种方法是通过观测大气中颗粒物的浓度对沙尘天气进行监测。2012 年中国环境监测总站开始构建中国区域的空气质量监测网，至今已建立了覆盖 300 多个城市、2000 多个站点的观测站网（<https://air.cnemc.cn:18007/>）。这些站点观测有细颗粒物 $PM_{2.5}$ 、可吸入颗粒物 PM_{10} 的浓度，能准确捕捉沙尘天气期间的大气颗粒物浓度，具有重要的应用价值，近年来已被广泛用于沙尘天气过程的分析（Yang et al., 2021；杨晓军等，2021；王蕾等，2022；张璐等，2022）。 $PM_{2.5}$ 和 PM_{10} 不仅包含沙尘粒子，而且包括大气中的其他颗粒物，如黑碳、有机碳、

硫酸盐、以及二次有机气溶胶等，因此需要首先依据 $PM_{2.5}$ 和 PM_{10} 的浓度特征对沙尘天气识别。沙尘天气期间，大气颗粒物（ $PM_{2.5}$ 、 PM_{10} ）浓度急剧增加，观测的 $PM_{2.5}$ 、 PM_{10} 浓度，特别是 PM_{10} 浓度，主要由沙尘贡献（Song et al., 2007; Wang et al., 2008; 李晓岚和张宏升, 2012; 吴成来和林朝辉, 2014），但是目前并没有针对沙尘天气识别的统一标准。相比于其他污染物，沙尘粒子的粒径更大，对 PM_{10} 浓度的贡献更显著，因而 PM_{10} 浓度往往作为识别沙尘天气最重要的特征因子。此外，综合 $PM_{2.5}$ 和 PM_{10} 浓度观测能提供污染物粒径分布的额外信息，为沙尘天气识别提供重要线索。因此，充分利用 $PM_{2.5}$ 和 PM_{10} 浓度进行沙尘天气识别也成为重要的研究内容（Tong et al., 2012; Fan et al., 2021）。

为识别沙尘天气，生态环境部于 2016 年发布《受沙尘天气过程影响城市空气质量评价补充规定》（环办监测〔2016〕120 号），提出了一套沙尘天气的识别方法（以下简称传统方法）。此后，生态环境部进一步于 2019 年发布《关于做好环境空气质量评价中扣除沙尘天气影响工作的函》（环测便函【2019】113 号），对该方法进行了补充和完善。传统方法对沙尘天气过程的判定标准如下：

（1）沙尘天气影响的起始时间为满足下列任一条件的整点小时：

（i）城市 PM_{10} 小时平均浓度大于等于前 6 个小时 PM_{10} 平均浓度的 2.0 倍且大于 150 微克/立方米；

（ii）城市 $PM_{2.5}$ 与 PM_{10} 小时浓度比值小于等于前 6 个小时比值平均值的 0.5 倍且 PM_{10} 小时浓度大于 150 微克/立方米。

（2）沙尘天气影响的结束时间为满足下列任一条件的整点小时：

（i）城市 PM_{10} 小时平均浓度首次降至与沙尘天气前 6 个小时 PM_{10} 平均浓度的 1.1 倍以下；

（ii）城市小时 $PM_{2.5}$ 空气质量指数（AQI）分指数首次超过小时 PM_{10} AQI 分指数。

上述传统方法对沙尘天气的识别具有较好的指导意义，已被用于沙尘天气监测业务和空气质量评价（Yang et al., 2021; 王蕾等, 2022）。尽管该方法具有较好的物理基础，且实践起来较方便，但也存在一定的局限性，其中 PM_{10} 浓度临界值（即 2 倍、150 微克/立方米、50%、10%等）、参考基准时间段（即沙尘天气前 6 个小时）的选取具有较大主观性。这些量值的选取缺少客观依据，而且可能

无法体现不同区域、不同季节大气污染特性的差异，其合理性也存疑。因此，在实际监测业务中，还需要根据卫星和地面气象台站等观测信息对传统方法识别的结果进行核对，经人工审核后确定为沙尘天气（以下简称业务方法）。由于审核人员经验的不一致性和其他人为因素干预，业务方法所识别的沙尘天气往往存在较大的主观性（郑湔等，2020）。鉴于此，特别是结合当前广泛应用的聚类分析等数据挖掘方法，开发新的沙尘天气识别方法。

郑湔等（2020）基于 K-means 发展了一种沙尘天气的快速识别方法，对中国环境监测总站观测的西安市 2018 年 2 月至 4 月逐小时 $PM_{2.5}$ 浓度、 PM_{10} 浓度、AQI 进行聚类分析，并将结果与业务方法的识别结果进行对比，发现 K-means 方法识别率较高，其识别的沙尘天气时次与业务方法差别小于 10%。郑湔等（2020）提出的 K-means 识别方法对沙尘识别技术研究具有指示意义，但也存在需要进一步解决的问题，如识别结果需要人工判断、第二次聚类的设置存在较大主观性等。本文进一步对该方法进行发展，提出一个更为通用的沙尘天气客观识别方法，并基于识别结果对沙尘天气的颗粒物浓度特征进行分析。

2 K-means 聚类算法

2.1 算法概述

聚类是数据分析的一种重要手段和工具，聚类过程指将数据对象的集合分成若干相似对象组成的簇的过程，目标是使同一簇中对象比较相似，不同簇间对象较为相异，在数据挖掘、机器学习、统计学等领域被广泛研究，此类研究主要集中在基于距离的聚类分析（孙吉贵等，2008；郑湔等，2020）。

K-means 聚类算法是由麦克奎因于 1967 年提出的一种迭代求解的划分式聚类算法，算法通常采用误差平方和函数作为优化的目标函数，是目前为止应用最广、最成熟的一种聚类分析算法，具有简单快速、适用于处理大规模数据的优点（徐大川等，2018；欧陈委，2011）。算法的输入通常为 K 值及样本集，通过该算法将样本进行聚类，具有相似特征的样本聚为一类，最后将使误差平方和达到最小的 K 个聚类输出（徐大川等，2018）。

2.2 算法流程

K-means 聚类算法的基本思想是从数据集中随机选择 K 个数据对象作为初始中心，计算每个数据对象到各中心的距离，根据最近邻原则，所有数据对象将会被划分到离它最近的那个中心所代表的簇中，分别计算新生成的各簇中数据对象的均值并更新为各簇新的中心，比较相邻迭代次数内的中心，如果变化量未超过规定的阈值，则算法收敛，输出结果；反之，根据新的中心对所有数据对象重新进行划分，直到满足算法的收敛条件为止。K-means 聚类算法的主要工作流程如图 1 表示。K-means 可采用的距离指标包括欧式距离、曼哈顿距离、夹角余弦等，不同指标可适用于不同的应用场景，本文采用欧式距离作为距离指标，因为本文聚类分析所用的 PM_{10} 浓度和 $PM_{2.5}$ 浓度具有相同的量纲。

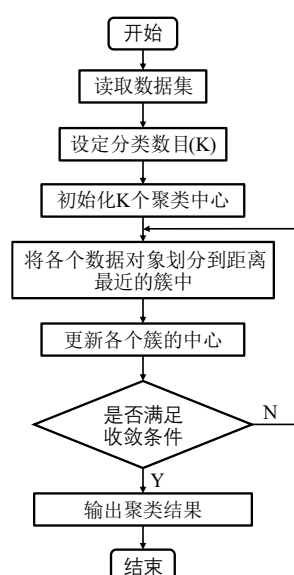


图 1 K-means 聚类算法流程

Fig.1 The flow chart of the K-means clustering algorithm

3 K-means 聚类算法识别沙尘天气

3.1 数据介绍

本文采用的观测数据为中国环境监测总站观测的西安市 2018 年 2 月至 4 月逐小时空气质量数据。虽然中国环境监测总站还提供了空气污染指数 (AQI)，但由于其已经部分考虑了 PM_{10} 小时质量浓度和 $PM_{2.5}$ 小时质量浓度的观测信息，为避免重复，本文不采用 AQI 资料。对于 $PM_{2.5}$ 浓度高于 PM_{10} 浓度的情况，本文视为异常数据，对于观测数据中的异常部分及缺失部分，本文不予考虑。总体

来看, 2018 年 2 月至 4 月共有 2119 组数据用于分析, 而其余异常数据共 17 组, 占有数据量 2136 组的 0.8%。为方便记录, 每组数据以 ($\text{PM}_{2.5}$ 浓度, PM_{10} 浓度) 形式统计。K-means 聚类结果易受到极端值的影响, 但本文所用资料中 $\text{PM}_{2.5}$ 、 PM_{10} 浓度最高分别为 $326\mu\text{g}/\text{m}^3$ 、 $852\mu\text{g}/\text{m}^3$, 量级与通常污染期间的浓度量级相当, 因此本文不对极端值做特别处理。对于其他区域或其他时段的污染物浓度观测中, 沙尘天气期间的 PM_{10} 浓度值有时可达数千微克每立方米, 对此, 可将 PM_{10} 浓度大于 1000 微克每立方米的时段判定为沙尘天气影响时段, 再对剩余数据再进行聚类分析。

3.2 不同分类数目下的聚类结果对比

为找到分类数目的最优值, 本文分别对不同 K 值进行了多次试验, 其中初始聚类中心采用随机选取的方法。K 值分别取 2、3、4、5, 各进行 100 次试验, 记录每次试验的初始聚类中心、最终聚类中心、各聚类包含的数据量, 统计不同 K 值条件下各个聚类结果的差异程度。对于 $K>5$ 的情况, 其所设置类别较多, 不利于对识别结果进行判断, 因此本文选择 $K\leq 5$ 的情况进行对比和分析。计算最终聚类中心的均值、标准差, 将两者的比值作为其变异系数, 用于表示 100 组试验最终聚类中心的差异程度, 如果差异程度高, 说明聚类结果稳定性低; 反之则稳定性高。

分析聚类结果中各个类别的数据特征, 判断其是否与沙尘天气期间的颗粒物浓度的物理特征一致, 即可吸入颗粒物浓度较高且以粗颗粒物为主, 本文中粗颗粒物为空气动力学直径为 $2.5\sim 10\mu\text{m}$ 的粒子, 因此 PM_{10} 浓度与 $\text{PM}_{2.5}$ 浓度相减即为粗颗粒物浓度。 $\text{PM}_{2.5}$ 浓度与 PM_{10} 浓度的比值越小, 代表粗颗粒物的比重越大。当该比值小于 0.5, 即对应 PM_{10} 与 $\text{PM}_{2.5}$ 浓度比值大于 2 时, PM_{10} 质量以粗颗粒物为主; 反之, PM_{10} 质量则以 $\text{PM}_{2.5}$ 为主。相比于沙尘中的其他污染物, 粗颗粒物含量较大, 因此可将较小的 $\text{PM}_{2.5}$ 与 PM_{10} 浓度比值作为沙尘天气的判定条件。

对于 $K=4$ 下沙尘天气识别的完整过程, 进行详细的阐述, 而其他 K 值下的识别过程与之类似。不同 K 值下 100 次试验的结果总结见表 1。可以看出:

(1) $K=3$ 和 $K=5$ 下最终聚类中心的差异程度较高, 其对应的 PM_{10} 浓度和 $\text{PM}_{2.5}$ 浓度变异系数最大超过 10%, 说明聚类结果相对不稳定;

(2) K=4 下最终聚类中心的差异程度相对较低，其中变异系数多数低于 2%，只有第四类的变异系数达到 5-6%，说明聚类结果较稳定；

(3) K=2 时，最终聚类中心的差异程度最小，其中变异系数均低于 1%，说明聚类结果较稳定。

对比沙尘天气的识别效果，K=2 时没有识别出沙尘天气，K=4 时可识别出沙尘天气，K=3 时部分情况下可识别出沙尘天气，K=5 时得到 2 个不同聚类结果，分别有 1 个或 2 个类别可以认为是沙尘天气。综合考虑聚类结果差异程度及沙尘天气识别效果两个因素，K=4 时可以达到最好的沙尘天气识别效果。以下将继续对 K=4 时识别出来的沙尘天气进行分析。

表 1 不同 K 值条件下 100 次试验聚类结果的差异程度及沙尘天气识别效果
Table 1 The degree of difference in clustering results from 100 experiments for each K-value conditions and the efficiency for identifying dust weather

	100 次试验所得到的最终聚类中心 的变异系数	最终聚类中心 差异程度	沙尘天气识别情况	沙尘天气 识别效果
K=2	(0, 0.33%) (0.28%, 0.14%)	较低	未识别出沙尘天气数据	较差
K=3	(5.79%, 6.68%) (11.36%, 7.21%) (10.62%, 14.36%)	较高	部分情况下可识别出沙尘天气，但这种结果出现的概率，为 20%，即 100 组中出现 20 组	较差
K=4	(1.08%, 1.24%) (1.43%, 0.91%) (0.59%, 2.11%) (5.76%, 6.56%)	较低	可识别出沙尘天气 结果：聚类 4 为沙尘天气	较好
K=5	(2.32%, 2.92%) (3.54%, 3.96%) (16.38%, 4.27%) (13.33%, 6.31%) (2.10%, 9.64%)	较高	可识别出沙尘天气 结果 1（95%概率）：聚类 5 为沙尘天气 结果 2（5%概率）：聚类 4、聚类 5 均为沙尘天气	较好

- a. 对应于每一组数据（PM_{2.5} 浓度，PM₁₀ 浓度），表格中数据分别为其对应聚类中心的变异系数，以%记录。
- b. 根据沙尘天气的典型特征，即 PM₁₀ 浓度较大且 PM_{2.5} 浓度与 PM₁₀ 浓度比值较小，判断是否为沙尘天气，如 K=4 时第三个类别的最终聚类中心为（98，484），相比其他三个类别，PM₁₀ 浓度最大且 PM_{2.5} 与 PM₁₀ 浓度的比值最小，判断该类别为沙尘天气。

3.3 K=4 时聚类结果分析

表 2 为 K=4 时的各个类别的聚类结果，图 2 为各个聚类中 PM_{2.5} 和 PM₁₀ 浓度的分布情况。K=4 时，4 个最终聚类中心分布在（36，74）、（73，155）、（98，484）、（151，247）附近。聚类分析的思想是将距离较近的点划分为同一类，则

同一类别的点将分布在（PM_{2.5} 浓度，PM₁₀ 浓度）坐标中距离相近的位置，即这四个类别将 PM_{2.5} 和 PM₁₀ 浓度的分布划分为 4 个区域（图 2）。根据各个聚类 PM_{2.5} 浓度、PM₁₀ 浓度与两者比值的范围，以及污染天气的典型特征，即污染天气存在较高的 PM_{2.5} 浓度或 PM₁₀ 浓度、沙尘天气期间 PM₁₀ 浓度较大且 PM₁₀ 浓度与 PM_{2.5} 浓度比值较大，四个聚类的数据类型可分别对应无污染型、雾霾污染型、沙尘污染型、混合型污染（注：此处所说“混合型”指数据组本身含有沙尘和雾霾数据，而非沙尘和雾霾污染物的混合）：

聚类 1：PM_{2.5} 浓度低于 93μg/m³、PM₁₀ 浓度在 126μg/m³ 以下，空气质量较好，没有明显的雾霾污染或沙尘污染，可认为是无污染型；

聚类 2：PM_{2.5} 和 PM₁₀ 浓度范围分别为 13~141μg/m³、98~266μg/m³，而 PM₁₀ 浓度与 PM_{2.5} 浓度比值变化较大，在 1.01~14.15 之间，可见污染程度较轻或以雾霾污染为主（PM_{2.5} 质量比重大），属于雾霾污染或轻度污染型；

聚类 3：PM_{2.5} 浓度高于 34μg/m³、PM₁₀ 浓度在 356μg/m³ 以上，PM₁₀ 浓度与 PM_{2.5} 浓度比值超过 2.73，可见粗颗粒物质量比重超过 PM_{2.5}，符合沙尘污染特征；

聚类 4：PM_{2.5} 浓度、PM₁₀ 浓度以及两者比值的离散程度均较大，数据点兼具雾霾污染和沙尘污染的特征，难以直接认为是雾霾或者沙尘型，因此可认为是混合型。该类数据仍有部分数据符合沙尘天气的数据特点，以下将在第一次聚类的基础上进行二次聚类。

表 2 K=4 时西安市 2 至 4 月 PM_{2.5}、PM₁₀ 小时质量浓度第一次聚类结果。该结果为 K=4 时 100 次试验里第 1 次试验的结果，其他试验与该结果略有差别，其差别特征可见表 1

Table 2 The clustering results for PM_{2.5} and PM₁₀ hourly mass concentrations from February to April in Xi'an at K=4. This result is derived from the first experiment of the 100 experiments at K=4. Note that there is slight difference in the results of other experiments and the statistics of the difference can be seen in Table 1

	聚类 1	聚类 2	聚类 3	聚类 4
最终聚类中心(μg/m ³)	(36, 74)	(73, 155)	(98, 484)	(151, 247)
PM _{2.5} 浓度范围(μg/m ³)	[8, 93]	[13, 141]	[34, 151]	[30, 326]
PM ₁₀ 浓度范围(μg/m ³)	[12, 126]	[98, 266]	[356, 852]	[172, 413]
PM ₁₀ 与 PM _{2.5} 浓度比值范围	[1, 7.92]	[1.01, 14.15]	[2.73, 14.09]	[1.01, 9.17]
数据组数	732	937	59	391
描述	无污染型	雾霾污染或轻度污染型	沙尘污染型	混合污染型

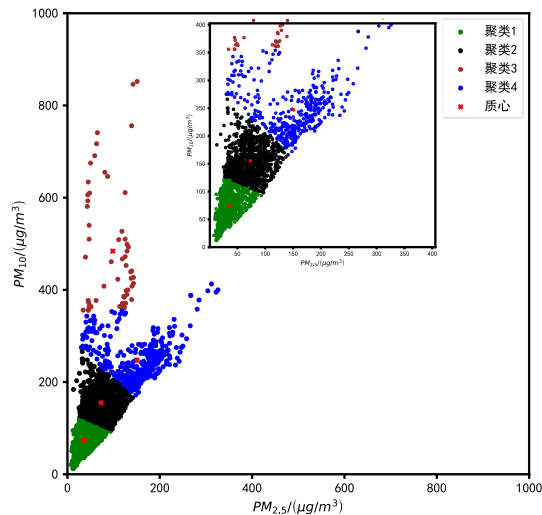


图 2 K=4 时 $PM_{2.5}$ 、 PM_{10} 小时质量浓度聚类结果（为了更清楚展示，图中小图坐标轴为 $0-400\mu g/m^3$ ）

Fig.2 The clustering results for $PM_{2.5}$ and PM_{10} hourly mass concentrations at $K=4$. For a clear display, the concentration is shown in the range of $0-400\mu g/m^3$ in the plot embedded.

进一步对聚类 4（混合型污染数据）进行分类。为找到第二次聚类时分类数目（ K_2 ）的最优值，对 K_2 值分别取 2、3、4，各进行 100 次试验，记录每次试验的初始聚类中心、最终聚类中心、各聚类包含的数据组数，统计不同 K_2 值条件下各个试验最终聚类中心的变异系数。比较不同 K_2 值条件下的聚类结果稳定性和沙尘数据识别效果，具体情况如表 3 所示。根据表 3，可以看出：

(1) $K_2=2$ 或 4 时，最终聚类中心对应的 PM_{10} 浓度和 $PM_{2.5}$ 浓度变异系数可接近于 40%，不同试验下聚类结果差异程度高；

(2) $K_2=3$ 时聚类结果较稳定，最终聚类中心对应的 PM_{10} 浓度和 $PM_{2.5}$ 浓度变异系数可接近于 4%，此时最终聚类中心在（146，214）、（203，273）、（71，289）附近，较好地将聚类 4 中 $PM_{2.5}$ 和 PM_{10} 浓度的分布划分为 3 个区域（图 3）。

类似于上述 $K=4$ 时第一次聚类结果的分析，可以判定第二次聚类得到的第三个类别（即聚类 4-3）较符合沙尘天气的典型特征，该类别可以认为是沙尘天气，另 2 个类别（即聚类 4-1、聚类 4-2）分别为雾霾污染型、雾霾污染型（见表 4）。

表 3 K=4 下对聚类 4（混合型污染）进行第二次聚类时，不同 K_2 值条件下各个聚类结果的聚类中心稳定性及沙尘天气识别效果

Table 3 Variations in the cluster centers and efficiency in the identification of dust weather for each K_2 value,

derived from the second classification on the cluster 4 (“mixed” pollution type, $K=4$)

100 次试验所得到的最终聚类中 心的变异系数	最终聚类中 心差异程度	沙尘天气识别情况	沙尘天气 识别效果
----------------------------	----------------	----------	--------------

$K_2=2$	(18.95%, 0.79%)(42.43%, 6.56%)	较高	结果 1: 未识别出沙尘天气 结果 2: 聚类 4-2 为沙尘天气	较差
$K_2=3$	(2.04%, 1.16%)(1.25%, 1.23%) (3.41%, 3.25%)	较低	结果: 聚类 4-3 为沙尘天气	较好
$K_2=4$	(5.34%, 1.14%)(25.97%, 4.03%) (42.93%, 5.06%)(60.10%, 5.55%)	较高	结果 1: 聚类 4-3 为沙尘天气 结果 2: 聚类 4-4 为沙尘天气 结果 3: 聚类 4-2 和聚类 4-4 为沙尘天气	较差
...				

表 4 聚类 4 进行 $K_2=3$ 第二次聚类分析后 $PM_{2.5}$ 、 PM_{10} 小时质量浓度聚类结果

Table 4 Cluster results for the $PM_{2.5}$ and PM_{10} hourly mass concentrations, derived from the second classification ($K_2=3$) on cluster 4 after the first classification at $K=4$

	聚类 4-1	聚类 4-2	聚类 4-3
最终聚类中心($\mu g/m^3$)	(146, 214)	(203, 273)	(71, 289)
$PM_{2.5}$ 浓度范围($\mu g/m^3$)	[90, 195]	[140, 236]	[30, 126]
PM_{10} 浓度范围($\mu g/m^3$)	[172, 270]	[222, 413]	[228, 354]
PM_{10} 与 $PM_{2.5}$ 浓度比值范围	[1.11, 2.58]	[1.01, 1.99]	[2.3, 9.17]
数据组数	193	127	71
描述	雾霾污染型	雾霾污染型	沙尘污染型

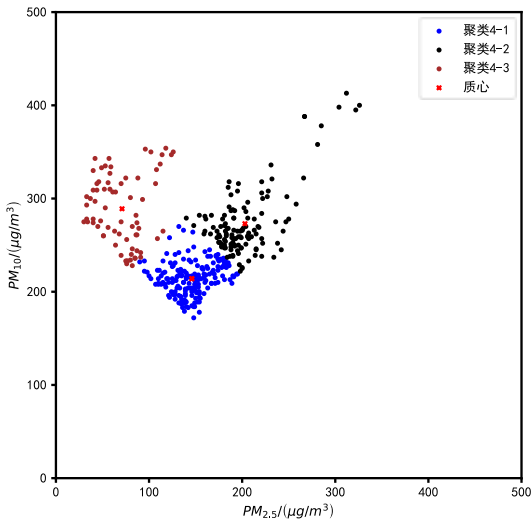


图 3 $K=4$ 下对聚类 4 进行第二次聚类后的 $PM_{2.5}$ 、 PM_{10} 小时质量浓度分布，其中 K_2 取 3

Fig.3 Distribution of $PM_{2.5}$ and PM_{10} hourly mass concentration for the second classification ($K_2=3$) on cluster 4 after the first classification at $K=4$

4 沙尘天气特征分析

4.1 沙尘天气发生时段

本节将对第 3 节 $K=4$ 下所识别出的沙尘天气进行分析，结合聚类 3 和聚类 4-3 的结果，进而得到所有的沙尘天气时次。总体来看，在 2018 年 2 月至 4 月西

安市环境监测数据中，被识别为沙尘天气的数据量有 130 组，共有受沙尘天气影响的时段 15 个，沙尘影响天数为 15 天。具体的沙尘天气影响时段如表 5 所示。图 4 为西安市 2018 年 2 月至 4 月逐小时 PM₁₀ 和 PM_{2.5} 质量浓度的时间序列图，图像中阴影部分表示该观测点受沙尘影响的时段。可以看出，沙尘天气期间，PM₁₀ 和 PM_{2.5} 浓度快速增加，达到峰值后又快速降低，沙尘过程持续的时长范围为 1~42 小时，变化较大。区别于通常从沙尘气团演变的角度所定义的沙尘天气过程，本文中的沙尘天气影响时段表示的是受到明显沙尘影响的时期，而在同一个沙尘天气过程中，可能会有多个影响时段。事实上，上述这些沙尘天气影响时段中，有一些时段之间的间隔较短，如表 5 中第 7 个和第 8 个时段、第 9 个和第 10 个时段、第 11 个和第 12 个时段之间的间隔，在 3 个小时以内。这些彼此间隔较短的沙尘天气影响时段，极有可能是受到同一个沙尘天气过程的影响，在沙尘沉降和外来输送的共同作用下，颗粒物浓度出现了短暂的低值时段，后又有所上升。

表 5 西安市 2 月至 4 月受沙尘天气影响时段
Table 5 The period from February to April in Xi'an affected by sandstorm weather

序号	开始时间	结束时间	持续时间/h	影响天数/d
1	2 月 9 日 14:00	2 月 10 日 06:00	17	2
2	3 月 15 日 20:00	3 月 16 日 05:00	10	2
3	3 月 16 日 10:00	3 月 16 日 14:00	5	
4	3 月 21 日 21:00	3 月 21 日 22:00	2	1
5	3 月 29 日 06:00		1	1
6	3 月 29 日 13:00	3 月 29 日 17:00	5	
7	4 月 1 日 22:00		1	1
8	4 月 2 日 01:00		1	1
9	4 月 2 日 06:00	4 月 2 日 07:00	2	
10	4 月 2 日 09:00	4 月 2 日 10:00	2	
11	4 月 3 日 00:00	4 月 3 日 05:00	6	2
12	4 月 3 日 07:00		1	
13	4 月 3 日 22:00	4 月 4 日 18:00	21	
14	4 月 5 日 21:00	4 月 6 日 10:00	14	2
15	4 月 10 日 10:00	4 月 12 日 03:00	42	3
合计	/	/	130	15

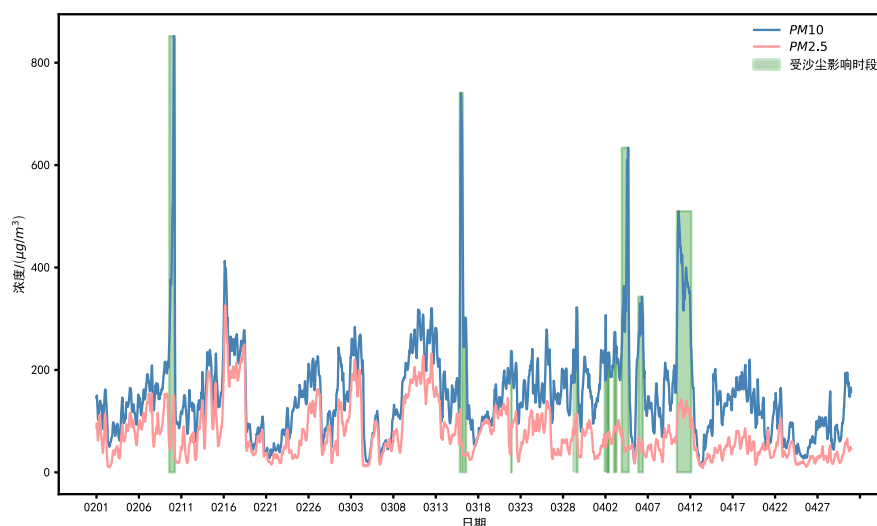


图4 西安市 2018 年 2 至 4 月大气污染物浓度时间序列图（沙尘天气具体起止时间见表 5）

Fig.4 Time series of atmospheric pollutants concentration in Xi'an from February to April 2018. The starting and ending time for each dust event is listed in Table 5

4.2 沙尘天气期间颗粒物浓度的分布特征

进一步分析沙尘天气下 PM_{10} 浓度、 $PM_{2.5}$ 浓度以及两者比值的统计特征。综合表 2 和表 4，可以得到：沙尘天气期间 PM_{10} 浓度大于 $228\mu g/m^3$ 、 PM_{10} 浓度与 $PM_{2.5}$ 浓度比值大于 2.30，也即 $PM_{2.5}$ 质量占 PM_{10} 比重小于 43.5%。通过图 5 中上述两个条件所对应直线的分布，可以看出，利用上述两个条件也可以将聚类分析后得到的沙尘天气数据较好地区分出来。气象台站通常根据水平能见度对沙尘天气进行判别，采用以能见度小于 10km 为沙尘天气发生的主要判据，根据 Song et al. (2007) 观测拟合得到的 PM_{10} 浓度和能见度关系可知，10km 能见度对应的 PM_{10} 浓度为 $205\mu g/m^3$ 。本文的结果 ($228\mu g/m^3$) 与 $205\mu g/m^3$ 大体相当。此外，43.5%作为上限，也能体现沙尘天气期间粗颗粒质量占比较大的特征。在以往研究中，部分学者将 $PM_{2.5}$ 占 PM_{10} 浓度的比重作为沙尘天气判断标准，Tong et al. (2012)将 $PM_{2.5}$ 占 PM_{10} 浓度的比值低于 39%作为美国沙尘天气的判别条件之一；Fan et al. (2021)采用 $PM_{2.5}$ 占 PM_{10} 浓度的比值低于 40%作为判断我国沙尘天气发生的标准。本文所给出的上限值 43.5%与这些结果比较接近。

从图 5 可以看出，如果以聚类后的结果作为参考，上述两个条件所区分出的数据组中有沙尘污染型以及轻度污染或雾霾污染型两种类型，说明这两个条件存在对沙尘天气的“误判”，误判的数据组数为 16，占聚类分析所得结果 (130) 的 12.31%。改变以上两个条件对应的临界值（即 $228\mu g/m^3$ 、2.30），从图 6 可以

看出，当 PM_{10} 浓度大于 $270\mu g/m^3$ 、 PM_{10} 浓度与 $PM_{2.5}$ 浓度比值大于 2.58（ $PM_{2.5}$ 质量占 PM_{10} 比重小于 38.8%），其区分出的数据组均为沙尘污染型数据，但无法区分出聚类结果中所有的沙尘数据，遗漏的沙尘天气数据组数为 25，占聚类分析识别总数（130）的 19.23%。

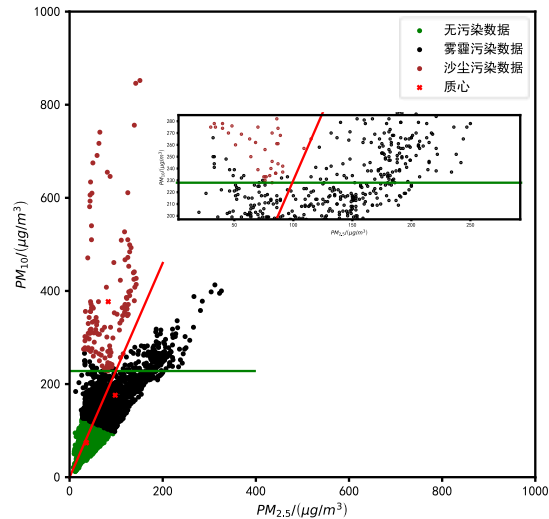


图 5 综合两次聚类结果识别的沙尘天气 $PM_{2.5}$ 、 PM_{10} 小时浓度分布。图中直线分别为 $y=2.30x$ 、 $y=228$ ，用于表示沙尘天气期间的 $PM_{2.5}$ 、 PM_{10} 浓度临界值分布特征

Fig.5 Distribution of hourly $PM_{2.5}$ and PM_{10} concentration in dust weather identified by combining the first and second clustering results. The straight lines in the figure denote $y=2.30x$ and $y=228$, used to represent the distribution characteristics of the critical values of $PM_{2.5}$ and PM_{10} concentrations during dust weather.

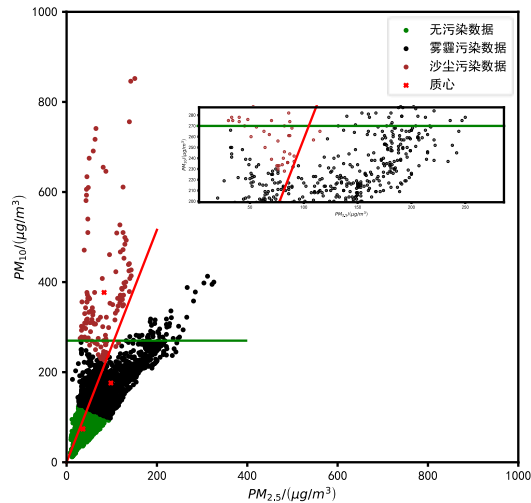


图 6 综合两次聚类后直线 $y=2.58x$ （红实线）与直线 $y=270$ （绿实线）对沙尘污染型数据的区分效果

Fig.6 The efficiency of straight line $y=2.58x$ (red solid line) and $y=270$ (green solid line) in distinguishing dust pollution data after combining the first and second clustering results

尝试将 PM_{10} 浓度在 $228\sim270\mu g/m^3$ 之间、 PM_{10} 浓度与 $PM_{2.5}$ 浓度比值在 2.30~2.58 之间找个“最优值”，使得上述“误判”和“遗漏”的比例相比于聚类分

析的结果最小。PM₁₀ 浓度临界值在 228~270μg/m³ 区间内，以 1μg/m³ 的间隔逐步增加，即分别取 228、229、230、...、270，PM₁₀ 浓度与 PM_{2.5} 浓度比值临界值在 2.30~2.58 间以 0.01 逐步增加，即分别取 2.30、2.31、2.32、...、2.58，计算沙尘天气的区分效果，可以得到，当上述误判和遗漏的比例最小时，对应的区分值分别为 233μg/m³、2.30(区分效果见图 7)，此时遗漏和误判的数据组数量占聚类分析所识别的沙尘型污染数据组数量的比例为 10%。表 6 列出了不同判断值下对应的分析结果。

表 6 不同 PM₁₀ 浓度、PM₁₀ 浓度与 PM_{2.5} 浓度比值的临界值对沙尘型污染天气的区分效果（以聚类分析识别结果 N=130 为参考）

Table 6 The efficiency of the critical values for PM₁₀ concentration and ratio of PM₁₀-concentration-to-PM_{2.5} concentration in distinguishing dust events (comparison with N=130 from the objective clustering analysis)

判断依据		识别的沙尘型污染数据量(Na)/组	遗漏的沙尘型污染数据量(Nb)/组	误判的沙尘型污染数据量(Nc)/组	遗漏和误判的沙尘型污染数据所占比例($\frac{Nb+Nc}{N}$)
PM ₁₀ 浓度	PM ₁₀ 与 PM _{2.5} 浓度比值				
≥228μg/m ³	≥2.30	146	0	16	12.31%
≥270μg/m ³	≥2.58	105	25	0	19.23%
≥233μg/m ³	≥2.30	139	2	11	10%

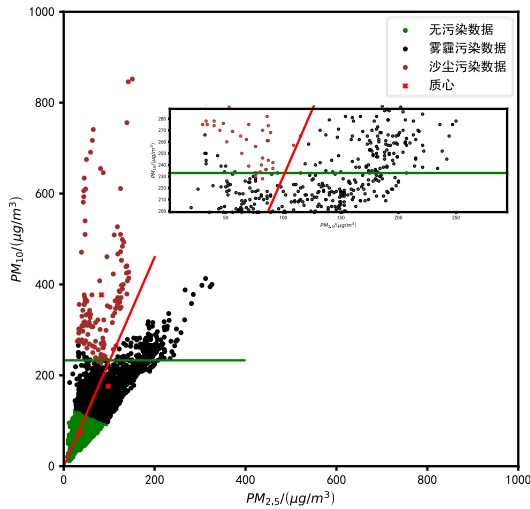


图 7 综合两次聚类后直线 $y=2.30x$ （红实线）与直线 $y=233$ （绿实线）对沙尘污染型数据的区分效果

Fig.7 The efficiency of straight line $y=2.30x$ (red solid line) and $y=233$ (green solid line) in distinguishing dust pollution data after combining the first and second clustering results

4.3 本文方法与以往方法对比

如前文所述，当前生态环境部根据沙尘天气的特点已经提出了沙尘天气的客观识别方法即传统方法。在此基础上，通常业务上还需参考多源资料并经人工审核后，最终给出沙尘天气的识别结果，即为业务方法。鉴于人工审核中存在的一些主观性，郑湔等（2020）基于 K-means 方法提出了一套沙尘天气的客观识别方

法（以下简称 Z2020）。这里将本文的结果与上述方法的结果进行对比，其中业务方法结果取自于郑湔等（2020）给出的结果（见其文中的表 1）。从图 8 可以看出，业务方法给出的 5 个沙尘影响时段均包含 PM_{10} 浓度快速增加且 PM_{10} 浓度远高于 $PM_{2.5}$ 浓度的特征，因而识别结果比较符合沙尘天气的物理特征，结果较准确。以下将以业务方法的结果为基准，分析几个方法的识别效果及优缺点。

首先，不同方法识别的沙尘天气持续时间的总体分布存在明显差别（表 5、表 7、表 8）。业务方法识别出 5 个沙尘影响时段，持续时间为 14~35 小时，变化较小（表 8）。这说明业务方法参考了卫星及地面气象台站等多源资料并经人工审核后，侧重于筛选出持续较长时间的沙尘天气过程，确保了较可靠的识别效果。与业务方法对比，传统方法、Z2020、本文识别的沙尘天气持续时间变化分别为 2~121 小时、1~40 小时、1~42 小时，可以看出这三个方法除了识别持续时间较长的沙尘天气过程，也会识别出持续时间较短的过程。

表 7 基于传统方法识别的西安市 2 月至 4 月受沙尘天气影响时段

Table 7 The period from February to April in Xi'an affected by sandstorm weather obtained by the traditional method

序号	开始时间	结束时间	持续时间/h	影响天数/d
1	2 月 9 日 13:00	2 月 10 日 07:00	19	2
2	3 月 4 日 08:00	3 月 4 日 11:00	4	1
3	3 月 15 日 21:00	3 月 16 日 17:00	21	2
4	3 月 21 日 15:00	3 月 21 日 18:00	4	1
5	4 月 5 日 15:00	4 月 8 日 17:00	75	4
6	4 月 14 日 17:00	4 月 19 日 17:00	121	6
合计	/	/	244	16

表 8 基于业务方法识别的西安市 2 月至 4 月受沙尘天气影响时段

Table 8 The period from February to April in Xi'an affected by sandstorm weather obtained by the operational

序号	开始时间	结束时间	持续时间/h	影响天数/d
1	2 月 9 日 13:00	2 月 10 日 07:00	19	2
2	3 月 15 日 21:00	3 月 16 日 17:00	21	2
3	4 月 4 日 05:00	4 月 4 日 18:00	14	1
4	4 月 5 日 15:00	4 月 6 日 16:00	26	2
5	4 月 10 日 10:00	4 月 11 日 20:00	35	2
合计	/	/	115	9

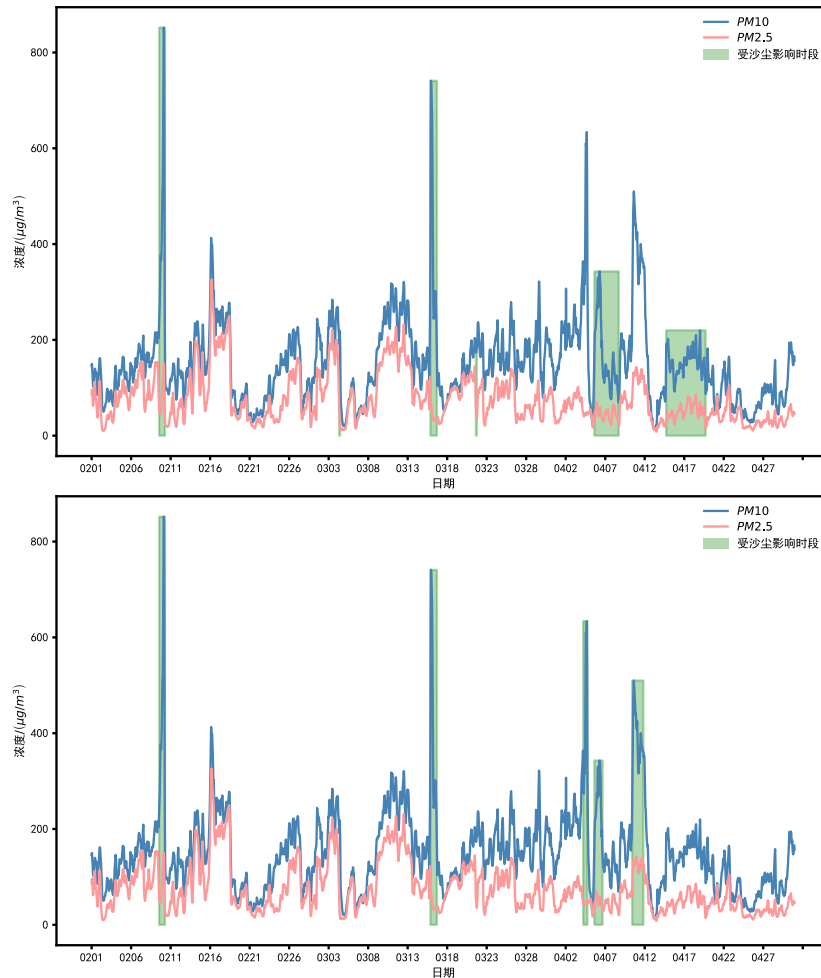


图 8 同图 4，上图、下图中阴影分别为传统和业务方法识别的沙尘天气影响时段（沙尘天气具体起止时间分别见表 7、表 8）
Fig. 8 As Fig. 4, with the shaded color indicating the dust events obtained from the traditional method (top) and the operational method (bottom), respectively. The starting and ending time for each dust event is listed in Table 7 and 8, respectively.

其次，传统方法对主要的沙尘天气存在漏判和误判。传统方法识别出了业务方法所给出 5 个沙尘影响时段中的 3 个，但未识别出 4 月 4 日和 4 月 10~11 日两次明显的沙尘影响时段（表 7、表 8），这主要是由于这 2 个过程中， PM_{10} 浓度增加、 $PM_{2.5}$ 浓度与 PM_{10} 浓度的比值减少均不如其他 3 个过程快速，不符合传统方法中沙尘天气起始时间的判定条件。对于 4 月 5~6 日沙尘天气过程，传统方法所识别的沙尘天气持续时间明显较长，这主要是由于沙尘天气开始前 6 个小时 PM_{10} 平均浓度仅为 $63\mu g/m^3$ ，相对较低，导致判定沙尘天气结束时间的浓度值也较低。另外，相比于业务方法，传统方法将 4 月 14~19 日弱的污染过程判定为沙尘天气。事实上，该过程中 PM_{10} 浓度为 $87\sim 220\mu g/m^3$ 。对于西安，该浓度范围在 2018 年 2~4 月较为常见，较难确定该时段是否发生沙尘天气。这也是业务中人

工审核阶段不考虑该时段作为沙尘天气影响时段的原因。

再次, 本文和 Z2020 均能较好识别出主要的沙尘天气过程。相比于业务方法, 本文和 Z2020 均识别出了主要的沙尘天气过程, 但沙尘时段的起始和结束时间存在差别, 说明沙尘天气的识别方法在判定其起始和结束时间的判定中存在不确定性。Z2020 识别结果中除了 2018 年 3 月 16 日 06:00 时的沙尘影响时刻未在本文识别结果中外, 其余 110 组沙尘影响时刻均在本文识别结果中, 并且本文所识别的沙尘影响时刻总数量比 Z2020 所识别的多 19 组。相应地, 本文所得到的沙尘天气期间 PM_{10} 浓度值、 PM_{10} 浓度与 $\text{PM}_{2.5}$ 浓度比值均低于 Z2020 的所得值。

表 9 为不同方法识别的沙尘天气影响时段和颗粒物浓度统计特征。对于主要沙尘天气过程的识别效果, 本文和 Z2020 相当, 均优于传统方法。从沙尘天气的数据特征来看, 传统方法所识别沙尘天气的 PM_{10} 浓度的下限、 $\text{PM}_{2.5}$ 浓度与 PM_{10} 浓度比值的下限均低于其他方法的下限值, 其下限值分别为 $69\mu\text{g}/\text{m}^3$ 、2.04, 这和传统方法主要依据 PM_{10} 浓度、 $\text{PM}_{2.5}$ 浓度与 PM_{10} 浓度比值与时间之间的变化速率进行判定的标准有关。业务方法的两个下限值分别为 $133\mu\text{g}/\text{m}^3$ 、2.73, 相较于传统方法的下限值, 两值均有所提高, 本文所得到的 PM_{10} 浓度的下限值为 $228\mu\text{g}/\text{m}^3$, 高于业务方法的所得值, $\text{PM}_{2.5}$ 浓度与 PM_{10} 浓度比值的下限为 2.30, 低于业务方法的所得值, 数据仍符合粗颗粒质量占比较大的特征。Z2020 的两个下限值分别为 $260\mu\text{g}/\text{m}^3$ 、2.73, 相较于上述三种方法的下限值, 该值均为最高, 其所识别的主要沙尘天气过程的持续时间最短。

此外, 表 9 总结了各个方法在操作流程、效率、主观性影响等方面的优缺点。传统方法具有清晰的物理基础, 对沙尘天气的识别有着指导意义, 但标准中的颗粒物浓度临界指标及参考基准时段的选取具有较大主观性, 导致识别结果异常, 如漏判沙尘天气或识别的沙尘天气持续时间偏长, 方法适用性存在明显不足。业务方法虽然能消除传统方法的不足, 但人工审核具有较大的主观性, 且工作量大, 不适合海量数据处理, 同时受限于多源观测资料, 其对持续时间长的沙尘天气过程识别效果较好, 但对持续时间短的沙尘天气较难识别。Z2020 与本文均为基于 K-means 聚类的客观识别方法, 适用于海量数据的处理, 但 Z2020 以 AQI 指数、 $\text{PM}_{2.5}$ 浓度、 PM_{10} 浓度为数据指标, 重复利用 $\text{PM}_{2.5}$ 及 PM_{10} 浓度的观测信息, 一定程度上增加了聚类的复杂性, 不利于对聚类结果的释用。此外, 在识别结果

的判断、第二次聚类的设置方面，Z2020 相较于本文存在更多的人工参与，其结果受主观性影响更大。

表 9 本文方法与以往方法识别效果对比

Table 9 Comparison of recognition performance between our method and previous methods

	沙尘天气影响时段	沙尘天气统计特征	优点	不足
传统方法	共识别出 244 组数据、持续 6 个时段、影响 16 天	PM ₁₀ 浓度大于 69 $\mu\text{g}/\text{m}^3$ 、PM ₁₀ 与 PM _{2.5} 浓度比值大于 2.04	具有较为清晰的物理基础、操作方便	选用的临界值适用性不足、易漏判沙尘天气，或识别的沙尘天气持续时间偏长
业务方法	共识别出 115 组数据、持续 5 个时段、影响 9 天	PM ₁₀ 浓度大于 133 $\mu\text{g}/\text{m}^3$ 、PM ₁₀ 与 PM _{2.5} 浓度比值大于 2.73	参考卫星观测和气象台站等多种资料对传统方法结果进行修正，对沙尘过程识别较准确	人工审核受主观因素大，且工作量大、不适合海量数据处理、较难识别持续时间短的沙尘过程
郑 昊 等 (2020)	共识别出 111 组数据、持续 9 个时段、影响 13 天	PM ₁₀ 浓度大于 260 $\mu\text{g}/\text{m}^3$ 、PM ₁₀ 与 PM _{2.5} 浓度比值大于 2.73	客观方法、适用于海量数据处理、识别结果可靠	采用的 AQI 指数包含污染物浓度信息，即重复使用了观测信息，且聚类结果的判别仍需要较多的主观判断
本文	共识别出 130 组数据、持续 15 个时段、影响 15 天	PM ₁₀ 浓度大于 228 $\mu\text{g}/\text{m}^3$ 、PM ₁₀ 与 PM _{2.5} 浓度比值大于 2.30	客观方法、具有清晰的物理基础、操作简便、适用于海量数据处理、识别结果可靠	难以识别弱沙尘天气影响时段

本文方法在物理基础、操作流程、效率及适用性等方面均具有明显的优势，但很难识别出弱沙尘天气的影响时段，也是目前所有沙尘天气识别方法的共性。识别方法通常根据其特定的临界条件进行沙尘天气判定，对于不符合临界条件但在临界条件附近的观测值，均判定为非沙尘天气，客观来讲，这样的判定结果存在不合理性。例如，本文方法在 4 月 2 日的 6~7 时、9~10 时均识别出沙尘天气，但两时段间隔的 4 月 2 日 8 时未识别出，该时刻对应的 PM₁₀ 浓度为 225 $\mu\text{g}/\text{m}^3$ （稍低于方法中 228 $\mu\text{g}/\text{m}^3$ 的临界值），PM_{2.5} 占 PM₁₀ 浓度的比例为 31%（满足方法中小于 43.5% 的判定条件），数据符合沙尘天气典型特征，合理的结果是将该时刻判定为沙尘天气影响时段。为解决弱沙尘天气的判定问题，可以采用以下方法：一是集成多种方法进行综合分析，弥补单一方法的不足，如通过时间序列分析方法对聚类分析结果进行修正，可以避免出现上述将 4 月 2 日 8 时判定为非沙尘天气时段的问题，从而更好地识别沙尘天气过程；二是辅以大气化学成分、大气沉降等其他要素的观测资料（韩力慧等，2005；薛一波等，2023），也即充分利用各种观测资料进行分析，特别是选取更精细化的观测指标来进行识别，如可利用污染物中特定元素富集因子的变化识别污染来源，能更准确地识别沙尘天气。

5 结语

本文基于中国环境监测总站观测的 $\text{PM}_{2.5}$ 和 PM_{10} 浓度观测资料，提出了一种基于 K-means 聚类算法的沙尘天气客观识别方法，并对识别的结果进行分析和检验。识别过程具体如下：

(1) 确定分类数目 K：采用随机初始聚类中心进行聚类，根据最终聚类中心的差异程度和沙尘天气的整体效果，可以确定 4 种天气类型，即无污染、轻度污染或雾霾型、沙尘型、混合型；

(2) 二次聚类：对无法直接判断污染类型的混合型天气进行第二次聚类，确定分类数目 K_2 ，进一步识别出沙尘型、非沙尘型；

(3) 判定聚类结束：对比混合型二次聚类后的分类结果，若两种类型的区别较为明显（即非沙尘型明显不同于沙尘型，可认为是雾霾型污染），则聚类结束，否则，不断进行第二步进行分类。

(4) 最终分类：综合上述识别效果，得到无污染天气、沙尘天气、轻度污染或雾霾天气，共 3 类事件。

西安市 2018 年 2~4 月小时 $\text{PM}_{2.5}$ 和 PM_{10} 浓度的聚类分析表明，本文方法简单清晰，易于操作，沙尘天气识别结果可靠。此外，得到西安市沙尘天气典型特征： $\text{PM}_{2.5}$ 占 PM_{10} 浓度的比例的上限为 43.5%、 PM_{10} 浓度的下限为 $228\mu\text{g}/\text{m}^3$ ，符合沙尘天气的物理特征，即可吸入颗粒物浓度较高、粗颗粒物含量较高。在以往研究中，部分学者将 $\text{PM}_{2.5}$ 占 PM_{10} 浓度的比例作为沙尘天气判断标，选取的值包括 39%(Tong et al., 2012)、40%(Fan et al., 2021)，本文所给出的上限值为 43.5%，接近于上述研究给出的结果。另一方面，气象台站通常采用 10km 能见度作为沙尘天气的判断条件，其所对应的 PM_{10} 浓度为 $205\mu\text{g}/\text{m}^3$ (Song et al., 2007; Wu et al., 2021)，因此本文得到的下限值，即 $228\mu\text{g}/\text{m}^3$ ，也接近于该判定条件。可见基于上述沙尘天气聚类识别方法能揭示沙尘污染的典型特征，可为沙尘天气的研究提供参考依据。

相比于传统方法、业务方法、郑湔等 (2020)，本文方法在物理基础、操作流程、效率及适用性等方面均具有明显优势，且适用于海量数据处理，具有广阔的应用前景。一方面，基于该方法得到的污染天气聚类中心以及典型特征指标

可用于沙尘天气的实时监测,为研判沙尘天气分布范围和发展趋势并进行预警提供决策支撑;另一方面,中国环境监测总站已积累了近十年的污染物浓度观测资料,基于这些数据应用本文方法识别出沙尘天气,有望建立一套我国沙尘天气数据集,从而有助于深入认识我国沙尘天气的时空演变特征,并为沙尘模式和预测系统的发展及评估检验提供关键的数据支撑。

本文方法也存在着以往识别方法普遍存在的问题,即难以识别弱沙尘天气。这需要集成多种方法,并综合利用不同要素观测资料进行改进(陈思宇等,2023)。同时,由于本文方法目前只在西安市2018年2~4月间进行试验,考虑到污染天气具有较大的时空变化特征,故对于该方法在其他地区及更长时间尺度的适用性,未来还需进一步检验和分析。

致谢 感谢陕西省环境监测中心站郑湔博士提供本文研究的观测数据以及研究中给予的帮助。

参考文献 (References)

- 石广玉, 赵思雄. 2003. 沙尘暴研究中的若干科学问题[J]. 大气科学, 27(4): 591-606. Shi Guangyu, Zhao Sixiong. 2003. Several scientific issues of studies on the dust storms[J]. Chinese Journal of Atmospheric Sciences(in Chinese), 27(4): 591-601.
- 姜大海, 王式功, 尚可政. 2011. 沙尘暴危险度的定量评估研究[J]. 中国沙漠, 31(6): 1554-1562. Jiang Dahai, Wang Shigong, Shang Kezheng. 2011. Quantitative assessment of sandstorm risk[J]. Journal of Desert Research(in Chinese), 31(6): 1554-1562.
- 石广玉, 檀赛春, 陈彬. 2018. 沙尘和生物气溶胶的环境和气候效应[J]. 大气科学, 42(3): 559-569. Shi Guangyu, Tan Saichun, Chen Bin. 2018. Environmental and climatic effects of mineral dust and bioaerosol[J]. Chinese Journal of Atmospheric Sciences(in Chinese), 42(3): 559-569.
- Y. Shao, C.H. Dong. 2006. A review on East Asian dust storm climate, modelling and monitoring[J]. Global and Planetary Change, 52:1-22.
- Chenglai Wu, Zhaohui Lin, Xiaohong Liu, et al. 2016. A process-oriented evaluation of dust emission parameterizations in CESM: Simulation of a typical severe dust storm in East Asia[J]. Adv. Model. Earth Syst., 8: 1432-1452.
- Chenglai Wu, Zhaohui Lin, Xiaohong Liu, et al. 2018. Can climate models reproduce the decadal change of dust aerosol in East Asia? [J]. Geophysical Research Letters, 45: 9953-9962.
- Chenglai Wu, Zhaohui Lin, Xiaohong Liu. 2020. The global dust cycle and uncertainty in CMIP5 (Coupled Model Intercomparison Project phase 5) models[J]. Atmos. Chem. Phys., 20: 10401-10425.
- Xunming Wang, Diwen Cai, Siyu Chen, et al. 2021. Spatio-temporal trends of dust emissions triggered by desertification in China[J]. CATENA, Volume 200, 105160.
- 吴成来, 林朝辉. 2014. WRF/Chem 模式中两种起沙参数化方案对东亚地区一次强沙尘暴过程模拟的影响[J]. 气候与环境研究, 19(4): 419-436. Wu Chenglai, Lin Zhaohui. Impact of two different dust emission schemes on the simulation of a severe dust storm in East Asia using the WRF/Chem model [J]. Climatic and Environmental Research (in Chinese), 19(4): 419-436.
- 周自江, 王锡稳, 牛若芸. 2002. 近 47 年中国沙尘暴气候特征研究[J]. 应用气象学报, 13(2): 193-200. Zhou Zijiang, Wang Xiwen, Nu Ruoyun. 2002. Climate characteristics of sandstorm in China in recent 47 years[J]. Journal of Applied Meteorological Science(in Chinese), 13(2): 193-200.
- Chenglai Wu, Zhaohui Lin, Yaping Shao, et al. 2022. Drivers of recent decline in dust activity over East Asia[J]. Nat Commun 13, 7105(2022).
- 杨晓军, 张强, 叶培龙, 等. 2021. 中国北方 2021 年 3 月中旬持续性沙尘天气的特征及其成因[J]. 中国沙漠, 41(3): 245-255. Yang Xiaojun, Ye Peilong, et al. 2021. Characteristics and causes of persistent sand-dust weather in mid-March 2021 over Northern China[J]. Journal of Desert Research(in Chinese), 41(3): 245-255.
- 高泽田, 胡秀清, 张小曳, 等. 2022. 基于 FY-4A 卫星遥感数据分析 2021 年 3 次沙尘暴特征[J]. 气象科技, 50(4): 536-544. Gao Zetian, Hu Xiuqing, Zhang Xiaoye, et al. 2022. Characteristic analysis of three sand-dust storm process in 2021 based on FY-4A satellite remote sensing data[J]. Meteorological Science and Technology(in Chinese), 50(4): 536-544.
- 张璐, 宋桂英, 范凡, 等. 2022. 2021 年 3 月 14-16 日中国北方地区沙尘暴天气过程诊断及沙尘污染输送分析[J]. 环境科学学报, 42(9): 351-363. Zhang Lu, Song Guiying, Fan Fan, et al. 2022. Diagnosis of sandstorm weather process and analysis of sand pollution transportation in northern China from 14th to 16th, March 2021[J]. Acta Scientiae Circumstantiae(in Chinese), 42(9): 351-363.

- 张鹏, 王春姣, 陈林, 等. 2018. 沙尘气溶胶卫星遥感现状与需要关注的若干问题[J]. 气象, 44(6): 725-736. Zhang Peng, Wang Chunjiao, Chen Lin, et al. 2018. Current status of satellite-based dust aerosol remote sensing and some issues to be concerned[J]. Meteorological Monthly(in Chinese), 44(6): 725-736.
- 钱正安, 蔡英, 刘景涛, 等. 2006. 中蒙地区沙尘暴研究的若干进展[J]. 地球物理学报, 49(1): 83-92. Qian Zhengan, Cai Ying, Liu Jingtao, et al. 2006. Some advances in dust storm research over China-Mongolia areas[J]. Chinese Journal of Geophysics(in Chinese), 49(1): 83-92.
- 张丽娟, 张建辉, 赵爱梅, 等. 2022. 利用环境减灾二号 A/B 卫星红外相机数据监测沙尘暴[J]. 航天器工程, 31(3): 159-165. Zhang Lijuan, Zhang Jianhui, Zhao Aimei, et al. 2022. Monitoring sand dust storms using data of infrared camera on HJ-2A/B satellites[J]. Spacecraft Engineering(in Chinese), 31(3): 159-165.
- 邱昀, 李令军, 鹿海峰, 等. 2018. 基于 Himawari 8 的沙尘遥感监测与应用[J]. 中国环境科学, 38(9): 3305-3312. Qiu Yun, Li Lingjun, Lu Haifeng, et al. 2018. Dust remote sensing monitoring and application based on Himawari 8[J]. China Environmental Science(in Chinese), 38(9): 3305-3312.
- 李彬, 卢士庆, 孙小龙, 等. 2018. 可见光波段灰度熵和红外亮温差的沙尘遥感判识[J]. 遥感学报, 22(4): 647-656. Li Bin, Lu Shiqing, Sun Xiaolong, et al. 2018. Dust detection algorithm based on the entropy of visible band brightness and brightness temperature difference[J]. Journal of Remote Sensing(in Chinese), 22(4): 647-656.
- Lili Yang, Shuwen Zhang, Zhongwei Huang, et al. 2021. Characteristics of dust events in China from 2015 to 2020[J]. Atmosphere 2021, 12, 952.
- 王蕾, 蒙瑞丽, 曹磊, 等. 2022. 2016-2020 年沙尘天气对陕西省空气质量的影响特征[J]. 中国沙漠, 42(4): 130-138. Wang Lei, Meng Ruili, Cao Lei, et al. 2022. Influence of dust events on air quality of Shanxi province from 2016 to 2020[J]. Journal of Desert Research(in Chinese), 42(4): 130-138.
- Zhenxin Song, Jinyan Wang, Shigong Wang. 2007. Quantitative classification of northeast Asian dust events[J]. J. Geophys. Res., 112, D04211.
- Yaqiang Wang, Xiaoye Zhang, Shuili Gong, et al. 2008. Surface observation of sand and dust storm in East Asia and its application in CUACE/Dust[J]. Atmos. Chem. Phys., 7: 9115-9138.
- 李晓岚, 张宏升. 2012. 2010 年春季北京地区强沙尘暴过程的微气象学特征[J]. 气候与环境研究, 17(4): 400-408. Li Xiaolan, Zhang Hongsheng. 2012. Micrometeorological characteristics during severe dust storms in spring of the Beijing area in 2010 [J]. Climatic and Environmental Research(in Chinese), 17 (4): 400-408.
- 郑昊, 薛惠锋, 李养养, 等. 2020. 基于 K-means 聚类的沙尘天气快速识别技术研究[J]. 中国环境监测, 36(5): 130-137. Zheng Hao, Xue Huifeng, Li Yangyang, et al. 2020. Research on the fast recognition technology of sand dust weather based on K-means Clustering[J]. Environmental Monitoring in China(in Chinese), 36(5): 130-137.
- 孙吉贵, 刘杰, 赵连宇. 2008. 聚类算法研究[J]. 软件学报, 19(1): 48-61. Sun Jigui, Liu Jie, Zhao Lianyu. 2008. Clustering algorithms research[J]. Journal of Software, 19(1): 48-61.
- 徐大川, 许宜诚, 张冬梅. 2018. K-均值算法的初始值方法综述 [J]. 运筹学学报, 22(2): 31-40. Xu Dachuan, Xu Yicheng, Zhang Dongmei. 2018. A survey on the initialization methods for the K-means algorithm[J]. Operations Research Transactions(in Chinese), 22(2):31-40.
- 欧陈委. 2011. K-均值聚类算法的研究与改进[D]. 2011. 长沙理工大学硕士学位论文. Ou Chenwei. 2011. Research and improvement on K-means clustering algorithm[D]. M. S. thesis(in Chinese), Changsha University of Science and Technology.
- 王昊雷. 2015. K 均值聚类算法研究与应用[D]. 哈尔滨工程大学硕士学位论文. Wang Haolei. 2015. Research and application of K-means clustering algorithm[D]. M. S. thesis(in Chinese), Harbin Engineering University.
- Chenglai Wu, Zhaohui Lin, Xiaohong Liu, et al. 2021. Description of dust emission parameterization in CAS-ESM2 and its simulation of global dust cycle and East Asian dust events[J]. Journal of Advances in Modeling Earth Systems, 13, e2020MS002456.
- 韩力慧, 庄国顺, 孙业乐, 等. 2005. 北京大气颗粒物污染本地源与外来源的区分—元素比值 Mg/Al 示踪法估算矿物气溶胶外来源的贡献 [J]. 中国科学 (B 辑化学), 35(3): 237-246.
- 薛一波, 黄双燕, 张小嘴, 等. 2023. 新疆 2018 年冬季雨雪风沙强降水事件 [J]. 干旱区研究, 40(5): 681-690. Xue Yibo, Huang Shuangyan, Zhang Xiaoxiao, et al. 2023. Study on the strong winter airborne dustfall mixed rain and snow events in Xinjiang, China in 2018 [J]. Arid zone research (in Chinese), 40(5): 681-690.
- 陈思宇, 杜世康, 毕鸿儒, 等. 2024. 沙尘天气识别和预报方法研究[J]. 中国沙漠, 44(1): 1-11. Chen Siyu, Du Shikang, Bi Hongru, et al. 2024. Review on identification and forecasting of dusty weather [J]. Journal of Desert Research(in Chinese), 44(1): 1-11.
- D.Q.Tong, M.Dan, T.Wang, et al. 2012. Long-term dust climatology in the western United States reconstructed from routine aerosol ground monitoring[J]. Atmos. Chem. Phys., 12: 5189-5205.
- Hao Fan, Chuanfeng Zhao, Yikun Yang, et al. 2021. Spatio-Temporal Variations of the PM_{2.5}/PM₁₀ Ratios and Its Application to Air Pollution Type Classification in China[J]. Sec. Atmosphere and Climate, Volume 9 - 2021.