Advances in Applied Mathematics 应用数学进展, 2022, 11(1), 231-237
Published Online January 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.111029

Hans汉斯

基于非负矩阵分解方法的主要城市环境质量状况分析

郭　菲

贵州大学数学与统计学院，贵州　贵阳

收稿日期：2021年12月17日；录用日期：2022年1月6日；发布日期：2022年1月20日

摘　要

根据水环境、大气环境、固体废物、声环境四方面的环境质量建立评价指标体系，利用非负矩阵分解方法对2020年全国31个主要城市的环境质量状况进行聚类分析评价。结果表明在2020年全国31个主要城市的环境质量状况中，环境质量状况在不同城市之间表现有差异，并且聚类结果呈现一定的空间分布特征。研究揭示了城市的环境质量状况与空间分布情况，能为城市生态环境状况发展建设提供参考。

关键词

环境质量评价，非负矩阵分解，聚类分析

Analysis of Environmental Quality of Major Cities Based on Non-Negative Matrix Factorization

Fei Guo

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Dec. 17th, 2021; accepted: Jan. 6th, 2022; published: Jan. 20th, 2022

Abstract

Establish an evaluation index system based on the environmental quality of water environment, atmospheric environment, solid waste, and acoustic environment, and use the non-negative matrix decomposition method to perform cluster analysis and evaluation on the environmental quality of 31 major cities across the country in 2020. The results show that in the environmental quality status of 31 major cities across the country in 2020, the environmental quality status varies

文章引用: 郭菲. 基于非负矩阵分解方法的主要城市环境质量状况分析[J]. 应用数学进展, 2022, 11(1): 231-237.
DOI: 10.12677/aam.2022.111029

between different cities, and the clustering results show certain spatial distribution characteristics. The research reveals the environmental quality and spatial distribution of the city, which can provide a reference for the development and construction of the urban ecological environment.

Keywords

Environmental Quality Assessment, Non-Negative Matrix Factorization, Cluster Analysis

1. 引言

我国城市在实现中高质量发展的同时，资源枯竭、经济结构失衡、环境污染等社会问题频现，环境问题日益受到重视[1]。2021 年我国“十四五”规划指出，我国发展不平衡不充分问题仍然突出，重点领域关键环节改革任务仍然艰巨，生态环保任重道远。应当深入打好污染防治攻坚战，建立健全环境治理体系，推进精准、科学、依法、系统治污[2]，不断改善空气、水环境质量，有效管控土壤污染风险，构建生态文明体系，推动经济社会发展全面绿色转型，建设美丽中国，以促进可持续发展[3] [4] [5]。

本文通过建立反映环境质量状况的指标体系，运用非负矩阵分解方法，对 2020 年全国 31 个主要城市的环境质量状况进行了聚类分析评价研究，得出其环境质量状况分类与环境质量状况的影响因素，由此为城市环境污染治理与可持续发展提供科学依据。

2. 环境质量评价体系构建

本文旨在分析全国 31 个主要城市的环境质量情况，为更能充分考虑城市工业、农业、居民生活方面的环境状态，从水环境、空气环境、固体废物和声环境四个方面选取了与环境质量密切相关的 16 个指标[6]，见表 1。其中，指标选取包含工业源和城镇生活源两方面。

Table 1. Environmental pollution degree evaluation index system
表 1. 环境污染程度评价指标体系

指标	指标内容	符号	单位
水环境	工业化学需氧量排放量	X_1	吨
	工业氨氮排放量	X_2	吨
	生活化学需氧量放量	X_3	吨
	生活氨氮排放量	X_4	吨
固体废物	工业固体废物产生量	X_5	万吨
	工业固体废物综合利用量	X_6	万吨
	工业固体废物处置量	X_7	万吨
	工业固体废物贮存量	X_8	万吨

Continued

大气环境	工业二氧化硫排放量	X_9	吨
	工业氮氧化物排放量	X_{10}	吨
	工业烟(粉)尘排放量	X_{11}	吨
	生活二氧化硫排放量	X_{12}	吨
	生活氮氧化物排放量	X_{13}	吨
	生活烟尘排放量	X_{14}	吨
声环境	环境噪声等效声级	X_{15}	dB(A)
	道路交通等效声级	X_{16}	dB(A)

3. 数据处理与研究方法

3.1. 数据来源与处理

本文所有原始数据均来源于《中国统计年鉴 2021》，由于各个指标间存在量纲上的不统一，在对数据进行非负矩阵分解聚类前，必须对数据进行处理，且各指标属于不同类型，有的指标的属性值越大越好(例如：一般工业固体废物综合利用量)，而有些指标的属性值越小越好(例如：工业化学需氧量排放量)，所以采用 min-max 标准化方法进行指标属性值的标准化，消除变量间在量纲和类型上的不同以及对聚类结果的影响。

对正向指标：

$$X_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

对负向指标：

$$X_i = \frac{x_{\max} - x_i}{x_{\max} - x_{\min}} \tag{2}$$

3.2. 非负矩阵分解

聚类分析通过距离或相似性对数据进行分组，是数据探索、压缩和展现的重要工具。本文采用非负矩阵分解聚类分析方法对环境质量进行评价。聚类分析是依据研究对象的个体特征，对其进行分类的方法，指标之间高度相关性导致了传统聚类分析方法无法获得良好的分类效果，已有研究采用聚类分析进行评价多采用主成分聚类分析方法进行评价，但主成分方法其拟合结果未必能最有效地提取最优类别信息，进而影响到后续聚类分析效果[7]。而聚类算法本质上可描述为矩阵分解问题，非负矩阵分解(Non-negative matrix factorization, NMF)是一种基于矩阵分解的降维手段，能实现高维的数据矩阵降维处理[8]。

NMF 的基本思想可以简单描述为：对于任意给定的一个非负矩阵 X，NMF 算法能够寻找到一个非负矩阵 U 和一个非负矩阵 V，使得满足 $X \approx UV$，从而将一个非负的矩阵分解为左右两个非负矩阵的乘积[9]。NMF 是一种相对新颖的数据维度约简技术。它将非负性约束引入矩阵分解过程中，原始矩阵一般可分解为

$$X = UV + E \approx UV \tag{3}$$

其中，$X, U, V \geq 0$，假设阶矩阵，非负矩阵 U 和 V 分别为矩阵和矩阵，可分别视为基矩阵和权重系数矩

阵。$m\times n$ 的噪声矩阵 E 表示逼近误差，一般要求 $\|E\|_F$ 尽可能小并且快速收敛。r 为矩阵 X 的秩，通常根据实际情况选取并满足 $(m+n)r<mn$，即利用少量的基向量便可表示高维数据。这意味着 V 可替代 X，由此实现了对 X 的降维，进而获得反映数据本质特征的低秩矩阵 U，UV 能够近似拟合 X。

3.3. 非负矩阵分解迭代算法

式(3)中提出的近似问题通常被表述为如下优化问题

$$\min_{U,V\geq 0}\|X-UV\|_F^2 \tag{4}$$

该问题可以利用约束交替最小二乘法[10] (Paatero 和 Tapper，1994)、乘法更新规则[11] [12] (Lee 和 Seung，2001)、投影梯度方法[13] (Hoyer, 2004)等算法求解。事实上，NMF 算法可被解释为 E 服从不同概率分布假设下的最大似然算法。基于 NMF 的聚类特性，聚类中 U 解释为聚类中心矩阵，V 反映聚类信息。

4. 31 个主要城市环境质量评价

4.1. 基于 NMF 的聚类分析

本文研究的对象为全国 31 个主要城市，在水环境、大气环境、固体废物、声环境四方面，共 16 个指标，分析 2020 年各城市的环境质量状况。本文对反映 2020 年 31 个主要城市环境质量状况的数据整理标准化后，利用 R 完成非负矩阵分解的聚类分析。在 NMF 算法中，是通过随机初始 U、V，开始的迭代算法，由于损失函数 $f(U,V)$ 可能有局部最小，所以不同的初始 U、V 可能会得到不同的结果，如果算法的结果收敛到一个比设定 k 更低的低秩，说明结果不是最优的，需要重新选择 k 值。NMF 方法中判断 rank 重要的标准是 cophenetic 的大小，如图 1，本文通过多次迭代计算 cophenetic，获得最优 k 值为 4。

给定矩阵 X 和一个参数 K，NMF 算法将高维的地区环境质量状况矩阵分解成一个系数矩阵 U 和一个基矩阵 V。在基矩阵中，为第 k 类行为在第 n 个地区上的值。在系数矩阵中，为第 m 个地区中第 k 类指标的权重系数，则地区环境质量状况可视为地区多种指标线性加权的结果。同时，可根据权重系数进行聚类，当第 k 类行为的权重系数最大时，则认为该地区属于第 k 类，因此 K 也是聚类个数。

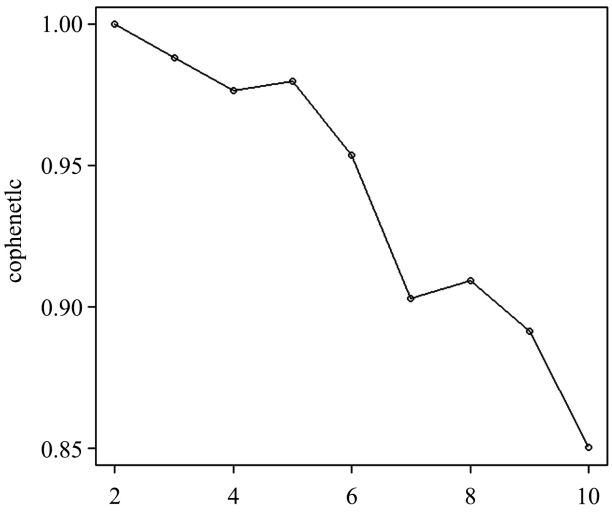


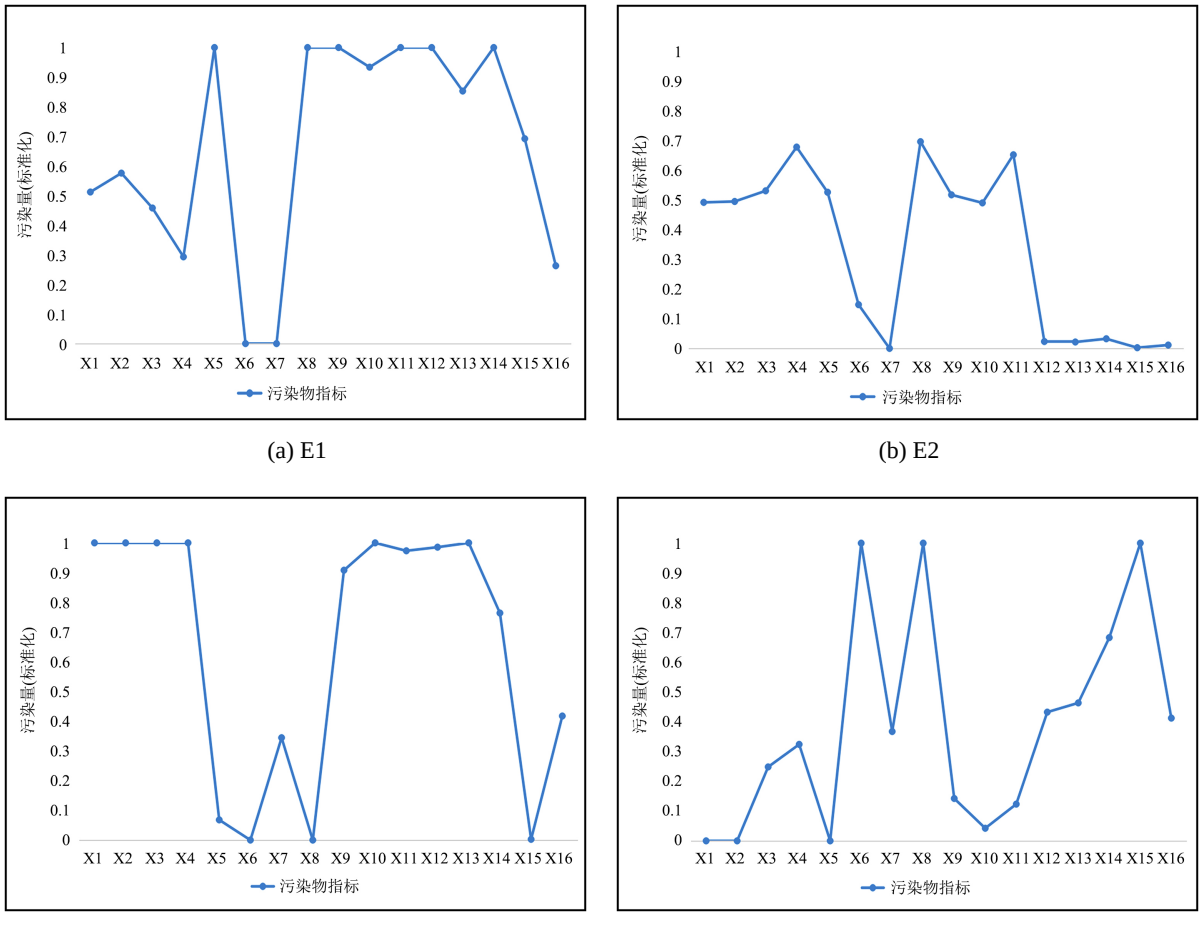
Figure 1. Determination of parameter K
图 1. 参数 K 的确定

综上，选取 $k = 4$，获取矩阵分解结果，即原始数据集聚类得到的簇数目为 4，每个地区包含的属性项维数为 4。

4.2. 环境质量状况矩阵分解

经过非负矩阵分解后，可得到维得基矩阵 V 和系数矩阵 U，其中 V 是一个 4×16 维的基矩阵，每一行代表一种环境质量状况(E1, E2, E3, E4)。

如图 2，总体上看，每种环境状况都表现为有一个主要的指标，不同的状况突出的环境指标不同，E1 在工业固体废物产生量、工业废物贮存量、工业二氧化硫排放量、工业氮氧化物排放量、工业烟(粉)尘排放量、生活二氧化硫排放量、生活氮氧化物排放量和生活烟尘排放量的值较高，主要反映了在 E1 环境质量状况中固体废物和大气环境的污染程度较高；E2 在生活化学需氧量放量、生活氨氮排放量、工业固体废物贮存量、工业二氧化硫排放量、工业氮氧化物排放量和工业烟(粉)尘排放量的值较高，主要反映了在这类环境状况中，生活废水和大气环境中工业方面的影响是较严重的；E3 水环境和大气环境中各项的指标值都较高，主要反映了在这类环境质量状况中，水污染和大气污染是较严重的；E4 在工业固体废物综合利用量、工业固体废物贮存量、环境噪声等效声级、生活化学需氧量放量和生活氨氮排放量的值较高，主要反映在这类环境质量状况中，声环境和生活废水污染影响较大。



(a) E1　(b) E2

(c) E3　(d) E4

Figure 2. Four environmental quality conditions
图 2. 四种环境质量状况

U 是一个 31×4 维的系数矩阵，矩阵中的元素为每个城市在每种环境状况下行为(E1, E2, E3, E4)的权重系数，由权重系数最大值将城市分为 4 类，结果见表 2。其中，第 2 类的城市数量最多，第 4 类次之，第 1、3 类相同。

Table 2. Clustering results of 31 major cities
表 2. 31 个主要城市聚类结果

类别	城市
第一类别	杭州、南昌、长沙、广州、南宁、成都、西宁
第二类别	北京、沈阳、长春、哈尔滨、武汉、海口、西安、兰州、乌鲁木齐
第三类别	太原、呼和浩特、合肥、福州、郑州、昆明、拉萨
第四类别	天津、石家庄、上海、南京、济南、重庆、贵阳、银川

4.3. 基于聚类分析的环境质量评价

通过非负矩阵分解得到全国 31 个主要城市的环境状况分类，从第一类到第四类的城市环境质量状况中，每一类都有不同的主要影响指标。为进一步说明聚类结果的空间布局，将聚类结果表示在地图上。

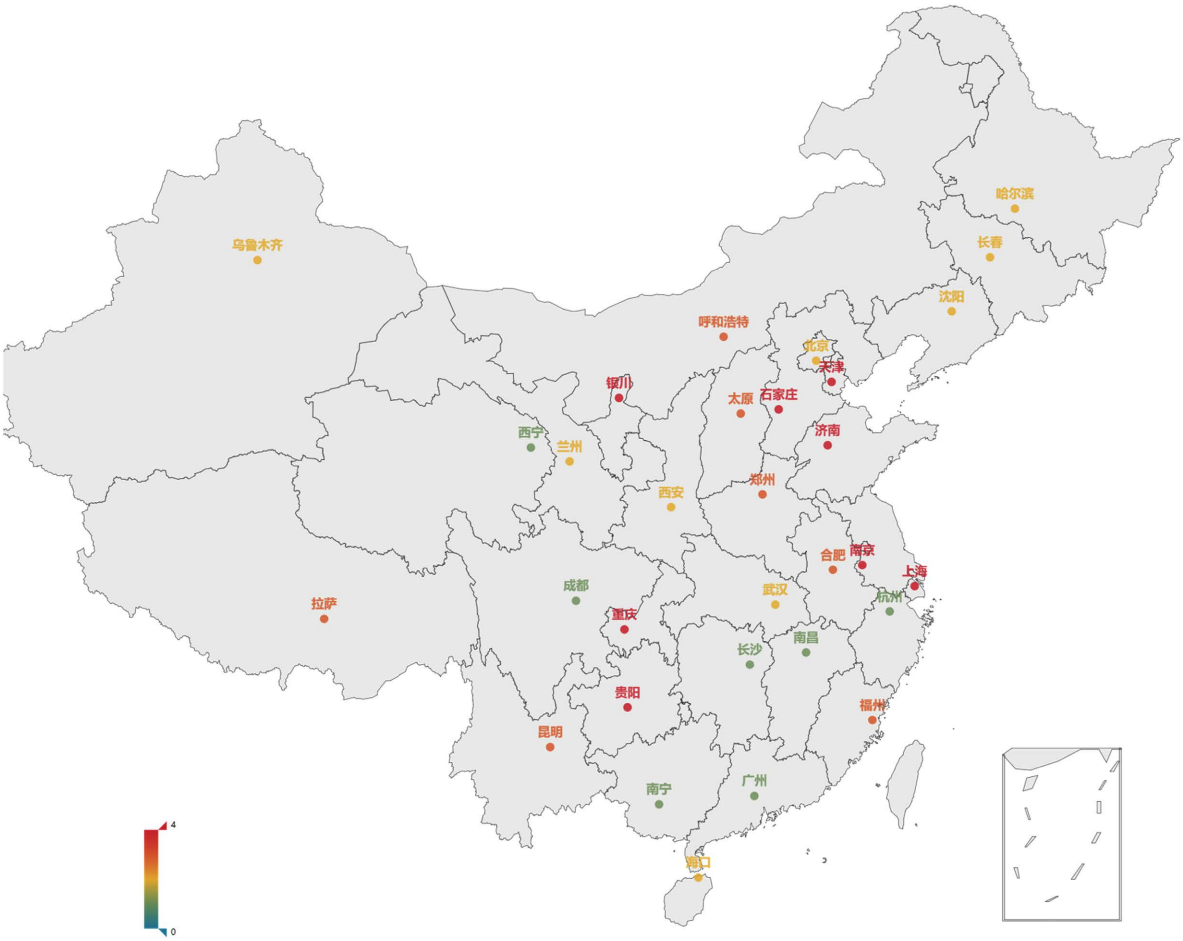


Figure 3. Spatial layout of clustering results
图 3. 聚类结果的空间布局

观察图 3 可知，31 个主要城市的环境质量状况分布特征总体表现出一定的区域特征，其中类别 1 主要集中分布在东南且呈现横线型结构，类别 2 和类别 3 主要分布在中西部及东北地区，而类别 4 则主要分布在东部地区，可以发现环境质量状况的分布和城市的地理分布具有一定的耦合性，反映了城市环境质量状况与城市发展行为、地理空间特征存在紧密的相互关联性。

5. 结论

基于 31 个主要城市环境数据和非负矩阵分解法对城市环境质量状况进行了聚类分析。研究发现：31 个主要城市的环境质量状况受到的主要污染指标不同，形成不同的环境质量状况；且在空间上表现出明显的空间分布特征。结合各城市的发展行为，这与实际的城市发展情况较为吻合。说明通过非负矩阵分解方法进行聚类分析是可行的，并且对各城市在发展生态环境质量方面具有一定的参考价值，可结合城市自身的发展方式、空间地理位置有针对性地防治污染，对突出问题重点处理，实现城市经济、生态高质量发展。但是，本文研究数据仅局限于 2020 年，有待对更多年份数据进行研究，分析环境质量状况的时空分布特征，同时对环境质量状况的指标选取还可以继续探索，尽可能对环境质量状况详尽全面分析。

参考文献

[1] 陈涛, 张越, 王玉阁, 李志英. 可持续发展视域下城市紧凑度与生态环境质量耦合协调关系研究——基于我国直辖市与省会城市的实证[J]. 生态经济, 2021, 37(10): 93-99+107.

[2] 李培, 王新, 柴发合, 王淑兰, 王镤一, 胡敏, 王瑞斌, 严刚, 任洪岩, 吴玉萍, 谢永明. 我国城市大气污染控制综合管理对策[J]. 环境与可持续发展, 2011, 36(5): 8-14.

[3] 卢敦, 张丽. 聚类分析与因子分析在天津港富营养化分析中的应用[J]. 数学的实践与认识, 2010, 40(11): 72-79.

[4] 黄恒君, 漆威. 海量半结构化数据采集、存储及分析——基于实时空气质量数据处理的实践[J]. 统计研究, 2014, 31(5): 10-16.

[5] 杨吉, 苏维词. 基于系统聚类分析的天河潭区域环境污染程度评价[J]. 环境工程, 2016, 34(8): 154-157+173.

[6] 陈军飞, 陈琳. 基于加权主成分距离聚类的江苏省环境质量评价[J]. 资源开发与市场, 2018, 34(10): 1383-1388.

[7] 黄恒君, 高海燕, 张梦瑶. 函数型聚类分析: 基于距离的一步法框架[J]. 数理统计与管理, 2019, 38(6): 986-995.

[8] 王宇辰. 非负矩阵分解算法综述[J]. 数字技术与应用, 2021, 39(2): 112-114.

[9] 高海燕, 黄恒君, 王宇辰. 基于非负矩阵分解的函数型聚类算法[J]. 统计研究, 2020, 37(8): 91-103.

[10] Paatero, P. and Tapper, U. (1994) Positive Matrix Factorization: A Non-Negative Factor Model with Optimal Utilization of Error Estimates of Data Values. *Environmetrics*, **5**, 111-126. https://doi.org/10.1002/env.3170050203

[11] Lin, CJ. (2007) Projected Gradient Methods for Nonnegative Matrix Factorization. *Neural Computation*, **19**, 2756-2779. https://doi.org/10.1162/neco.2007.19.10.2756

[12] Lee, D.D. and Seung, H.S. (1999) Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature*, **6755**, 788-791. https://doi.org/10.1038/44565

[13] Hoyer, P.O. (2004) Non-Negative Matrix Factorization with Sparseness Constraints. *Journal of Machine Learning Research*, **5**, 1457-1469.