Advances in Applied Mathematics 应用数学进展, 2017, 6(5), 711-715
Published Online August 2017 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2017.65085

Hans汉斯

Discussion on Correlation of Beijing-Tianjin-Hebei Region's Air Quality Index and Fuzzy Clustering Analysis

Hangliang Li, Huili Pei

Key Laboratory of Machine Learning and Computational Intelligence, College of Mathematics and Information Science, Hebei University, Baoding Hebei
Email: lihongliang003@126.com

Received: Aug. 5th, 2017; accepted: Aug. 20th, 2017; published: Aug. 24th, 2017

Abstract

This paper introduces a new statistical method by Bhattacharyya distance, studies thirteen cities' daily air quality index (AQI) data of the Beijing-Tianjin-Hebei region, between December 2013 and September 2016. The line charts generated by data, show that Beijing-Tianjin-Hebei regional AQI charts have a similarity. Then, based on Bhattacharyya distance, this article studies correlation of different cities' AQI, uses the fuzzy clustering method to cluster different regions' data, and analyzes the clustering results. The clustering results and the average of data show that most cities in the Beijing-Tianjin-Hebei region are under the condition of bad air quality.

Keywords

Beijing-Tianjin-Hebei Region, Air Quality Index, Fuzzy Clustering, Bhattacharyya Distance

京津冀地区空气质量指数相关性的研究及模糊聚类分析

李洪亮，裴慧丽

河北大学，数学与信息科学学院，机器学习与人工智能重点实验室，河北 保定
Email: lihongliang003@126.com

收稿日期：2017年8月5日；录用日期：2017年8月20日；发布日期：2017年8月24日

文章引用: 李洪亮, 裴慧丽. 京津冀地区空气质量指数相关性的研究及模糊聚类分析[J]. 应用数学进展, 2017, 6(5): 711-715. DOI: 10.12677/aam.2017.65085

摘　要

本文提出基于巴氏距离的统计方法，研究了京津冀地区十三个城市2013年12月至2016年9月每日空气质量指数(AQI)数据，对该地区的空气质量的相关性进行了分析。通过观察数据生成的折线图，发现京津冀各地区的AQI折线图具有一定的相似性。于是，本文基于巴氏距离研究了各地空气质量的相关性，利用模糊聚类方法对上述地区聚类，并分析了聚类结果。聚类结果及数据平均值显示京津冀地区大部分城市处于空气质量恶劣的条件下。

关键词

京津冀地区，空气质量指数，模糊聚类，巴氏度量

1. 引言

用近年来，空气质量问题一直困扰着中国北方很多城市。特别突出的是京津冀地区频繁出现严重雾霾天气。这种现象吸引了大量的学者进行研究[1] [2] [3]，近几年的相关研究数量也成快速增长趋势。模糊数学是一门实用性很强的数学科学[4] [5]，运用模糊数学方法研究京津冀地区的空气质量数据蕴含的规律，可以定量的验证人们对客观观察的猜测，同时，也能挖掘出一些人们没有观察到的一些规律。

2. 京津冀地区 AQI 数据定性分析及问题提出

通过记录 2013 年 12 月 2 日至 2016 年 9 月 30 日共计 1034 天，京津冀地区北京，天津，石家庄等共计 13 个城市的 AQI 数据[6]，以北京，天津，石家庄，张家口为例绘制数据折线图如图 1。

通过观察图 1 中数据，可以发现：北京，天津，石家庄的 AQI 数据折线图形状有一定的相似。这表明上述数据具有一定的相关性，各地区之间的 AQI 会相互影响。根据空气流动，可以猜测：相邻的城市之间，AQI 相互影响会比较大。同时，由于城市交界的复杂性，相邻城市哪一个对中心城市的影响最大，即相关性最大，这些问题需要具体的数据分析才能验证。

3. 京津冀地区 AQI 的相关性

巴氏距离是描述概率分布之间差异性的一种方式，可以定量的描述随机变量分布之间的差异。下面是离散型随机变量巴氏距离的定义：

定义 3.1 两个离散型随机变量 X,Y 的分布律为：$p_i,q_i,i=1,2,\cdots$。称 $BC(X,Y)=\sum_{i=1}^{\infty}\sqrt{p_iq_i}$ 为巴氏系数，称 $D_B(X,Y)=-\ln BC(X,Y)$ 为离散型随机变量 X,Y 的巴氏距离。

京津冀地区的 $N=13$ 个城市 $M=1034$ 天的 AQI 指数，记为 x_{ij}，$i=1,2,\cdots,M$，$j=1,2,\cdots,N$，表示第 j 个城市第 i 天的 AQI 指数。为了度量不同城市之间的差异性，对数据作如下操作：

$$p_{ij}=\frac{x_{ij}}{\sum_{i=1}^{M}x_{ij}}$$

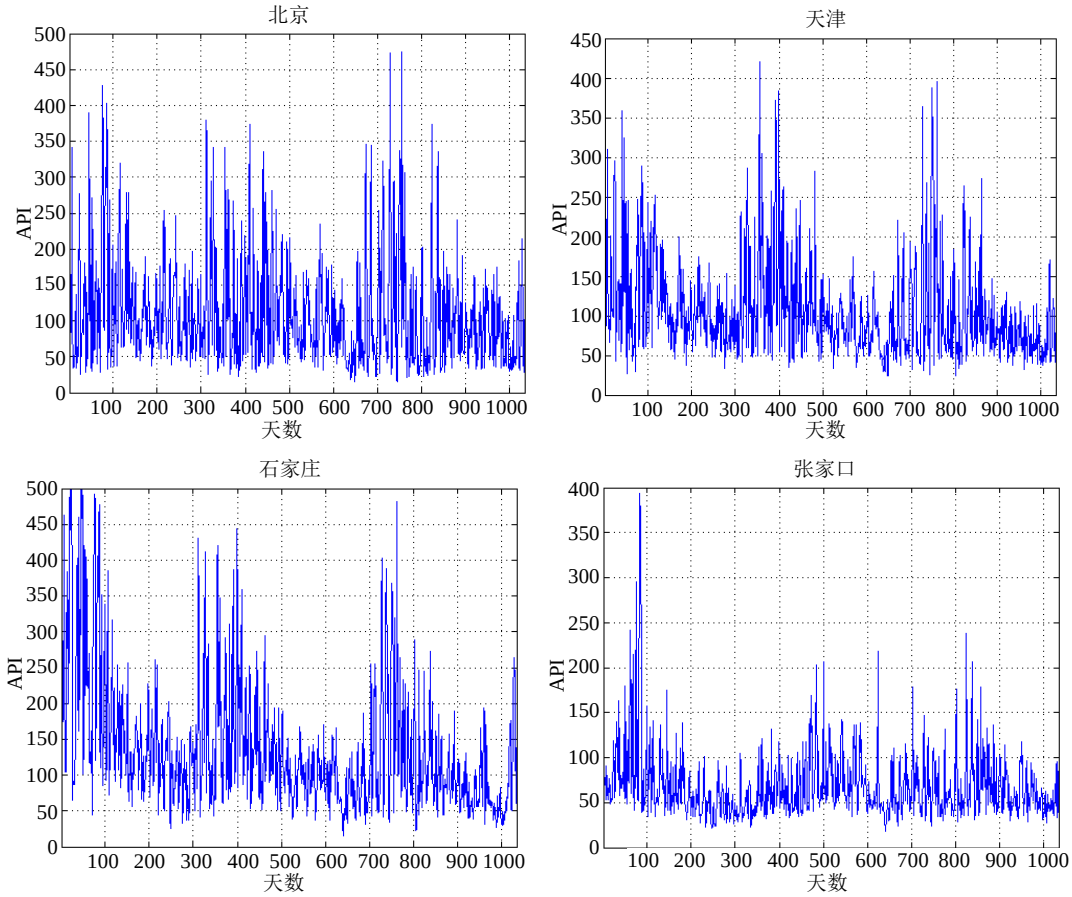

Figure 1. Line chart of AQI of Beijing, Tianjin, Shijiazhuang, Zhangjiakou
图 1. 北京，天津，石家庄，张家口的 AQI 数据折线图

通过这种变换，既保持原有单个城市的数据结构，又使数据成为某离散型随机变量的分布律，便于利用巴氏距离度量不同城市 AQI 的差别。计算不同城市间 p_{ij} 的巴氏度量，会得到13×13的巴氏距离矩阵，根据巴氏度量的对称性可知，该矩阵是对称矩阵。利用 MATLAB 计算，并添加城市名称信息得到表 1 巴氏度量矩阵 D_B。

通过对城市 j_1 所在行的非零数值取最小值 $a_{j_1 j_2}$，得到与城市 j_1 最相关的一个城市 j_2，记为(城市 j_1，城市 j_2)。得到如下结果：(石家庄，邢台)，(承德，北京)，(张家口，承德)，(秦皇岛，唐山)，(唐山，天津)，(廊坊，天津)，(保定，石家庄)，(沧州，衡水、天津)，(衡水，邯郸)，(邢台，石家庄)，(邯郸，邢台)，(北京，承德)，(天津，唐山)。可以看到除了衡水之外，其他 12 个城市 AQI 最相关的城市均与该城市相邻，证实了开始的猜测：相邻的城市之间，AQI 影响较大。但是，影响不一定是相互的。

4. 京津冀地区 AQI 的模糊聚类

由于巴氏距离矩阵本身是对称矩阵，而且巴氏距离取值于 $[0,1]$，因此巴氏距离矩阵表示了一种模糊相似关系。巴氏距离取值越小，表示两个分布越相似。为了利用模糊相似关系进行聚类，对巴氏距离矩阵做取余运算得到模糊相似矩阵：

$$FSR = DB^{c}$$

利用二次方法[5]，求得模糊相似矩阵 FSR 的传递闭包，即表 2，与 FSR 最接近的模糊等价矩阵 FER：

基于模糊等价矩阵，设置置信水平 $\lambda = 0.988$ 时，得到聚类结果：第一类：承德；第二类：张家口；第三类：秦皇岛；第四类：其他城市。

为了探讨上述聚类结果的实际意义，计算了各个城市 AQI 的平均值，如表 3。

该表显示显示：第四类城市的 AQI 平均值都已达到轻度污染程度，而前三类的承德，张家口，秦皇岛的 AQI 平均值在良好状态。

Table 1. Bhattacharyya distance matrix of 13 cities in Beijing-Tianjin-Hebei
表 1. 京津冀 13 个城市 AQI 巴氏度量矩阵

城市	石家庄	承德	张家口	秦皇岛	唐山	廊坊	保定	沧州	衡水	邢台	邯郸	北京	天津
石家庄	0	0.0282	0.0415	0.0302	0.0218	0.0185	0.0113	0.0176	0.0159	0.0067	0.014	0.0296	0.0204
承德	0.0282	0	0.019	0.0184	0.0151	0.0184	0.0257	0.0214	0.0257	0.0292	0.0302	0.0144	0.0185
张家口	0.0415	0.019	0	0.0342	0.0305	0.0358	0.0419	0.0349	0.0357	0.042	0.0415	0.0321	0.0343
秦皇岛	0.0302	0.0184	0.0342	0	0.0123	0.0207	0.0234	0.0202	0.0257	0.0287	0.0302	0.0284	0.014
唐山	0.0218	0.0151	0.0305	0.0123	0	0.0106	0.0172	0.0119	0.0176	0.0219	0.0227	0.0199	0.0071
廊坊	0.0185	0.0184	0.0358	0.0207	0.0106	0	0.0124	0.0145	0.0188	0.0208	0.025	0.0119	0.0091
保定	0.0113	0.0257	0.0419	0.0234	0.0172	0.0124	0	0.0141	0.0138	0.0132	0.0164	0.0261	0.0146
沧州	0.0176	0.0214	0.0349	0.0202	0.0119	0.0145	0.0141	0	0.0081	0.0155	0.0142	0.0293	0.0081
衡水	0.0159	0.0257	0.0357	0.0257	0.0176	0.0188	0.0138	0.0081	0	0.0107	0.0079	0.0338	0.0144
邢台	0.0067	0.0292	0.042	0.0287	0.0219	0.0208	0.0132	0.0155	0.0107	0	0.0074	0.0348	0.0203
邯郸	0.014	0.0302	0.0415	0.0302	0.0227	0.025	0.0164	0.0142	0.0079	0.0074	0	0.041	0.0207
北京	0.0296	0.0144	0.0321	0.0284	0.0199	0.0119	0.0261	0.0293	0.0338	0.0348	0.041	0	0.0219
天津	0.0204	0.0185	0.0343	0.014	0.0071	0.0091	0.0146	0.0081	0.0144	0.0203	0.0207	0.0219	0

Table 2. Fuzzy similar matrix of 13 cities in Beijing-Tianjin-Hebei
表 2. 京津冀 13 个城市 AQI 模糊相似矩阵

城市	石家庄	承德	张家口	秦皇岛	唐山	廊坊	保定	沧州	衡水	邢台	邯郸	北京	天津
石家庄	1	0.9856	0.981	0.9877	0.9919	0.9909	0.9887	0.9919	0.9921	0.9933	0.9926	0.9881	0.9919
承德	0.9856	1	0.981	0.9856	0.9856	0.9856	0.9856	0.9856	0.9856	0.9856	0.9856	0.9856	0.9856
张家口	0.981	0.981	1	0.981	0.981	0.981	0.981	0.981	0.981	0.981	0.981	0.981	0.981
秦皇岛	0.9877	0.9856	0.981	1	0.9877	0.9877	0.9877	0.9877	0.9877	0.9877	0.9877	0.9877	0.9877
唐山	0.9919	0.9856	0.981	0.9877	1	0.9909	0.9887	0.9919	0.9919	0.9919	0.9919	0.9881	0.9929
廊坊	0.9909	0.9856	0.981	0.9877	0.9909	1	0.9887	0.9909	0.9909	0.9909	0.9909	0.9881	0.9909
保定	0.9887	0.9856	0.981	0.9877	0.9887	0.9887	1	0.9887	0.9887	0.9887	0.9887	0.9881	0.9887
沧州	0.9919	0.9856	0.981	0.9877	0.9919	0.9909	0.9887	1	0.9919	0.9919	0.9919	0.9881	0.9919
衡水	0.9921	0.9856	0.981	0.9877	0.9919	0.9909	0.9887	0.9919	1	0.9921	0.9921	0.9881	0.9919
邢台	0.9933	0.9856	0.981	0.9877	0.9919	0.9909	0.9887	0.9919	0.9921	1	0.9926	0.9881	0.9919
邯郸	0.9926	0.9856	0.981	0.9877	0.9919	0.9909	0.9887	0.9919	0.9921	0.9926	1	0.9881	0.9919
北京	0.9881	0.9856	0.981	0.9877	0.9881	0.9881	0.9881	0.9881	0.9881	0.9881	0.9881	1	0.9881
天津	0.9919	0.9856	0.981	0.9877	0.9929	0.9909	0.9887	0.9919	0.9919	0.9919	0.9919	0.9881	1

Table 3. Average of AQI of 13 cities
表 3. 13 个城市 AQI 的平均值

石家庄	承德	张家口	秦皇岛	唐山	廊坊	保定	沧州	衡水	邢台	邯郸	北京	天津
137.05	78.2	67.848	82.94	120	119	148	108	141	148	134	110	107

5. 总结

通过对京津冀地区 13 个城市 1034 天的空气质量指数 AQI 的定量分析，证实了相邻地区的空气质量有影响，但这种影响不一定是相互的，如与保定 AQI 相关性最大的城市是石家庄，而与石家庄 AQI 相关性最大的城市是邢台。这在某种程度上反映了石家庄可能是污染源或空气质量较好的地区。但石家庄的 AQI 指数较高，我们倾向于其是污染源的可能性更大。

基于巴氏距离得出的各个城市 AQI 之间的模糊相似矩阵，利用传递闭包法对数据经行了聚类。当 $\lambda = 0.988$ 时，京津冀地区除了承德，张家口，秦皇岛三个城市，其他城市聚为一类，说明这些地区的 AQI 相关性较强。同时，由于上述地区 AQI 指数平均值较高。因此推测该地区出现雾霾的可能性较大。

京津冀地区城市的 AQI 之间有影响，但这种影响不一定是相互的，确实有一些城市的 AQI 是被动性的变差。因此，下一步的研究是对被动性污染城市与主动性污染城市的模糊聚类分析。

基金项目

国家自然科学基金项目(61572011)；河北省自然科学基金项目(F2016201161)；河北省高等学校科学技术研究重点项目(ZD2017005)；河北省教育厅青年基金(QN2014039)。

参考文献 (References)

[1] Wang, Y., Jia, C., *et al.* (2016) Chemical Characterization and Source Apportionment of PM2.5 in a Semi-Arid and Petrochemical-Industrialized City, Northwest China. *Science of the Total Environment*, **573**, 1031-1040. https://doi.org/10.1016/j.scitotenv.2016.08.179

[2] 徐恒鹏, 李岳, 史国良, 王玮, 轩淑艳. 基于模糊聚类的 PM2.5 拟合组分选择模型的研究[J]. 中国环境科学, 2016, 36(1): 12-17.

[3] 刘俊, 安兴琴, 朱彤, 翟世贤, 李楠. 京津冀及周边减排对北京市 PM2.5 浓度下降评估研究[J]. 中国环境科学, 2014, 34(11): 2726-2733.

[4] 吴从炘, 马明, 方锦暄. 模糊分析学的结构理论[M]. 贵阳: 贵州科技出版社, 1994.

[5] 谢季坚, 刘承平. 模糊数学方法及其应用[M]. 武汉: 华中科技大学出版社, 2006: 44-88.

[6] 中国空气质量在线监测分析平台. 空气质量历史数据[EB/OL]. https://www.aqistudy.cn/historydata/daydata.php?city=北京&month=201610, 2016-10-09.