编号：B477

基于惩罚样条的半参数模型下我国大城市空气质量的实证分析

论文题目：基于惩罚样条的半参数模型下我国大城市空气质量的实证分析

参赛学校：昆明理工大学

参赛成员（作者）：曾鑫、徐嘉阳、亢震

指导老师：彭俊　吴刘仓

目录

图目录

表目录

摘要

目的 暴露于空气被污染的环境中会增加死亡率和发病率，缩短预期寿命。本文拟基于惩罚样条的半参数模型对我国具有代表性的4个大城市(北京、上海、深圳和成都)的空气质量进行实证分析，探讨空气质量指数(AQI)的变化趋势，根据季节和地域的不同探究影响我国空气质量的不同因素, 并给出适当的政策建议。

对象和方法 搜集 2018 年 7 月-2021 年 5 月 AQI 及相关指标的日度数据并对原始数据进行检验和描述。基于 Box-Cox 转换方法修正原始数据中的偏度，使其近似服从正态分布，从而降低模型误差。分别从半参数可加模型、半参数交互模型、贝叶斯惩罚样条下的半参数曲线因子模型、稳健估计和半参分位数回归出发对 4 个城市的 AQI 进行了系统的分析，从边际非参数回归和基于贝叶斯惩罚样条下的半参数回归模型讨论了 AQI 的季节效应。

结果 本文的分析结果表明：北京市空气质量较上海、深圳和成都差，其中深圳在平均值和波动率方面表现最优。在半参数可加模型的 NO_2 含量对 AQI 的拟合中，上海、深圳和成都市都与北京市的 AQI 有显著的不同。可以合理地得出结论，上海市 AQI 平均值比北京市平均低 21.0440，而深圳和成都市的 AQI 平均值分别比北京市平均低 22.4551 和 13.7166；在半参数模型的稳健估计中，当 AQI 值一定时，非稳健估计都低估了 NO_2 和 $PM_{2.5}$ 的含量，尽管在某些值处这种低估非常小, 但仍然不可忽视；半参数分位数回归的结果表明, 0.01、0.05、0.25、0.5 和 0.75 分位数间隔很小, 表明每年的大多数时候我国的空气质量较好，同时得出结论：NO_2 和 $PM_{2.5}$ 含量对 AQI 的不同分位数的影响是不同的，即使 AQI 集中在值较小的区域内；考虑 AQI 变化趋势随季节而不同时，边际非参数回归的结果表明, 北京和成都空气质量最高的时间是秋季, 而深圳和上海则是夏季的空气质量更优。

结论 本文研究的模型的优点和特点在于：第一，本文使用半参数回归模型研究了我国的 AQI 数据，克服了传统的线性模型和部分时间序列分析中无法充分提取非线性信息的缺点；第二，基于分位回归的半参数模型对 AQI 离群值表现出稳健性。

关键词: AQI；惩罚样条基；半参数模型；贝叶斯分析；分位数回归

Abstract

Purpose Exposure to air pollution increases mortality and morbidity and shortens life expectancy.This paper based on the punishment of the spline semiparametric regression model for our country representative of the four big cities (Beijing, Shanghai, Shenzhen and Chengdu) the air quality of empirical analysis, and the trend of air quality index(AQI), depending on the season and region explore the different factors influencing the air quality in our country, and gives appropriate policy Suggestions.

Object and method The diurnal data of AQI and related indexes from July 2018 to May 2021 were collected, and the original data were tested and described.Based on the Box-Cox transformation method, the skewness in the original data is corrected to make it approximately follow the normal distribution, so as to reduce the model error.Respectively from semiparametric additive models, half parameter interaction model, bayesian punishment under the spline parametric curve factor model, the robust estimation and half and quantile regression based on the four cities of AQI system analysis, and from the marginal based on nonparametric regression and O 'Sullivan punishment spline under the semiparametric regression model discusses the AQI seasonal effect.

Result The analysis results show that the air quality in Beijing is worse than that in Shanghai, Shenzhen and Chengdu, and Shenzhen has the best performance in terms of average value and volatility.In the fitting of AQI with NO_2 content in the semi-parametric additive model, the AQI of Shanghai, Shenzhen and Chengdu is significantly different from that of Beijing.It can be reasonably concluded that the average AQI of Shanghai is 21.0440 lower than that of Beijing, while the average AQI of Shenzhen and Chengdu is 22.4551 and 13.7166 lower than that of Beijing, respectively.In the robust estimation of the semi-parametric model, when the AQI value is constant, the non-robust estimation all underestimate the content of NO_2 and

$PM_{2.5}$. Although this underestimate is very small at some values, it is still not negligible.In the semi-parameter quantile regression, the results show that the interval between 0.01, 0.05, 0.25, 0.5 and 0.75 quantiles is very small, indicating that the air quality in China is good in most of the time of the year. At the same time, it is concluded that NO_2 and $PM_{2.5}$ contents have different effects on the different quantiles of AQI, even if the AQI is concentrated in the region with small values.Considering that the variation trend of AQI varies with the seasons, in the marginal non-parametric regression, it is concluded that the air quality of Beijing and Chengdu is the highest in autumn, while that of Shenzhen and Shanghai is better in summer.

Conclusion The advantages and characteristics of the model studied in this paper are as follows: Firstly, the semi-parametric regression model is used to study the AQI data in China, which overcomes the shortcomings of the traditional linear model and partial time series analysis that cannot fully extract the nonlinear information; Second, the quantile-based semi-parametric regression is robust to the AQI outliers.

Keywords: AQI; Penalty spline basis; Semi-parametric model; Bayesian analysis; Quantile regression

一、引言

(一) 研究背景

近二十年来，随着我国经济的腾飞，环境问题也逐渐成为一个不可忽视的问题。同时，健康是我国人民实现全面发展以及生活幸福的重要前提，也是我国实现全面小康的重要内涵，更是民族昌盛和国家富强的重要标志。然而，空气污染问题已经深深地影响了人类的健康。人为向空气中排放氮氧化物的主要来源是固定来源(加热、发电)和机动车辆燃烧矿物燃料。在环境条件下，一氧化氮被臭氧等大气氧化剂迅速转化为二氧化氮。颗粒物空气污染是指悬浮在空气中的固体、液体或固液体颗粒的混合物。悬浮粒子的大小从几纳米到几十微米不等。最大的颗粒(粗粒级)是由较大颗粒的机械摩擦产生的，小于 1 微米的小颗粒大部分是由气体形成的，最小的(小于 0.1 微米，超细)是由凝结或形成新颗粒的化学反应成核而形成的。实际上，PM_{10} (粒径在 10 微米以下的颗粒，可以穿透到下呼吸系统)、$PM_{2.5}$ (粒径在 2.5 微米以下的颗粒的“可呼吸”颗粒，可以穿透到肺的气体交换区)都是影响空气质量的重要因素，其中，$PM_{2.5}$ 是我国空气质量面临的最严重的污染物之一，是我国雾霾中的主要成分。2021 年 3 月，我国北方部分地区遭受了强沙尘天气，$PM_{2.5}$ 和 PM_{10} 几乎都达到了历史最高水平。

在 1952 年的“英国伦敦烟雾事件”中，由于空气污染物浓度的急剧增加，仅在数天内死亡人数超过预期的 3 倍，估计导致超过 4000 人死亡。实际上，类似的事件曾在 1930 年比利时的默兹山谷以及其他地方发生过。于是，在过去的几十年，发达国家通过立法和转移工业等手段，大大减小了空气污染水平，传统污染物已经降低至很低的水平。然而，在许多发展中国家的大城市中，传统和现代各种污染物都暴露在极端环境中。实际上，在较为理想情况下，发达国家的经验教训可以帮助发展中国家走上一条更可持续、污染更少的工业化和现代化道路。然而，现有数据和研究结果都表明，全球竞争和人口增长的综合压力面前，许多发展中国家对空气质量的治理似乎没有回旋的余地。

已有众多研究结论表明，长期暴露于环境空气污染会增加死亡率和发病率，

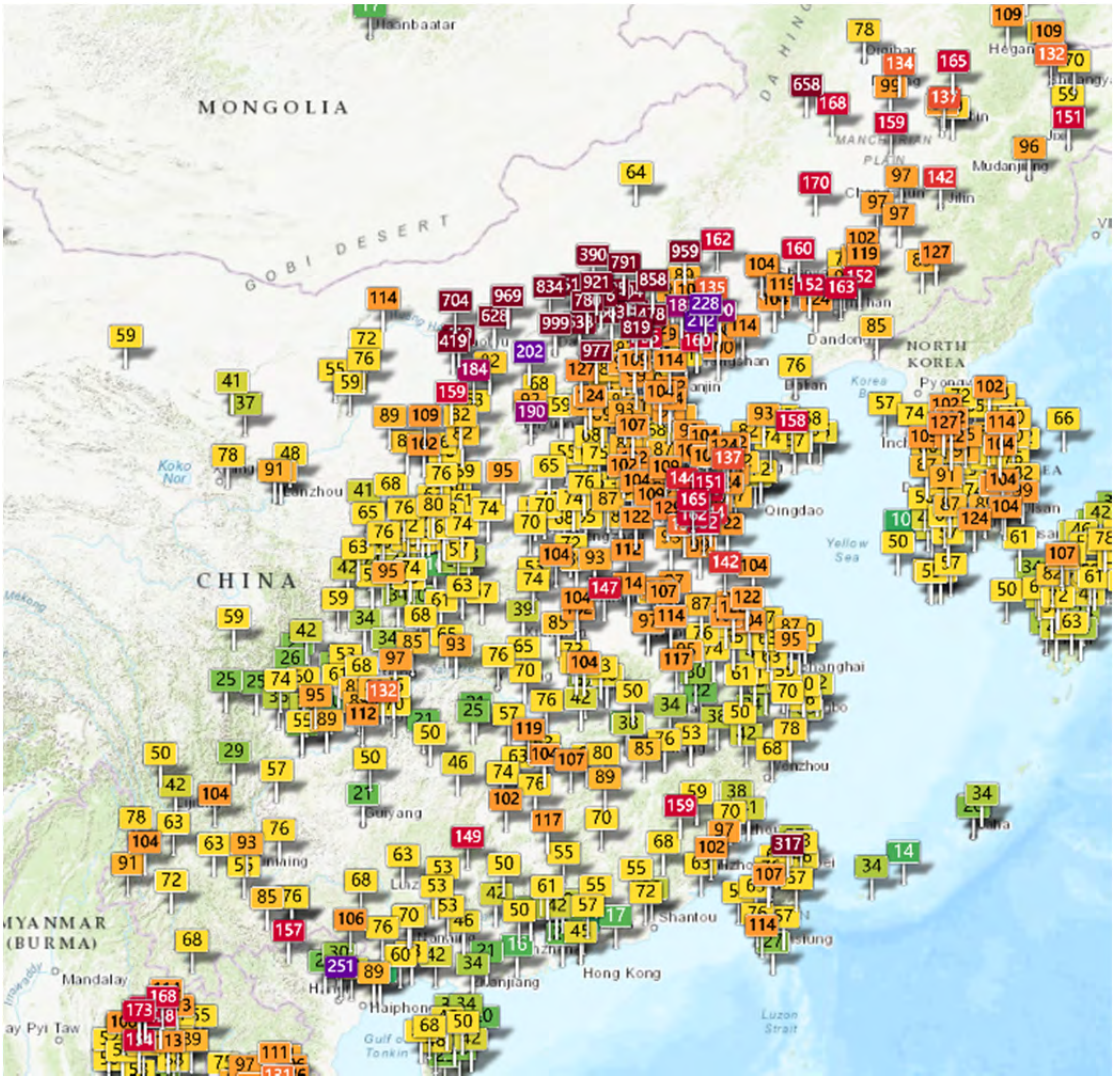

图 1 2021 年 5 月某日我国部分地区的实时 AQI

缩短人类的预期寿命。因此，空气质量一直吸引着当代空气污染科学家的关注。对空气质量的监测和评估对国家发展和人民幸福都起着至关重要的作用，尤其是对于我国这种发展中国家而言。

此外，由于影响空气质量 (AQI) 的因素众多，各指标与 AQI 的关系大多为非线性关系。同时，空气质量随着季节和地域的不同有着很大的差别，如图 1 为我国东部、东北部、中部和西南部地区的 2021 年 5 月某日的 AQI 实时监测系统，可以很容易看出，AQI 指数大于 100 的城市大多数分布在我国北方。因此，本文拟基于惩罚样条的半参数模型对我国四个大城市的空气质量进行实证分析，探讨空气质量指数的变化趋势，并根据时间和空间的不同探究影响我国空气质量的不

同因素。

(二) 国内外研究现状

现有的研究表明，暴露于环境空气污染会增加发病率和死亡率，是造成全球疾病负担的主要因素[1][2]。2015 年的一项研究结果显示，由于人口老龄化、非传染性疾病发病率的变化以及中低收入国家日益严重的空气污染，全球疾病负担在过去 25 年里有所增加。在全球疾病、伤害和危险因素负担(GBD 2015)研究中，估计了 195 个国家 1990 年至 2015 年 79 个危险因素造成的疾病负担，结果将空气污染确定为全球疾病负担的主要原因，特别是在低收入和中等收入国家中，因为这些国家在治理空气污染的公共政策方面所做的努力更小，甚至直接忽视了其重要作用[3]。

人们对环境微粒对健康的影响进行了大量的研究，从而达到了成功减少传统空气污染物的时代。在 20 世纪 70 年代后期发达国家出现的浓度，当时认为不太可能对健康产生不利影响[4]。然而，在此后的二十年里，空气污染再次成为一个主要的环境健康问题。一个原因是，尽管传统化石燃料燃烧造成的空气污染现在的浓度比 50 年前低得多，但其他成分的污染却日益突出。光化学空气污染发生在美洲、欧洲和亚洲的大部分地区，其特点是在温暖和晴朗的天气里臭氧浓度高。尽管目前新能源汽车已经得到了部分人的青睐，但机动车数量的不断增加所产生的氮氧化物直到今天仍在持续增加。其中有两项较为著名的研究，文献[5][6]两项研究都是基于 20 世纪 70 年代末至 80 年代末的数据，当时的空气污染浓度比过去低得多。而研究 AHSMOG 发现，直径小于 10 微米的颗粒物(PM_{10})对男性和女性的非恶性呼吸系统死亡以及男性的肺癌死亡有显著影响[7]。

实际上，现在二氧化硫的浓度已经显著下降，包括我国在内，由于前二十年间酸雨造成的影响，公共政策已经严格控制了 SO_2 污染物的排放标准。近年来，人们的注意力已经转移到臭氧(O_3)、二氧化氮(NO_2)和微粒($PM_{2.5}$,PM_{10})上。对于生活在发展中国家农村地区的大多数人来说，使用传统的化学燃料造成的空气污染的浓度较发达国家高出许多。实际上，APHEA(空气污染与健康:欧洲方法)的研究是基于较早的数据(APHEA-1) [8]；在 20 世纪 90 年代后期，一项新的系列研

究(APHEA-2)利用了 PM_{10} 的数据，APHEA-2 死亡率研究覆盖了 29 个欧洲城市超过 4300 万人，这些城市在 20 世纪 90 年代早期中期都进行了超过 5 年的研究[9]。从涉及 21 个城市的数据中的结果显示，PM_{10} 每增加 10 μg/m^3，所有原因的日死亡率增加 0.6% (95% 置信区间为 0.4% - 0.8%)。APHEA-2 医院入院研究涵盖了生活在 8 个欧洲城市的 3800 万人口，在 20 世纪 90 年代早期至中期进行了 3-9 年的研究[10]。在调查的 44 个医院中, 65 岁以上人群中, PM_{10} 浓度在每 10 μg/m^3 的哮喘和慢性阻塞性肺病(COPD)的入院率增加了 1.0%(95% 置信区间为 0.4%-1.5%)，心血管疾病(CVD)的入院率增加了约 0.5%(95% 置信区间为 0.2% - 0.8%)，同时该研究还发现了 NO_2 和死亡率之间的关联。

在对空气质量指数(AQI)的建模研究方面，高燕基于时间序列理论对 AQI 数据进行 ARMA 和 ARIMA 建模，探讨了 AQI 的季节变化的特点[11]；焦东方和孙志华基于 Box-Cox 变化研究了青岛市空气质量状况[12]。

目前，基于时间序列和线性回归观点分析空气质量指数的研究较多，但传统的回归模型需要满足部分假设，在这些假设成立的条件下预测模型才有效，而实际情况中假设往往并不满足。然而，半参数回归模型在某些设置下不需要考虑传统的线性模型的假设条件，因而回归具有稳健性。目前，半参数模型已经在某些领域得到了应用，可参考文献[13]，然而在生态环境领域的应用较少。因此，本文基于惩罚样条的半参数模型研究我国大城市的空气质量状况, 依据时间和空间的不同探究各自的影响因素，并给出适当的政策建议。

二、研究目标与建模思路

(一) 研究目标

目标 1. 选取 4 个有代表性的城市 (北京、上海、深圳和成都), 搜集 AQI 数据及其相应指标, 对数据进行检验和描述；

目标 2. 通过采用半参数回归模型、贝叶斯半参数回归模型、曲线因子模型和半参数分位数回归模型探究按城市分组、按月份分组数据中 AQI 的基本情况；

目标 3. 根据模型得出的结果, 对我国空气污染治理提出合理的政策建议。

(二) 建模思路

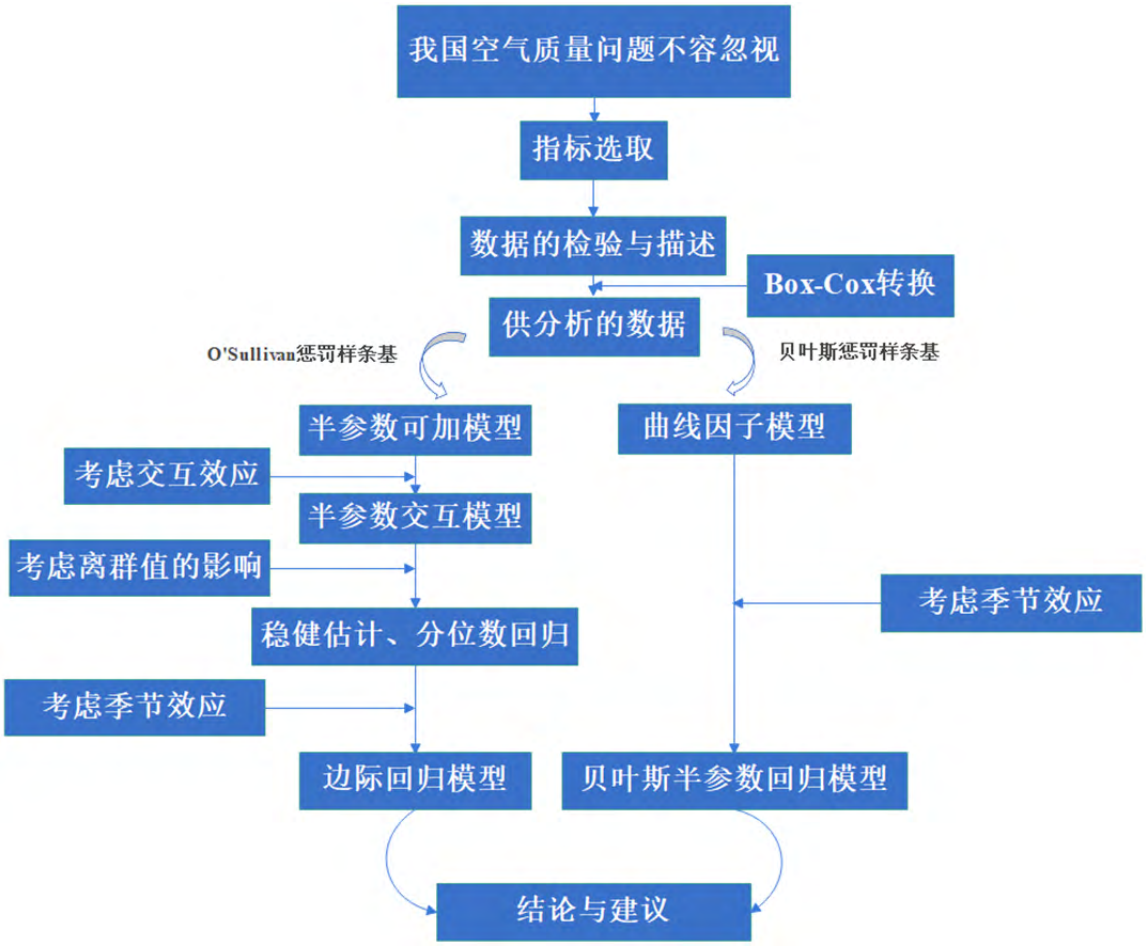

图 2 本文逻辑结构图

三、数据的检验与描述

(一) 数据的检验与 Box-Cox 转换

本文分析的原始数据来源于 https://www.aqistudy.cn/historydata/。变量说明表如下**表 1** 所示。共搜集 4 个具有代表性的城市 (北京、上海、深圳、成都) AQI 日度数据, 样本量共 4216, 原始数据见数据包的“ AQI ”, 分组数据见“ AQIgroup ”。

表 1 变量说明表

变量	符号	单位	搜集区间	单个城市样本总量
空气质量指数	AQI	$\mu g/m^3$		
二氧化氮含量	NO_2	$\mu g/m^3$		
细颗粒物含量	$PM_{2.5}$	$\mu g/m^3$		
可吸入颗粒物含量	PM_{10}	$\mu g/m^3$	2018.7.1-2021.5.19	1054
二氧化硫含量	SO_2	$\mu g/m^3$		
一氧化碳含量	CO	mg/m^3		
臭氧含量	O_3	$\mu g/m^3$		

由于原始数据不存在缺失，因而不需要进行插补处理。但包含 3 个 AQI 为 0 的点，具体为北京市 2018 年 12 月 3 日数据点、成都市 2018 年 12 月 4 日和 5 日数据点，我们将其删除。对未分组的原始数据即四个城市数据的 AQI 集合进行正态性检验，QQ 图和直方图如**图 3**。结果显示，AQI 具有右偏特征，显然是没有通过正态性检验的。由于在半参数模型的参数部分依然需要假设响应变量的正

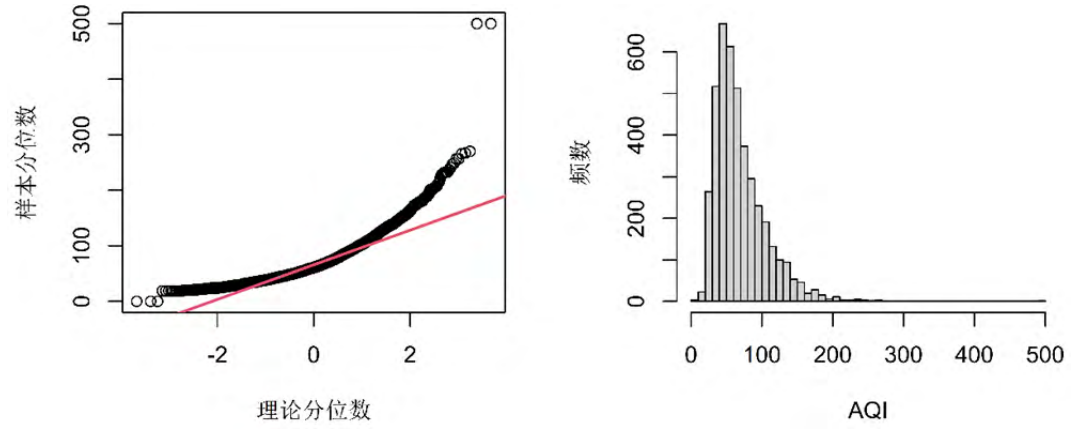

图 3 原始数据的正态性检验

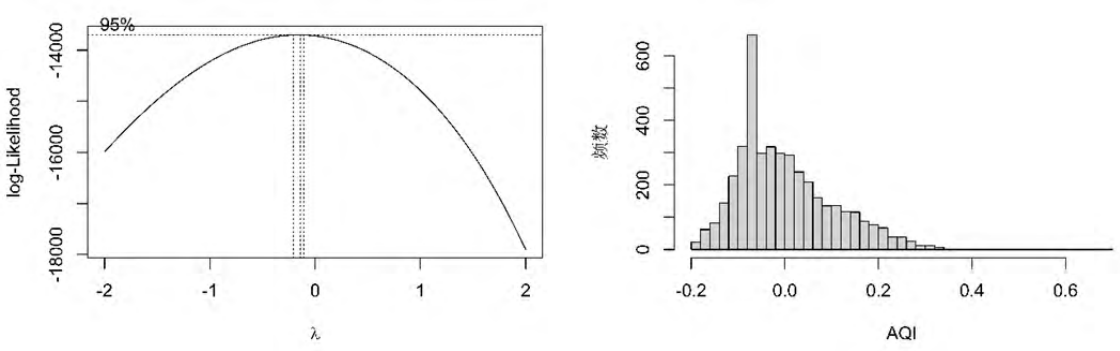

图 4 Box-Cox 转换结果

态性，因此我们需要对原始数据进行 Box-Cox 转换以降低数据偏度，使其服从正态分布或近似服从正态分布。

对数据进行 Box-Cox 转换(R 文件为 **Box-Cox**, 或程序文档 **Code 1**)的结果如**图 4**，左面板中，通过选择对数似然最大的转换参数值 0.14141，得到了右面板的转换结果的直方图。结果显示，数据的右偏得到了很好的修正，尽管数据依然不是严格的正态分布，但已经近似服从正态分布。在下一节中，我们将对转换后的数据进行进一步的描述。

(二) 数据的描述

在本小节中，我们对数据进行描述性分析，以初步了解我国 AQI 的基本情

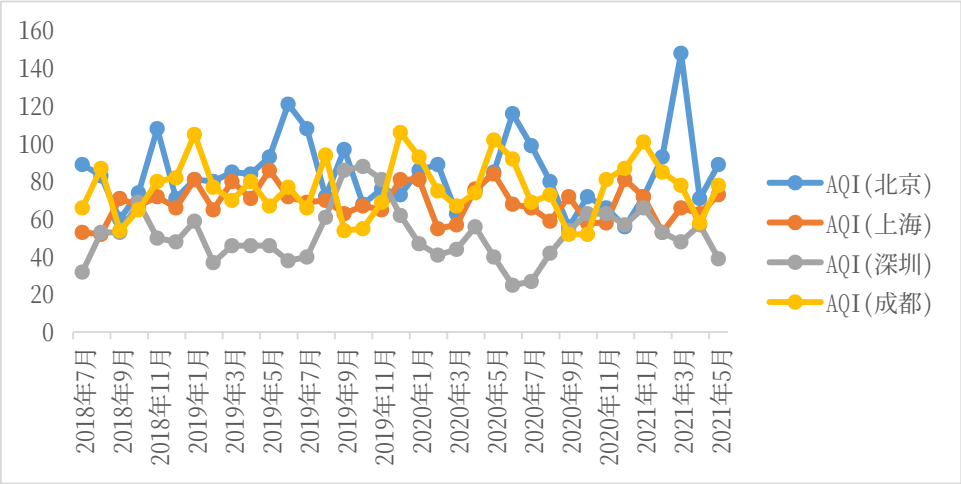

图 5 按城市分组的 AQI(年度数据)对比

况。为了查看近年来 AQI 的变化趋势，我们将日度数据求平均获得年度数据，AQI 的变化情况如图 5。容易看出，从 2018 年 7 月至 2021 年 5 月，北京市 AQI 高于其他三个城市，尤其是 2021 年 5 月受沙尘天气的影响，AQI 达到了近年来最高值，空气质量令人担忧。深圳市的空气质量最好，AQI 值多年保持最低；上海市 AQI 值较稳定，波动较小；成都市空气质量较上海和深圳差。

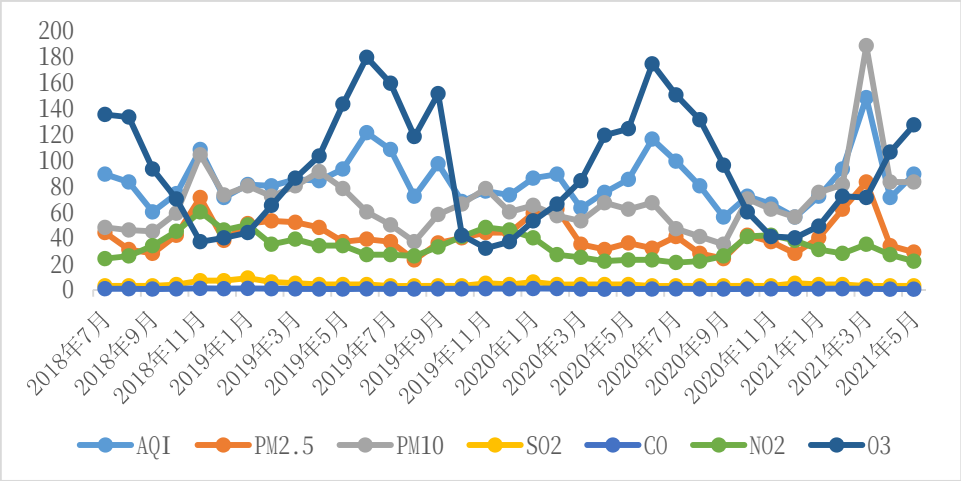

图 6 北京市 AQI(年度数据)与各指标变化趋势

此外，我们以北京市为例，展示了北京市 AQI(年度数据)与各指标的变化趋势如图 6。结果显示，各指标与 AQI 变化趋势大致相同。正如第一章所介绍的，目前，SO2 对空气质量的影响已经大大减小且目前由于机动车尾气排放加剧，NO_2 已经成为影响 AQI 的重要因素。同时考虑到 $PM_{2.5}$ 与 PM_{10} 对 AQI 的影响效果类似。通过对图 6 的解读，我们考虑使用 $PM_{2.5}$ 和 NO_2 同时结合季节和地域对我国 AQI 进行半参数建模的实证分析。

四、基于半参数建模下 AQI 的实证分析

(一) 三次 O'Sullivan 惩罚样条

惩罚样条平滑是一种拟合曲线到散点图的方法, 是半参数回归的主要组成部分 (半参数回归由参数回归和非参数回归两部分组成)。参数回归已很常见和周知, 我们将在后面的半参回归模型中进行简单介绍。现在我们讨论非参数回归模型

$$y_i = f(x_i) + \varepsilon_i, i = 1, \ldots, n, \tag{1}$$

其中协变量和响应变量的观测为(x_i, y_i)。这里的 f 是平滑的任意函数，ε_i 是满足 $E(\varepsilon_i) = 0$ 且独立的随机变量。关于 f 的一个简单的惩罚样条模型可以表示为

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k (x - \kappa_k)_+ \tag{2}$$

其中对于任何 $x \in R$、$\kappa_k s$ 和 $1 \le k \le K$，$x_+ = \max(x, 0)$ 是预先设定的值，我们通常取为关于 x 的分位数近似相等的间隔。K 通常取 10 到 50 之间的某个数字。式(2)的函数是在结 $\kappa_k s$ 处"绑在一起"的分段线。根据下列式(3)的约束优化问题估计系数 β_0、β_1 和 u_k，

$$\min imize\left[\sum_{i=1}^{n}\{y_i - f(x_i)\}^2 + \lambda \sum_{k=1}^{K} u_k^2\right] \tag{3}$$

实际上，这是一个平滑参数 $\lambda > 0$ 的惩罚最小二乘，即

$$\hat{f}(x; \lambda) = \hat{\beta}_0 + \hat{\beta}_1 x + \sum_{k=1}^{K} \hat{u}_k (x - \kappa_k)_+$$

其中 $\hat{\beta}_0$、$\hat{\beta}_1$ 和 $\hat{u}_k$ 由最小化式(3)得到。平滑参数 λ 对 $\hat{f}(\cdot; \lambda)$ 的影响和平滑参数的选择方法已经有相当多的文献进行了研究。上式所表示的线性基函数虽然简单直

观，但也有一些缺点，比如：回归函数拟合由于 f 是分段线性的限制而产生人为的打结 ;基函数是无界且非正交的, 这可能会导致数值计算时出错的问题。因此，使用 $\{(x-\kappa_k)_+\}^p$ 代替 $(x-\kappa_k)_+$ 是较好的选择，且 p 常常取 3，因为这是连续二阶导数的分段三次拟合，且通过以 B 样条为标准选择的截断多项式基函数的线性变换来实现了数值稳定性。

下面，我们使用三次 O'Sullivan 样条基函数(R 文件 **OSullivan-spline-basis**, 或程序文档 **Code 2**)对 AQI 进行惩罚估计，**图 7** 展现了三次 O'Sullivan 样条基函数下 AQI 的惩罚估计，其中 λ=100 且 K=20。紫色的菱形物表示了结所在的位置。**图 8** 为**图 7** 的放大视角，垂直范围设置为区间 (-40, 40)。可以很容易看出，基于三次 O'Sullivan 样条基函数的估计与普通的线性基函数的估计是不同的，O'sullivan 惩罚样条具有作为平滑样条的自然泛化的吸引力，几乎不受高样本量的影响[14]。由于此次分析的数据量较大，因此我们相信该基函数能在 AQI 分析中起到良好的效果。

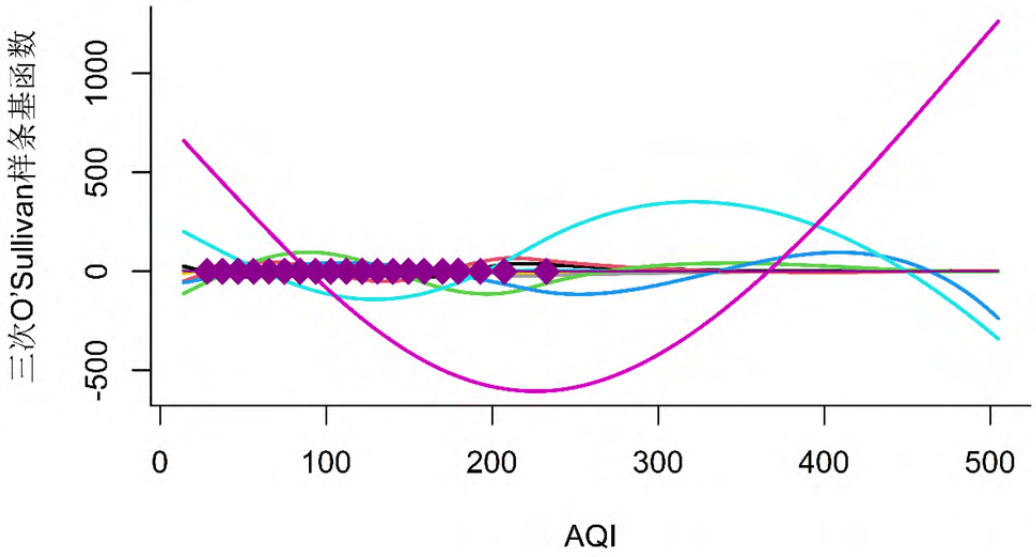

图 7 三次 O'Sullivan 样条基函数下 AQI 的惩罚估计

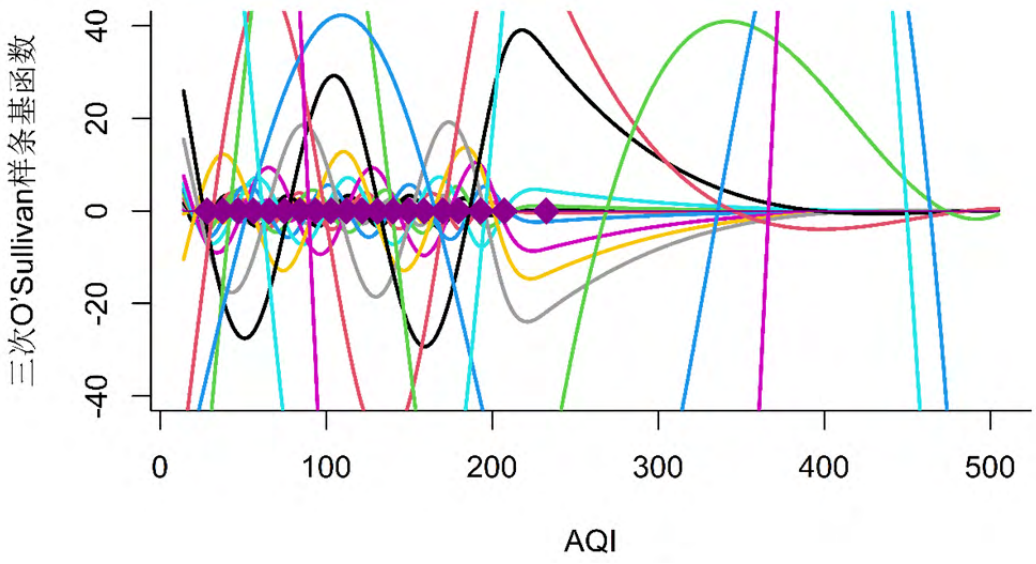

图 8 图 7 的放大视角

(二) 贝叶斯惩罚样条

惩罚样条也可以采用贝叶斯方法进行拟合，我们称之为贝叶斯惩罚样条。与频率的方法相比，贝叶斯方法的优点包括考虑了与方差成分相关的不确定性，以及处理异方差和数据缺失等复杂问题的能力。本文也将从基于贝叶斯的观点对我国 AQI 进行半参数建模。

贝叶斯惩罚样条非参数回归的形式为:

$$y_i \mid \beta, u, \sigma_\varepsilon^2 \overset{ind}{\sim} N(\beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u_K z_k(x_i), \sigma_\varepsilon^2), i = 1, \ldots, n,$$

$$u \mid \sigma_u \overset{ind}{\sim} N(0, \sigma_u^2 I), \beta_0, \beta_1 \overset{ind}{\sim} N(0, \sigma_\beta^2) \tag{4}$$

$$\sigma_u \sim Half - Cauchy(A_u), \sigma_\varepsilon \sim Half - Cauchy(A_\varepsilon)$$

其中 u 是包含 u_k 的 $K \times 1$ 向量。符号 $x \sim Half - Cauchy(A)$ 表示 x 有半柯西分布的密度函数，即

$$p(x) = 2/[A\pi\{1 + (x/A)^2\}], x > 0,$$

其中 $A > 0$ 是尺度参数。实际上，如果不对上述参数施加其先验分布，则贝叶斯

惩罚样条退化为混合效应惩罚样条。模型(4)的超参数为$\sigma_\beta > 0, A_u > 0, A_\varepsilon > 0$。这些超参数需要人为指定。对于一般的贝叶斯分析，明智的选择是根据数据特点指定合适的超参数。然而，对于模型(4)及其各种半参数回归模型来说，情况则发生了变化，即，一般推荐无信息先验。在本文中，我们将通过将σ_β、A_u和A_ε等超参数设置为较大的正数来加强无信息先验，本次分析我们采用了迭代 2000 次，丢弃前面 1000 次，通过后 1000 次迭代计算贝叶斯估计。

分别使用 AQI 数据的 NO_2 和 $PM_{2.5}$ 作为协变量，AQI 作为响应变量，通过使用贝叶斯惩罚样条对上述模型进行拟合(R 文件 **bayes-spline**，或程序程序文档 **Code 3**)，结果如**图 9 至图 12**。图 8 和图 10 分别显示了 NO_2 和 $PM_{2.5}$ 对我国 AQI 的非参数拟合，三条虚线分别为 0.25、0.5 和 0.75 分位点，浅绿色的代表了估计值的 95%置信区间，蓝色小圆点代表了样本点，由**图 9** 和**图 11**，可以得到下列 3 条结果：

1. 图 8 和图 10 中，代表 0.25、0.5 和 0.75 分位点的三条虚线的间距很近尤其是图 10。表明大多数被估计的 AQI 值较小，即每年的大多数时候我国的空气质量较好。

2. 95%置信区间在 NO_2 含量大于 80 μg/m^3, $PM_{2.5}$ 含量大于 100 μg/m^3 时逐渐增大。

3. 需要注意，左上角的两个小蓝点表示了 2021 年 3 月北京市遭受强沙尘天气时的 AQI。在**图 9** 中显示此时的 NO_2 含量是低于 0.25 分位点的；然而图 10 显示了此时的 $PM_{2.5}$ 已经超过了 0.95 分位点。

图 10 和**图 12** 显示了贝叶斯惩罚样条估计的结果摘要。第 2 列至第 4 列分别表示了算法的收敛，第 5 列显示了参数的后验分布的密度估计，最后一列则表示了参数的点估计及其 95%的置信区间估计。自相关图 (除第一个参数之外) 都表现出了迅速的结尾，表明这几个参数的收敛速度较快。但最终由 Brooks–Gelman–Rubin (RGB) 显示了所有参数的收敛性，同时路径图都显示了快速的变化。综合表明贝叶斯惩罚样条取得了良好的效果。

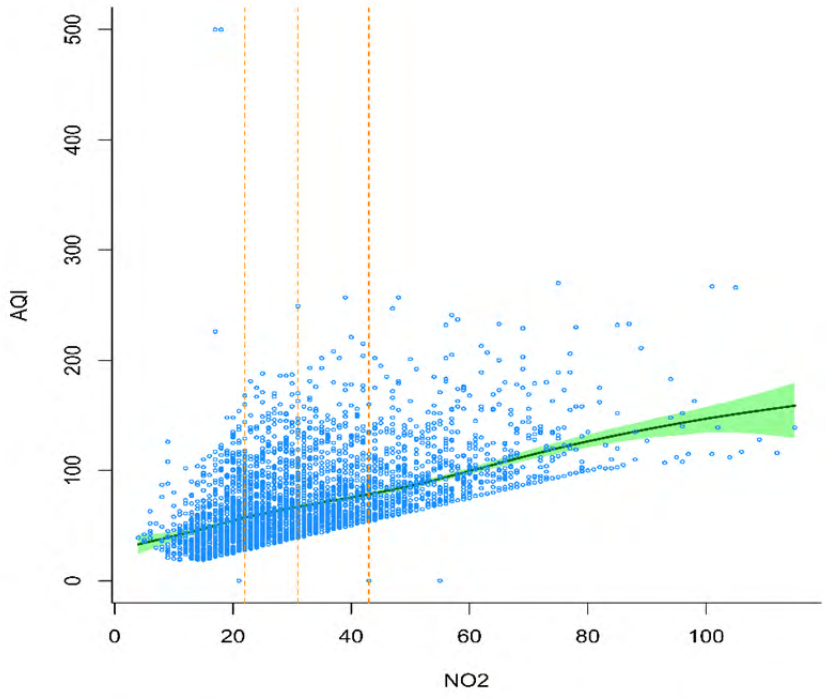

图 9 贝叶斯惩罚样条下 NO_2 对 AQI 的非参数拟合及 95%置信区间

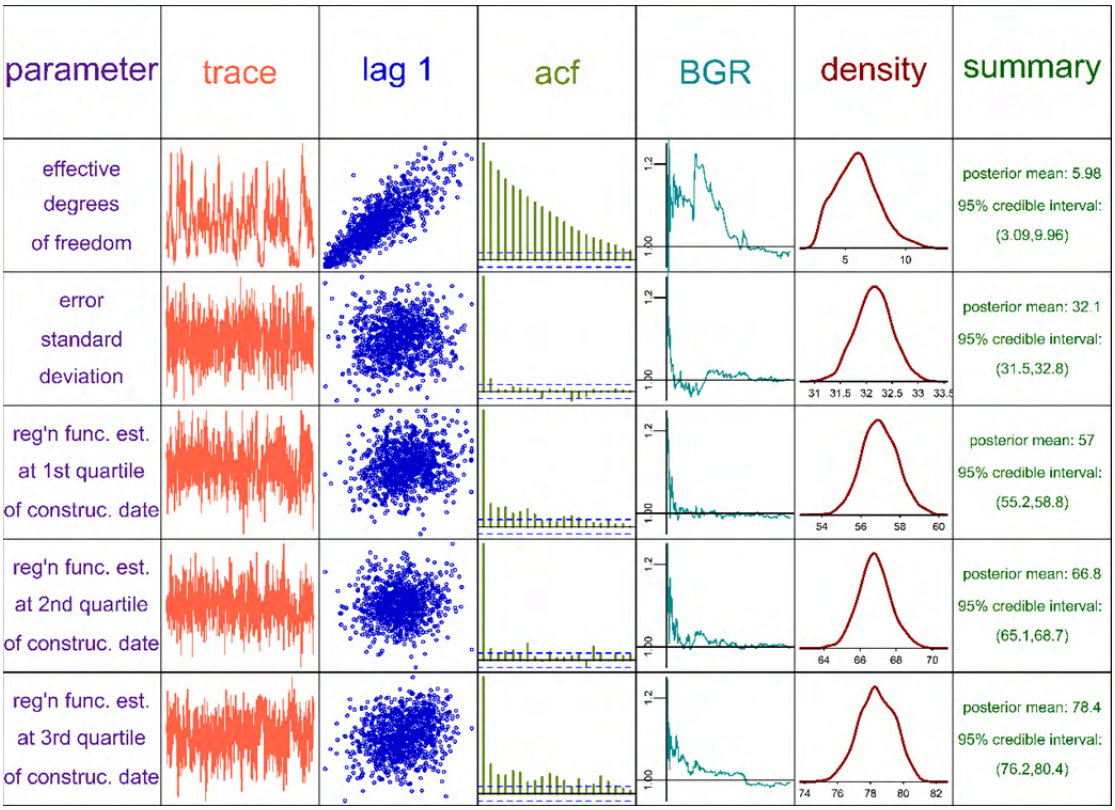

图 10 贝叶斯惩罚样条下 NO_2 对 AQI 的拟合摘要

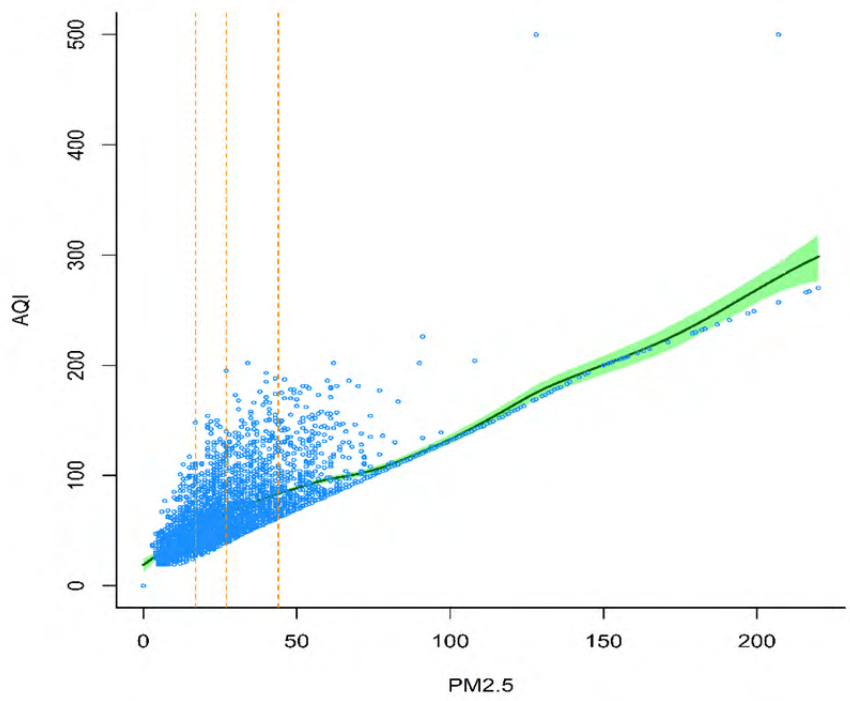

图 11 贝叶斯惩罚样条下 $PM_{2.5}$ 对 AQI 的非参数拟合及 95%置信区间

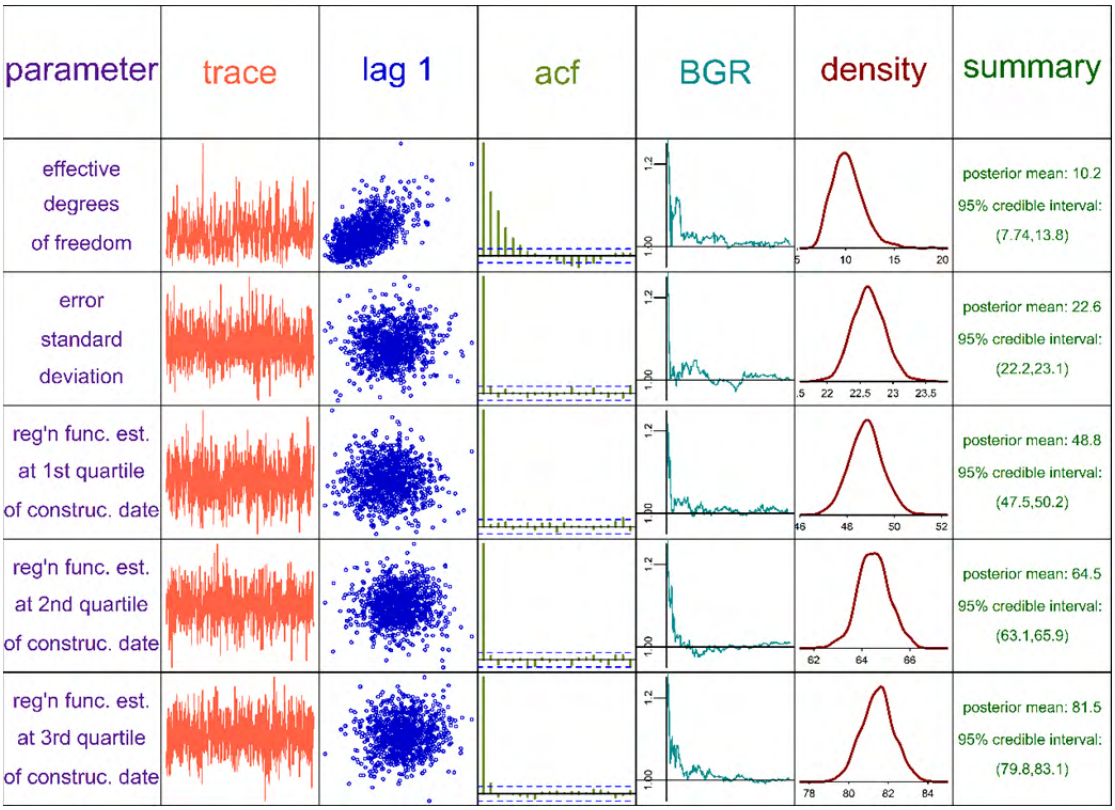

图 12 贝叶斯惩罚样条下 $PM_{2.5}$ 对 AQI 的拟合摘要

(三) 半参数可加模型

我们整理的数据集中包含一个名为 city 的分类变量, 该变量表示了我们所搜集的数据所在的城市为北京、上海、深圳和成都中的一个。可以提出的一个问题是，AQI 与 NO_2 和 $PM_{2.5}$ 的关系是否因地区而异。这可以通过扩展非参数回归模型(1)来达到，使每个城市(北京、上海、深圳、成都)都有各自的截距，即

$$AQI_i = \beta_1 I(shanghai_i) + \beta_2 I(shenzhen_i) + \beta_3 I(chengdu_i) + f(NO_{2_i}) + \varepsilon_i, \quad (5)$$

其中， $f()$ 为一个非参数回归， $I()$ 表示一个指示函数，为

$$I(shanghai) = \begin{cases} 1, \text{如果第} i \text{个} AQI \text{是来自于上海市的样本}, \\ 0, \text{否则}, \end{cases} \quad (6)$$

式(6)中的定义也类似地适用于 $I(shenzhen)$ 和 $I(chengdu)$，如果 $I(shanghai) = I(shenzhen) = I(chengdu) = 0$，则表明数据是来自北京市的样本。$PM_{2.5}$ 对 AQI 的拟合有类似的形式。拟合(R 文件 **semiadd**，或程序文档 **Code 4**)的结果为，

表 2 半参数可加模型的拟合结果

Formula	Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
Formula 1	Intercept	84.4801	0.9546	88.500	<2e-16 ***
	factor(city)chengdu	-13.7166	1.3799	-9.940	<2e-16 ***
	factor(city)shanghai	-21.0440	1.3626	-15.440	<2e-16 ***
	factor(city)shenzhen	-22.4551	1.3767	-16.310	<2e-16 ***
Formula 2	Intercept	76.8432	0.6987	109.983	<2e-16 ***
	factor(city)chengdu	-8.7003	0.9839	-8.843	<2e-16 ***
	factor(city)shanghai	-6.5554	0.9933	-6.600	4.63e-11 ***
	factor(city)shenzhen	-11.4243	1.0117	-11.292	<2e-16 ***

其中 Formula 1 为 AQI ~ factor(city) + s(NO_2, bs = "cr", k = 38)；Formula 2 为 AQI ~ factor(city) + s($PM_{2.5}$, bs = "cr", k = 38)。

从**表 2** 的输出结果中我们可以清晰地看到，两种拟合中，上海、深圳和成都市都与北京市的 AQI 有显著的不同。可以合理地得出结论，在 NO_2 含量对 AQI

的拟合中，上海市 AQI 平均值比北京市平均低 21.0440，而深圳和成都市的 AQI 平均值分别比北京市平均低 22.4551 和 13.7166，这与**图 5** 的描述性分析的结果完全一致，尽管图 4 没有给出具体的数值。由 Formula 2 可以得到类似的结果。

此外，我们还展示了以地域不同的拟合结果图及其子图。如**图 13 至图 16**。

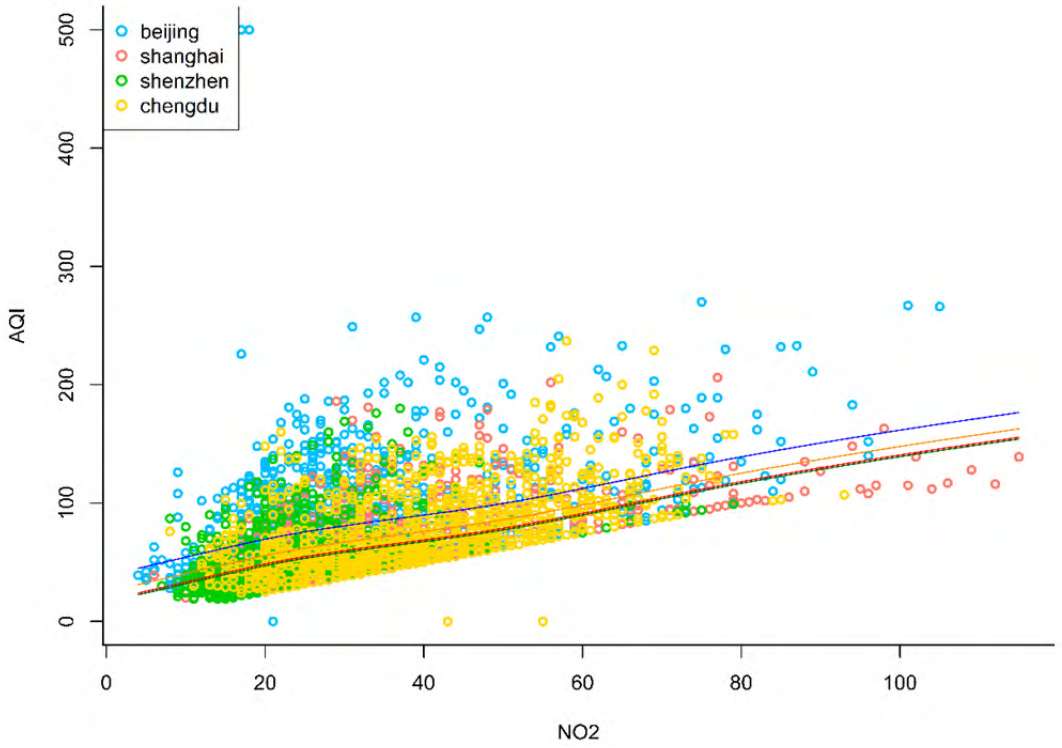

图 13 以城市为因子的半参数可加模型下 NO_2 对 AQI 的拟合结果

(四) 半参数交互模型

上述的半参数可加模型并没有考虑各个城市间数据的交互效应，接下来我们考虑一个更复杂的模型，即允许每个区域有自己的线性形式，并且仍然保持 AQI 与 NO_2 和 $PM_{2.5}$ 之间的半参数关系相同。这样一来，变量 AQI 与 NO_2 和 $PM_{2.5}$ 的观测可以以参数方式进行交互。模型可以写为

$$\begin{aligned} AQI_i = {} & f\left(NO_{2_i}\right) + \beta_1 I\left(shanghai_i\right) + \beta_2 I\left(shenzhen_i\right) + \beta_3 I\left(chengdu_i\right) \\ & + \beta_4 I\left(shanghai_i\right) \times NO_{2_i} + \beta_5 I\left(shenzhen_i\right) \times NO_{2_i} \\ & + \beta_6 I\left(chengdu_i\right) \times NO_{2_i} + \varepsilon_i \end{aligned} \tag{7}$$

拟合的结果（R 文件 **semiint**，或程序文档 **Code 5**）如**表 3**、**图 17** 和**图 18** 所示。

这

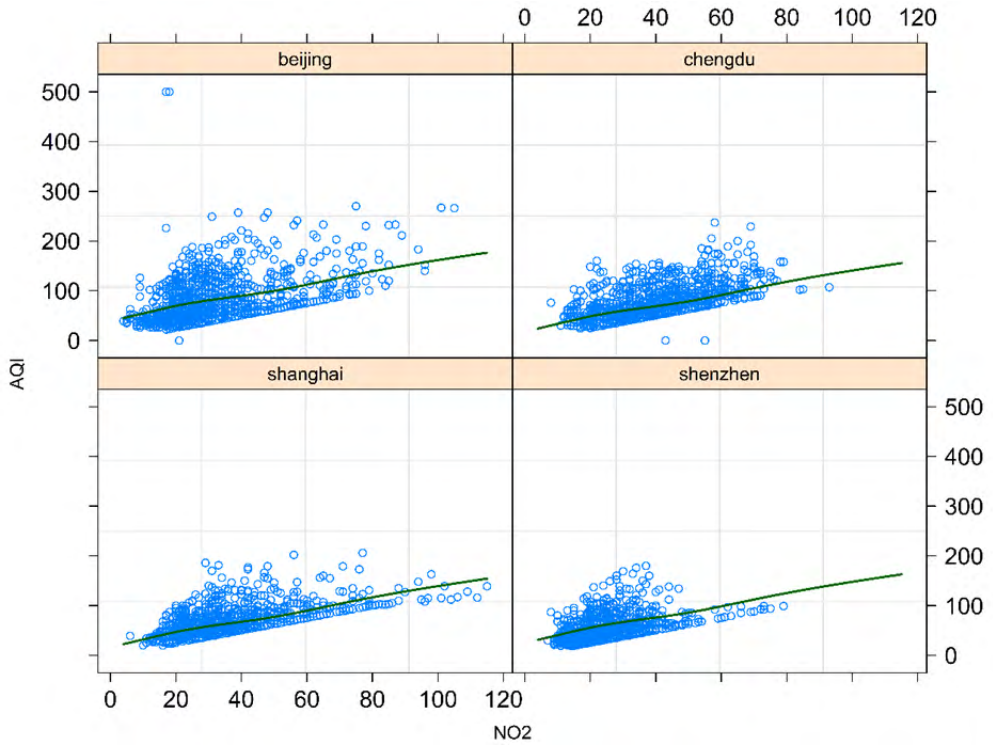

图 14 以城市为因子的半参数可加模型下 NO_2 对 AQI 的拟合结果子图

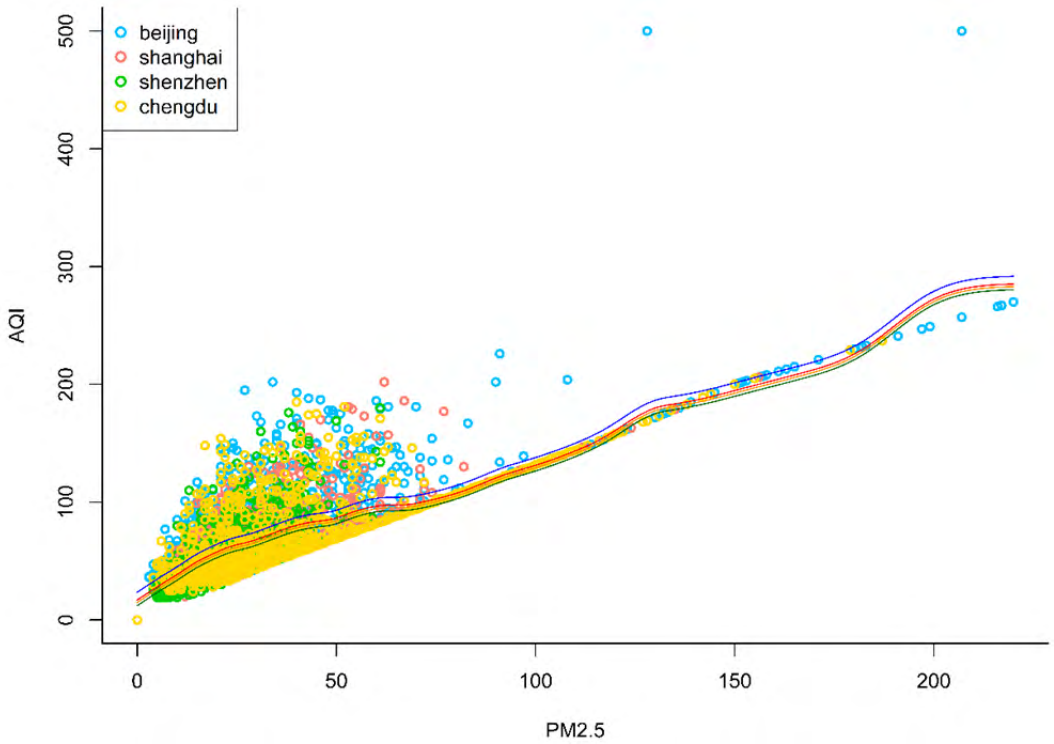

图 15 以城市为因子的半参数可加模型下 $PM_{2.5}$ 对 AQI 的拟合结果

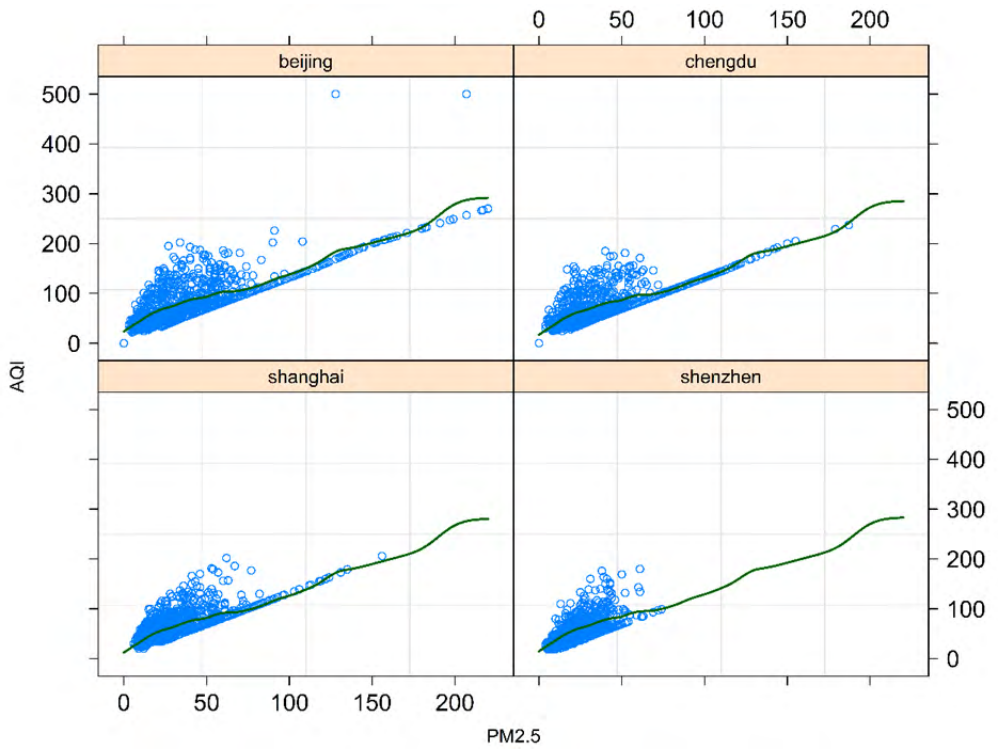

图 16 以城市为因子的半参数可加模型下 $PM_{2.5}$ 对 AQI 的拟合结果子图

里我们仅展示了拟合图，限于篇幅没有再展示其子图。如图 17，在 NO_2 的含量方面，北京市在 AQI 的低值和高值处都具有很高的 NO_2 含量，而深圳市和上海市的 NO_2 含量低值和高值处都较低。尤其需要注意的是成都市，低值时的 AQI 对应了最低的 NO_2 含量，然而高值时的 AQI 却对应了最高的 NO_2 含量，说明影响成都市 AQI 的重要因素是空气中 NO_2 的含量高低。由于空气中的 NO_2 主要来自于机动车尾气排放，因此，在提高空气质量方面，成都市应加大机动车治理，包括限号行驶和推广新能源汽车。图 18 中，在 $PM_{2.5}$ 的含量方面，北京市在低值 AQI 时拥有最大的 PM2.5 含量，高值 AQI 时 $PM_{2.5}$ 含量排名第 2，表明 PM2.5 是影响北京市空气质量的主要因素，因此，加强北方风沙治理，植树造林。通过，与北京市临近城市的工业也是导致北京市高 $PM_{2.5}$ 含量的一个因素，因此，工业技术的升级也是改进空气质量的重要方面。(注：在表 3 中，Formula 1 为 AQI ~ factor(city) * NO_2+ s(NO_2, bs = "cr", k = 37)；Formula 2 为 AQI ~ factor(city) * $PM_{2.5}$ + s($PM_{2.5}$, bs = "cr", k = 37))。

表 3 半参数交互模型的拟合结果

Formula	Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
Formula 1	Intercept	42.6762	2.7275	15.646	<2e-16 ***
	factor(city)chengdu	-29.2305	3.5922	-8.137	5.27e-16 ***
	factor(city)shanghai	-15.2667	3.2583	-4.685	2.88e-06 ***
	factor(city)shenzhen	-23.7037	3.4286	-6.914	5.44e-12 ***
	NO_2	1.2222	0.0753	16.228	<2e-16 ***
	factor(city)chengdu: NO_2	0.4039	0.0893	4.521	6.32e-06 ***
	factor(city)shanghai: NO_2	-0.1444	0.0815	-1.771	0.0767
	factor(city)shenzhen: NO_2	0.0297	0.1154	0.257	0.797
Formula 2	Intercept	38.8743	1.107	35.116	<2e-16 ***
	factor(city)chengdu	-7.0703	1.7377	-4.069	4.81e-05 ***
	factor(city)shanghai	-6.9405	1.7489	-3.969	7.35e-05 ***
	factor(city)shenzhen	-18.7986	1.9738	-9.524	< 2e-16 ***
	PM2.5	1.0961	0.0214	51.281	< 2e-16 ***
	factor(city)chengdu: $PM_{2.5}$	-0.0301	0.0345	-0.874	0.382
	factor(city)shanghai: $PM_{2.5}$	0.0176	0.0412	0.427	0.669
	factor(city)shenzhen: $PM_{2.5}$	0.3262	0.0697	4.683	2.92e-06 ***

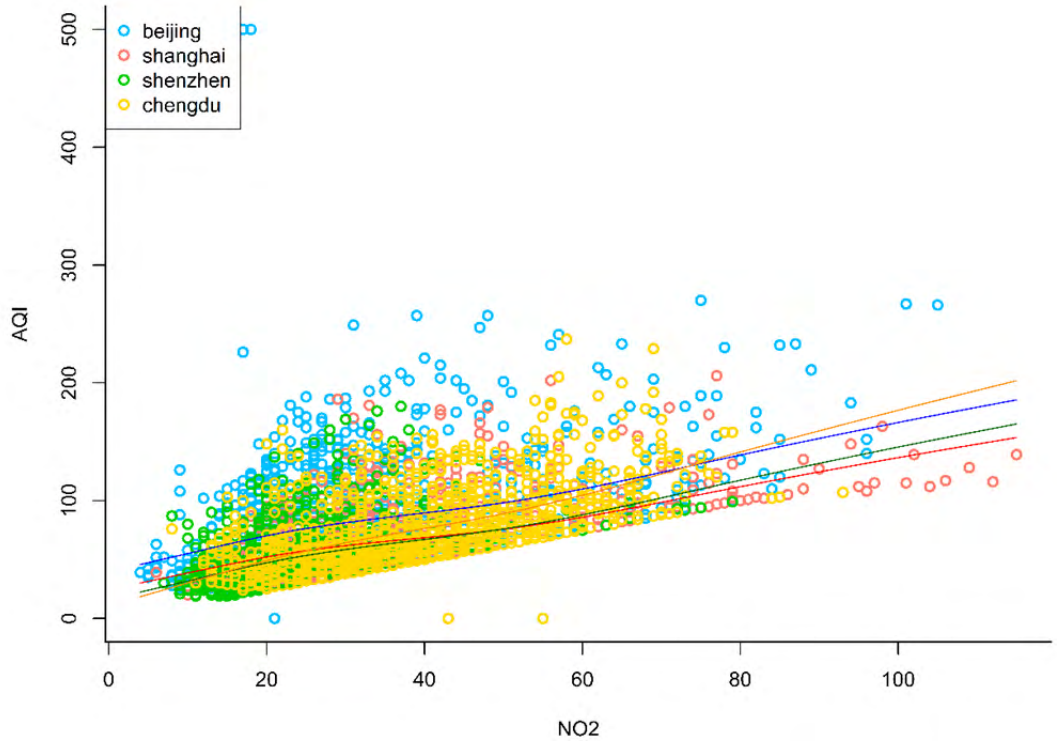

图 17 以城市为因子的半参数交互模型下 NO_2 对 AQI 的拟合结果

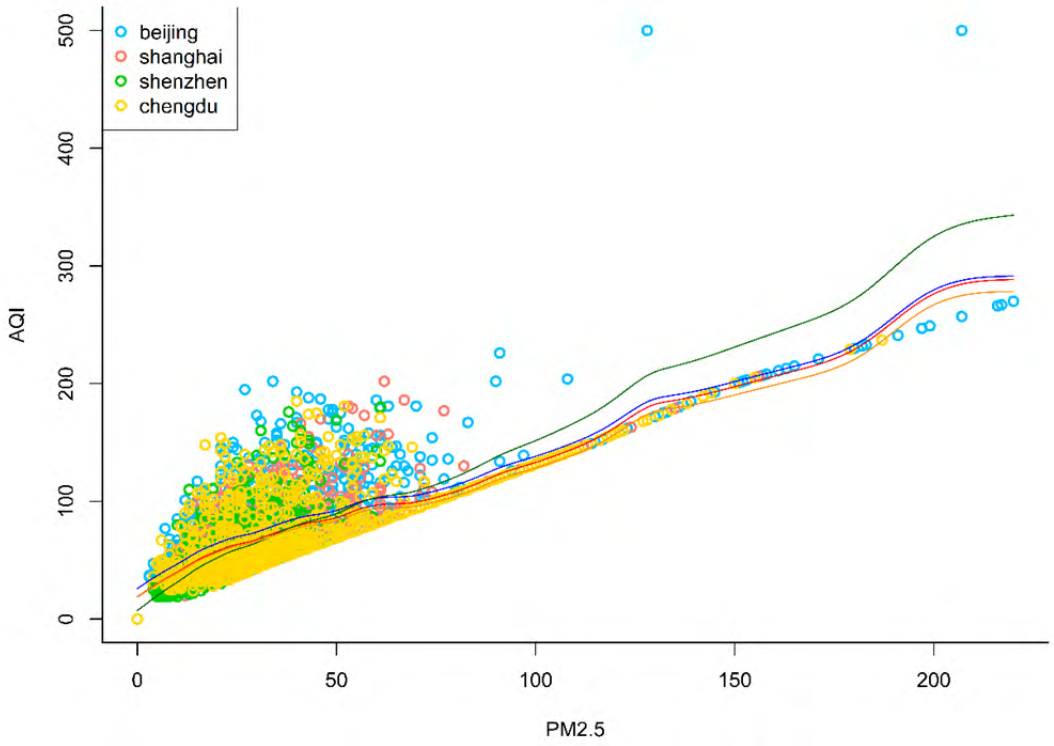

图 18 以城市为因子的半参数交互模型下 $PM_{2.5}$ 对 AQI 的拟合结果

(五) 贝叶斯半参数曲线因子模型

上述模型研究了各城市之间的 AQI 水平及 NO_2 和 $PM_{2.5}$ 对各城市 AQI 影响程度的不同。但某些情况下，我们需要研究某城市 AQI 与全国其他城市相比的水平，我们以深圳市对比全国为例。则我们可以使用到半参数曲线因子模型。模型的形式如下：

$$AQI_i = f_{1-I(shenzhen_i)}\left(NO_{2_i}\right) + \varepsilon_i, \tag{8}$$

其中

$$f_0(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_{0k} z_k(x)$$

$$f_1(x) = \beta_0 + \beta_0^{contrast} + (\beta_1 + \beta_1^{contrast})x + \sum_{k=1}^{K} u_{1k} z_k(x)$$

其中$u_{0k} \overset{ind}{\sim} N(0,\sigma_{u0}^2), u_{1k} \overset{ind}{\sim} N(0,\sigma_{u1}^2), \varepsilon_i \overset{ind}{\sim} N(0,\sigma_{\varepsilon}^2), 1 \le i \le n$，实际上，它附带了两个无约束的惩罚样条，一个用于深圳市，一个用于全国其他地区。对比函数

$$c(x) = f_1(x) - f_0(x) = \beta_0^{contrast} + \beta_1^{contrast} x + \sum_{k=1}^{K} (u_{1k} - u_{0k}) z_k(x) \tag{9}$$

表示了深圳市以外城市和深圳市的 AQI 之差。

拟合的结果(R 文件 **curvefactor**，或程序文档 **Code 6**)如**图 19 至图 22**。在图 19 和图 21 中，根据样本点是否属于深圳市，用不同的颜色进行标记。根据贝叶斯半参数曲线因子模型对数据进行惩罚样条拟合。红色实线是 f_0 的贝叶斯估计，蓝色实现是 f_1 的贝叶斯估计，虚线是估计值的 95%置信区间。拟合是基于马尔可夫链蒙特卡罗抽样。图 19 和图 21 中，曲线是对比函数 $c(x)$ 的估计，绿色阴影为估计值的 95%置信区间。由**图 19** 和**图 21** 可以看出，深圳市 AQI 空气中 NO_2 含量低于全国其他城市，表明深圳市空气质量更好。而 $PM_{2.5}$ 与全国其他大城市相比几乎相同。从**图 20** 和**图 22** 中可以得出相同的结果。

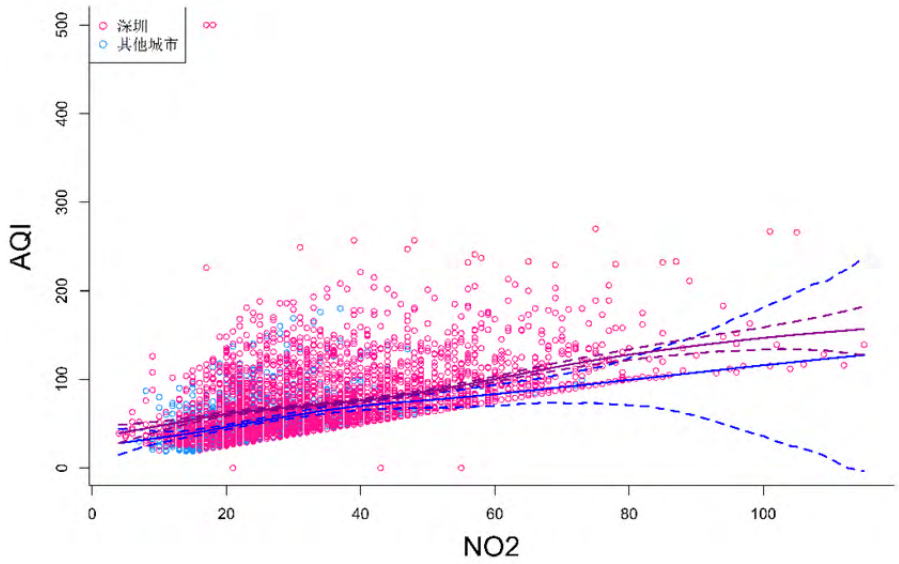

图 19 深圳市 NO_2 含量对 AQI 的贝叶斯半参数曲线因子模型拟合(对比全国)

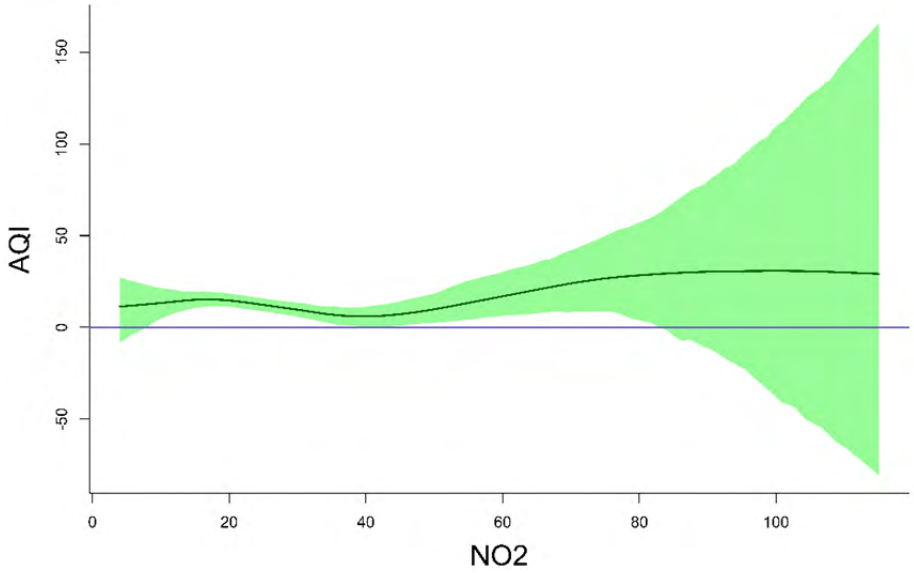

图 20 图 19 中对比函数的估计值及其 95%置信区间

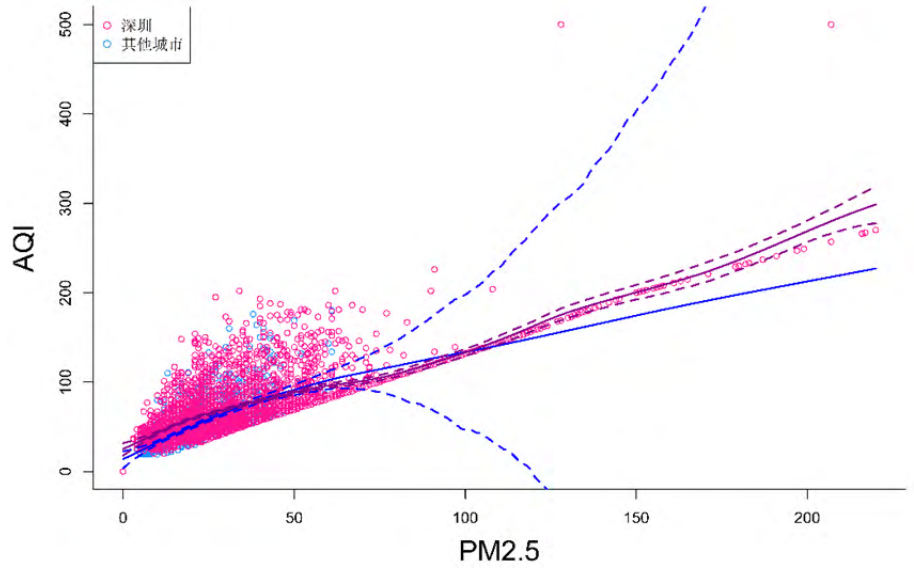

图 21 深圳市 $PM_{2.5}$ 含量对 AQI 的贝叶斯半参数曲线因子模型拟合(对比全国)

(六) 稳健估计和半参分位数回归

众所周知，当数据中存在离群值时，基于最小二乘的回归方法容易得到不稳定的估计。降低离群值影响的方法通常被称为稳健的。由于分位数回归的对象是分位数而不是均值，因此分位数回归也可以被归为稳健回归的一种。

很容易注意到，AQI 的原始数据是右偏的，尽管在经过 Box-Cox 转换后偏度得到了修正，但并未完全消除。众所周知，数据偏斜时，若仍然使用正态误差

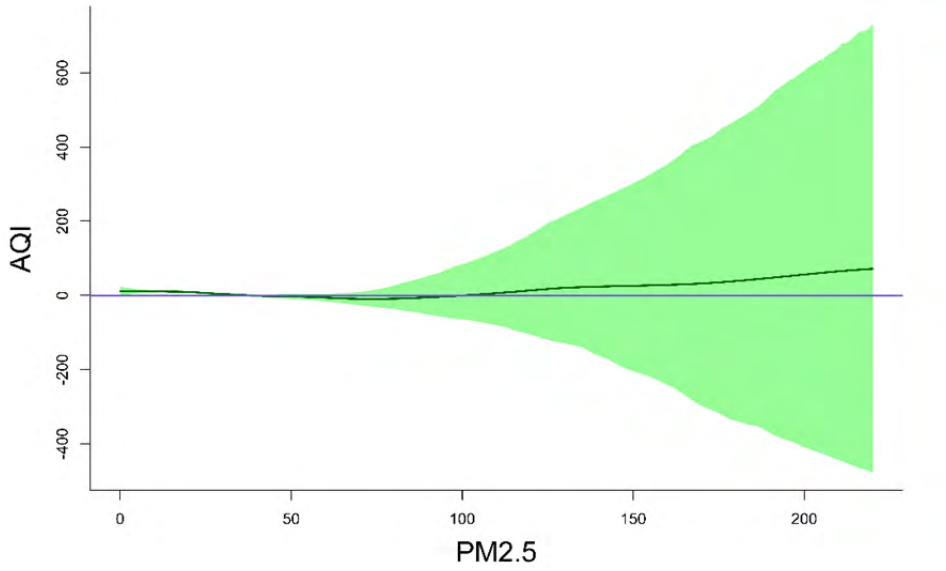

图 22 图 21 中对比函数的估计值及其 95%置信区间

的回归模型进行拟合，将有更多的数据点被认为是离群值。因此，本文的 AQI 研究中应当使用稳健回归的方法。

实际上, 稳健半参数回归的一种方法是将响应变量建模为重尾分布如 t 分布, 可根据调整自由度使其具有任意重尾。与 t 分布相对应的似然会在回归拟合上产生一种稳健性。注意具有位置参数 μ，尺度参数 σ > 0，自由度 ν > 0 的 t 分布具有下列的密度函数

$$p(x;\mu,\sigma,\nu)=\frac{\Gamma(\frac{\nu+1}{2})}{\sigma\sqrt{\pi\nu}\Gamma(\nu/2)[1+\{(x-\mu)/\sigma\}^2/\nu]^{\frac{\nu+1}{2}}}, \tag{10}$$

对于具有这种密度函数的随机变量 x，我们记为 $x\sim t(\mu,\sigma,\nu)$。则全国 AQI 数据基于半参数稳健回归(R 文件 **robusress**，或程序文档 **Code 7**)的拟合结果见**图 23** 和**图 24**。其中绿色实线为非稳健估计，红色实线为稳健估计，且左面板为全局视角，右面板为放大视角。由**图 23** 和**图 24** 可知，AQI 值一定时，非稳健估计都低估了 NO_2 和 $PM_{2.5}$ 的含量，尽管在某些值处这种低估非常小，但仍然不可忽视。

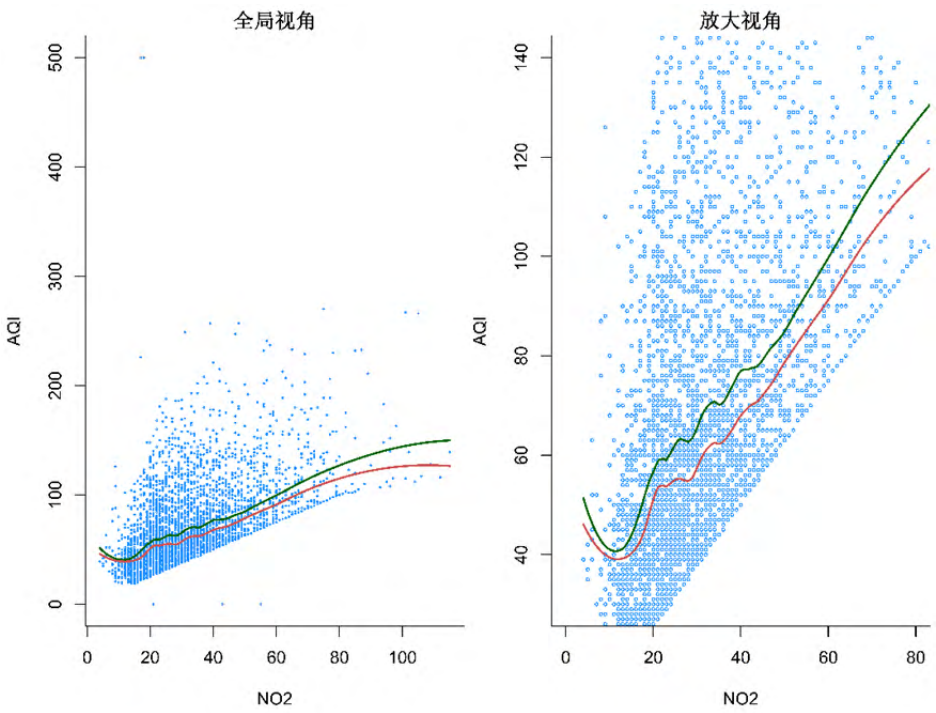

图 23 NO_2 对 AQI 数据的稳健回归结果

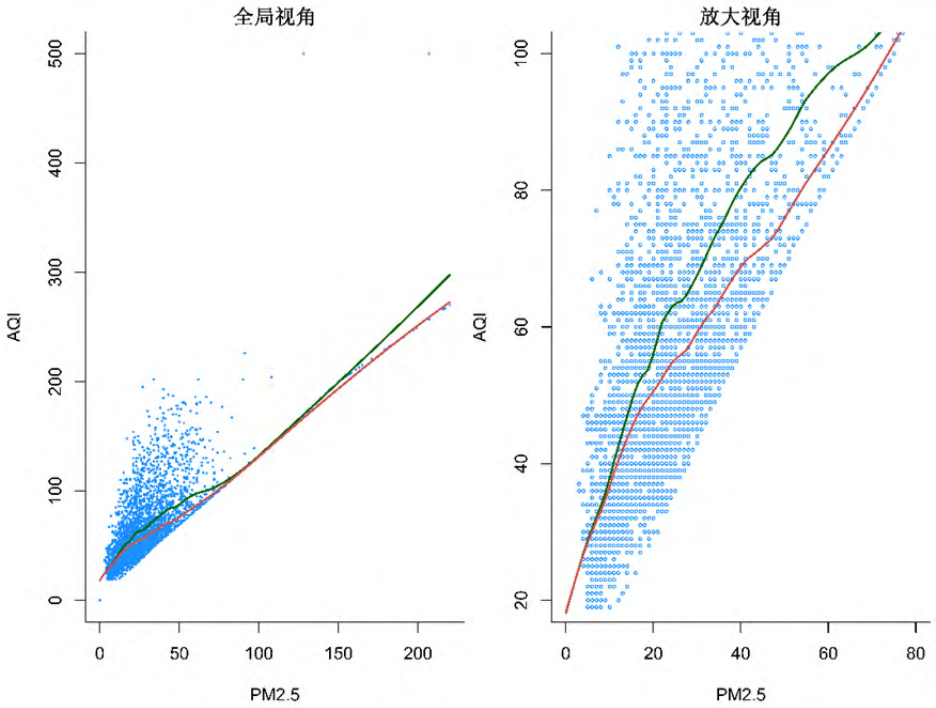

图 24 PM2.5 对 AQI 数据的稳健回归结果

限于篇幅，半参数分位数回归(R 文件 **semiqss**，或 **Code 8**)的理论我们不再详述，仅给出结果如**图 25** 和**图 26**。结果显示，0.01,、0.05、0.25、0.5 和 0.75

分位数间隔很小，表明大的 AQI 值较少，同时很容易得到 NO_2 和 $PM_{2.5}$ 含量对 AQI 的不同分位数的影响是不同的，即使 AQI 集中在值较小的区域内。

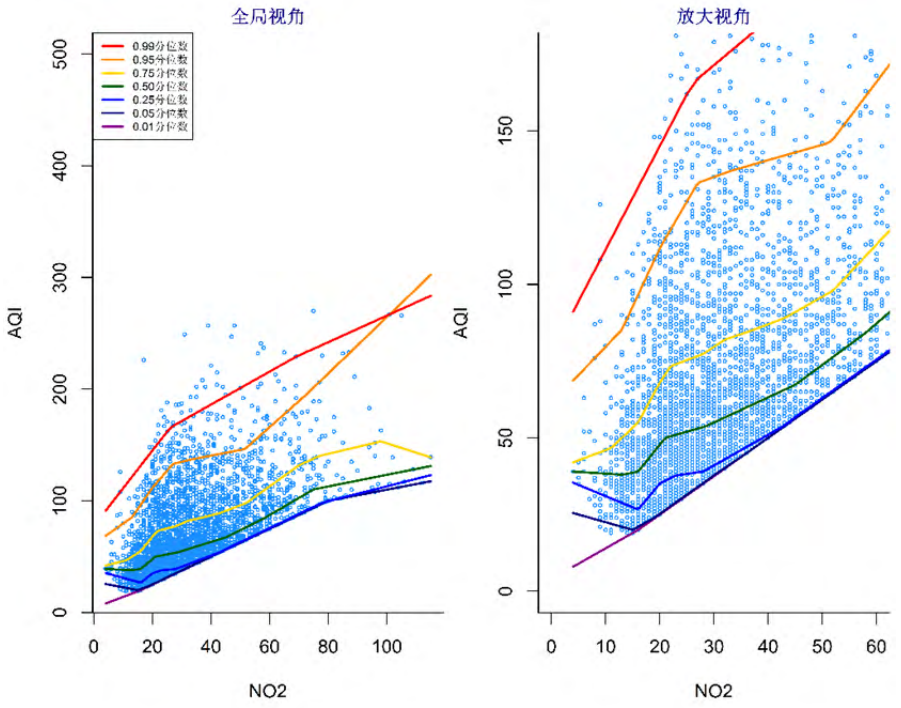

图 25 NO_2 对 AQI 数据的分位数回归结果

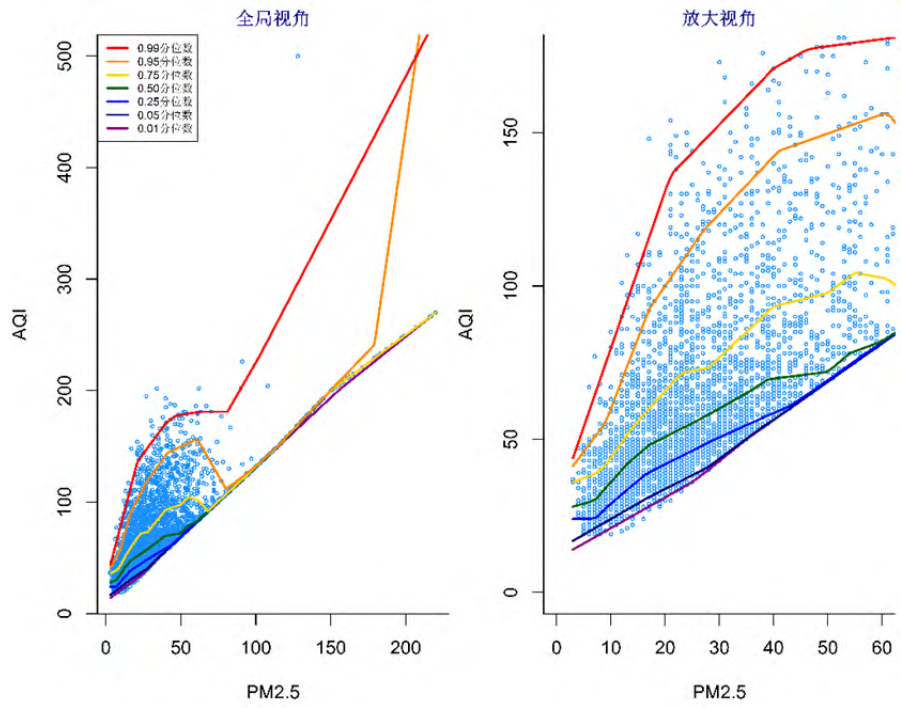

图 26 $PM_{2.5}$ 对 AQI 数据的分位数回归结果

(七) 考虑 AQI 季节效应的边际非参数回归和半参数回归建模

首先将数据按月份分组，按照季节的气象划分法：3～5 月为春季，6～8 月为夏季，9～11 月为秋季，12 月～次年 2 月为冬季划分季节。将每个月的 AQI 数据视为一组。则数据集为(x_{ij}, y_{ij})，其中$1 \le i \le m, 1 \le j \le n$。记$y_i$为第$i$组的响应变量向量。则边际非参数回归模型可以表示为

$$y_{ij} = f(x_{ij}) + \varepsilon_{ij}, Cov(\varepsilon_i) = \Sigma, 1 \le i \le m, 1 \le j \le n, \tag{11}$$

其中$\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in})^{\mathrm{T}}$为第$i$组的随机误差。假设均值函数$f$是光滑的，协方差矩阵$\Sigma$是一个非结构化的$n \times n$的协方差矩阵。模型(R 文件 **nonpar**，或程序文档 **Code 9**)的拟合结果如**图 27** 和**图 28**。由图可知，北京和成都空气质量最高的时间是秋季，而深圳和上海则是夏季的空气质量更优。

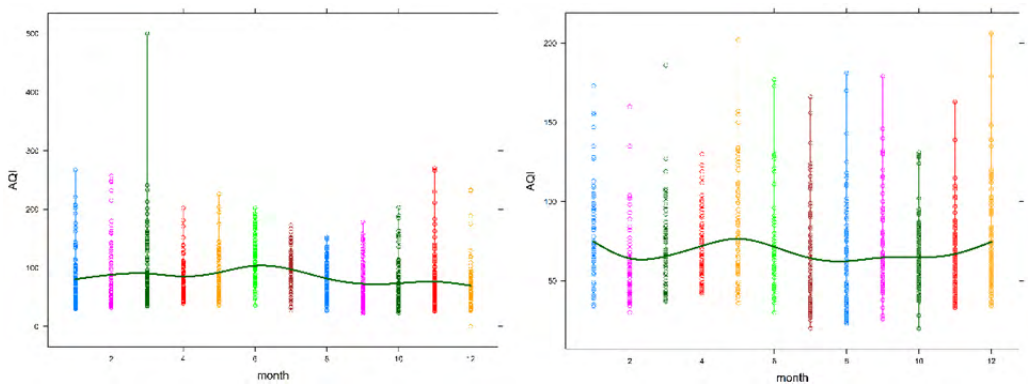

图 27 北京(左)和上海(右)按月分组数据的边际非参数拟合

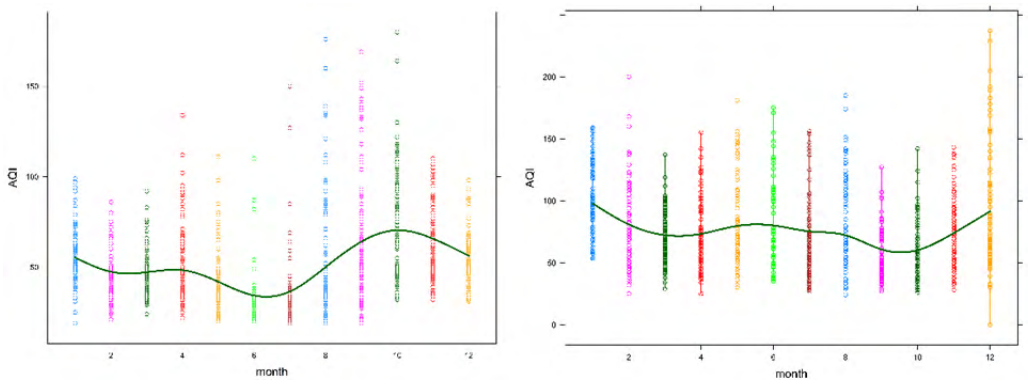

图 28 深圳(左)和成都(右)按月分组数据的边际非参数拟合

将数据分组后，依然可以从贝叶斯半参数回归的角度进行分析，考虑下列的贝叶斯半参数模型(R 文件 bayessemi，或程序文档 **Code 10**)：

$$\begin{aligned} AQI_{ij} \mid \beta, U_i, u_1, \ldots, u_K, \sigma_U, \sigma_u, \sigma_\varepsilon \overset{ind}{\sim} N(\beta_0 + U_i + \beta_{NO_2} NO_{2_{ij}} \\ + \sum_{k=1}^{K} u_K z_k(NO_{2_{ij}}) + \beta_1 shanghai_i + \beta_2 shenzhen_i + \beta_3 chengdu_i, \sigma_\varepsilon^2), \end{aligned} \tag{12}$$

其中 $U_i \mid \sigma_U \overset{ind}{\sim} N(0, \sigma_U^2)$，$u_k \mid \sigma_u \overset{ind}{\sim} N(0, \sigma_u^2)$，$\beta_0, \beta_{NO_2}, \beta_1, \beta_2, \beta_3 \overset{ind}{\sim} N(0, \sigma_\beta^2)$，$\sigma_U \sim Half-Cauchy(A_U)$，$\sigma_u \sim Half-Cauchy(A_u)$，$\sigma_\varepsilon \sim Half-Cauchy(A_\varepsilon)$。

半柯西分布已经在之前进行了介绍。这里，我们通过设置超参数 $\sigma_\beta = A_U = A_u = A_\varepsilon = 10^5$ 来加强无信息先验。拟合的结果为**图 29 至图 32**。**图 29** 和**图 30** 显示了 NO_2 和 $PM_{2.5}$ 含量对 AQI 的拟合曲线和估计值的 95%置信区间，其中同一颜色的散点连起来的线表示一个月，因此每一条线上有 28 至 31 个散点。**图 31** 和**图 32** 显示了贝叶斯估计结果的摘要。后验样本路径图和自相关图都显示了参数的收敛性，第 5 列显示了参数的后验分布的密度估计，最后一列则表示了参数的点估计及其 95%的置信区间估计。结果都显示了北京市 AQI 与其他 3 个城市之间的差异，其中北京市空气中的 NO_2 和 $PM_{2.5}$ 的平均含量高于其他 3 个城市，在其余 3 个城市中，深圳和上海表现较好，其中深圳市 AQI 表现了最小的波动率。

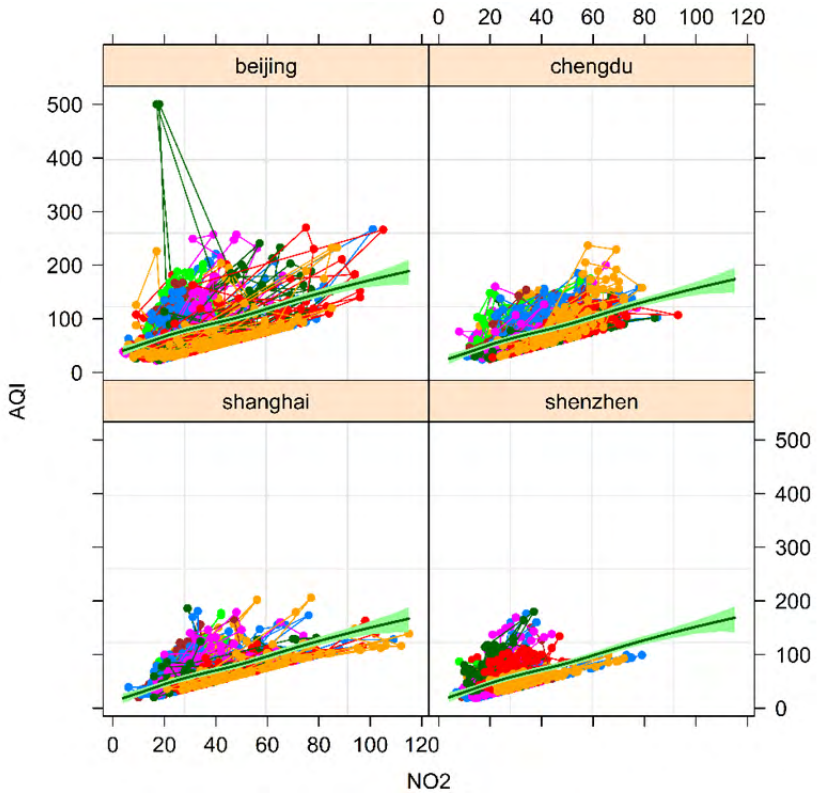

图 29 NO_2 含量对每个城市按月分组 AQI 的均值函数的贝叶斯估计

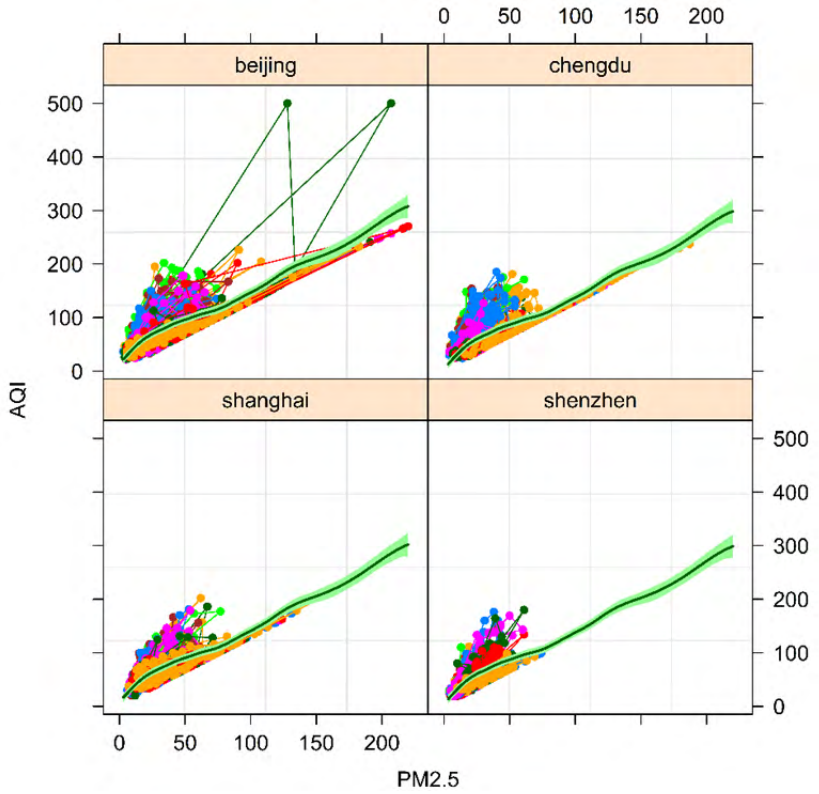

图 30 $PM_{2.5}$ 含量对每个城市按月分组 AQI 的均值函数的贝叶斯估计

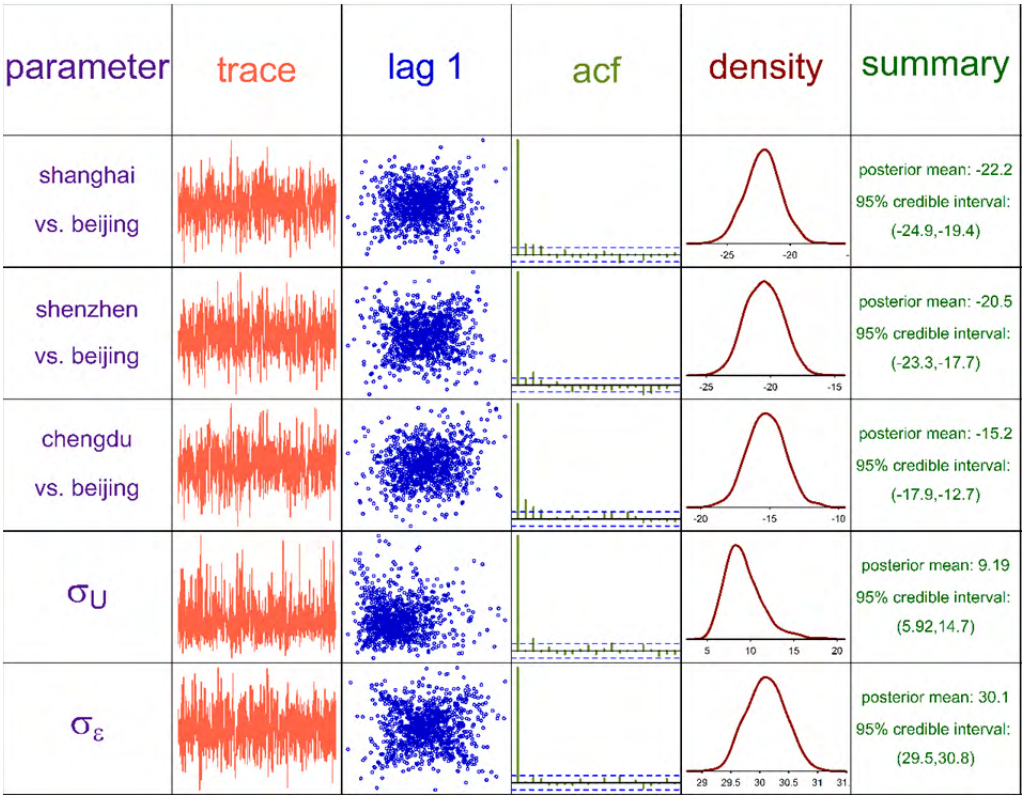

图 31 贝叶斯半参数回归模型下 NO_2 含量对 AQI 的拟合摘要

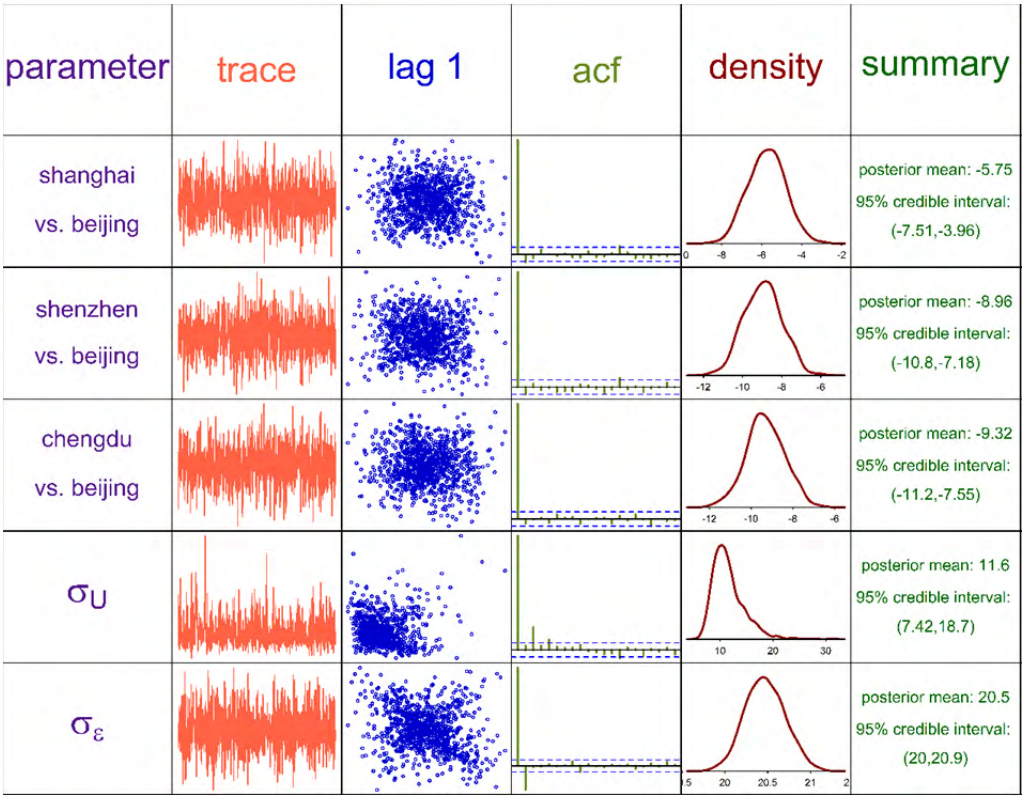

图 32 贝叶斯半参数回归模型下 $PM_{2.5}$ 含量对 AQI 的拟合摘要

五、结论与建议

(一) 结论

本文研究的特点在于使用半参数回归模型研究了我国的空气质量指数(AQI)数据, 克服了传统的线性模型和部分时间序列分析中无法充分提取非线性信息的缺点。同时，基于分位数的半参数回归对 AQI 离群值表现出稳健性。本文得出的主要结论如下：

1. 在半参数可加模型的 NO_2 含量对 AQI 的拟合中，上海、深圳和成都市都与北京市的 AQI 有显著的不同。可以合理地得出结论，上海市 AQI 平均值比北京市平均低 21.0440, 而深圳和成都市的 AQI 平均值分别比北京市平均低 22.4551 和 13.7166；

2. 考虑不同城市数据的交互作用时，在 NO_2 的含量方面，北京市在 AQI 的低值和高值处都具有很高的 NO_2 含量，而深圳市和上海市的 NO_2 含量在相同情况下都更低。需要注意的是，对于成都市，低值时的 AQI 对应了最低的 NO_2 含量，然而高值时的 AQI 却对应了最高的 NO_2 含量，说明影响成都市 AQI 的重要因素是空气中 NO_2 的含量高低。在 $PM_{2.5}$ 的含量方面，北京市在 AQI 的低值处拥有最大的 $PM_{2.5}$ 含量，高值 AQI 时 $PM_{2.5}$ 含量排名第 2，表明 $PM_{2.5}$ 是影响北京市空气质量的主要因素；

3. 考虑深圳市与全国水平的比较时，深圳市 AQI 空气中 NO_2 含量低于全国其他城市，表明深圳市空气质量高于其他城市；

4. 在半参数模型的稳健估计中, 当 AQI 值一定时, 非稳健估计都低估了 NO_2 和 $PM_{2.5}$ 的含量，尽管在某些值处这种低估非常小，但仍然不可忽视；在半参数分位数回归中，结果表明，0.01,、0.05、0.25、0.5 和 0.75 分位数间隔很小，表明每年的大多数时候我国的空气质量较好，同时得出结论：NO_2 和 $PM_{2.5}$ 含量对 AQI 的不同分位数的影响是不同的，即使 AQI 集中在值较小的区域内；

5. 考虑 AQI 变化趋势随季节而不同。在边际非参数回归中，得出结论：北

京和成都空气质量最高的时间是秋季，而深圳和上海则是夏季的空气质量更优；在贝叶斯半参数回归中，结果显示了北京市 AQI 与其他 3 个城市之间的差异，其中北京市空气中的 NO_2 和 $PM_{2.5}$ 的平均含量高于其他 3 个城市，在其余 3 个城市中，深圳和上海表现较好，其中深圳市 AQI 表现了最小的波动率。

(二) 建议

在本文的分析中，各项指标和结果都表明了北京市空气质量较上海、深圳和成都差。基于本文得出的结果，提出以下建议：

1. 对于北方城市，空气污染的主要因素是 $PM_{2.5}$。因此，一方面，加强北方风沙治理，植树造林，防范沙尘天气；另一方面，$PM_{2.5}$ 来自于工业大机器的排放的烟尘，如发电、冶金和石油化工等领域，因此工业技术的升级和工艺流程的改进也是治理空气质量的不可忽视的方面。同时，冬季燃煤使用时产生的 NO_2 也会影响空气质量，因此应尽快推广清洁能源；

2. 对于南方城市，空气污染的主要因素已由十年前的 SO_2 变为了 NO_2。汽车尾气中的 NO 排放到空气中极易被氧化为 NO_2，进而污染空气。因此大城市中应继续或加大力度推行燃油机动车限号行驶，同时加大新能源汽车的研发和推广，进而减少 NO_2 排放。

参考文献

[1] United States Environmental Protection Agency. 2009 final report: integrated science assessment for particulate matter. 2009.

[2] WHO. Air quality guidelines—global update 2005. http://www.who.int/phe/health_topics/outdoorair/outdoorair_aqg/en/.

[3] GBD 2015 Risk Factors Collaborators. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015[J]. Lancet, 2016, 388: 1659-724.

[4] Holland W W, Bennett A E, Cameron I R, et al. Health effects of particulate pollution: reappraising the evidence[J]. Am J Epidemiol, 1979, 110: 527-659.

[5] Dockery D W, Pope C A, Xu X, et al. An association between air pollution and mortality in six US cities[J]. N Engl J Med, 1993, 329:1753-59.

[6] Pope C A, Thun M J, Namboodiri M M, et al. Particulate air pollution as a predictor of mortality in a prospective study of US adults[J]. Am J Respir Crit Care Med, 1995, 151: 669-74

[7] Abbey D E, Nishino N, McDonnell W F, et al. Long-term inhalable particles and other air pollutants related to mortality in nonsmokers[J]. Am J Respir Crit Care Med, 1999, 159: 373-82.

[8] Katsouyanni K, Zmirou D, Spix C, et al. Short-term effects of air pollution on health: a European approach using epidemiological time-series data. The APHEA project: background, objectives, design[J]. Eur Respir J, 1995, 8: 1030-38.

[9] Katsouyanni K, Touloumi G, Samoli E, et al. Confounding and effect modification in the short-term effects of ambient particles on total mortality: results from 29 European cities within the APHEA2 project[J]. Epidemiology,

2001, 12: 521-31.

[10] Atkinson R W, Anderson H R, Sunyer J, et al. Acute effects of particulate air pollution on respiratory admissions: results from APHEA 2 project. Air Pollution and Health: a European Approach[J]. Am J Respir Crit Care Med, 2001, 164: 1860–66.

[11] 高燕. 基于 AQI 的济南市空气质量的分析及推断[D].山东师范大学,2020.

[12] 焦东方,孙志华.空气质量指数回归分析[J].中国海洋大学学报(自然科学版),2018,48(S2):228-234.

[13] 卢雨婷.科学、技术创新对经济增长的非线性影响研究——基于半参数模型的实证分析[J].商业经济,2021(05):124-126.

[14] Harezlak J, Ruppert D, Wand M P. Semiparametric Regression with R[M]. 2018.

附录

(一) 数据包

数据见文件夹“ 数据包 ”，包含两个表格。

(二) 程序包

R 代码见文件夹“ 程序包 ”，其中包含 10 个 R 文件和程序文档(共 25 页),程序文档目录如下：

程序文档目录