武漢理工大學

数学建模实战培训论文

基于断点回归的城市空气质量数据真实性判别分析

第 16 组

姓名	方向
肖善	建模
易雨谦	建模
林荣武	编程

2016 年 9 月 2 日

基于断点回归的城市空气污染数据真实性判别分析

摘要

空气污染问题一直是中国发展面临的主要问题，如何判断并提高空气质量数据的真实性更是国家及政府关注的重点。

本文针对三个区域典型城市的空气监测及气候数据，通过断点回归和多元线性回归法进行数据分析及相关性分析对空气数据的真实性进行了检测。

针对问题一，要求根据资料确定出异常数据，本文首先用断点回归分析，将典型城市的空气质量指数 AQI 作为回归的驱动变量，假设在 AQI 为 100 时数据被修改的可能性较大，然后通过 Matlab 及 R 软件从图像上观察出断点出现位置，所得到的断层（断点左右两边）现象极为明显，因此认为 100 可以作为数据的异常点及不真实性的反映。

针对问题二，在问题一的基础上要求考虑多种因素，首先，查找数据并绘制典型城市的气候数据图从而说明三个区域中每个城市都具有一致的气候条件；接着，对三个区域的城市 AQI 值进行方差分析得出几个典型城市的 AQI 数据存在异常现象，同时在 R 中对典型城市的 AQI 和 5 个污染物浓度进行多元回归分析，找到影响典型城市 AQI 的污染物浓度，并在此基础上在此做断点回归分析，最终得出污染物浓度异常数据出现在各个污染物浓度指标的临界值处

针对问题三，首先，通过问题一二的分析，结合我国政策与实际国情将空气质量数据的不真实情况进行了分类；接着，分别找到了不真实数据出现的原因；最后，针对不同断点出现的原因，我们提出了相应的提高数据真实性的对政府的解决措施。

针对问题四，首先以上海为例，查找相应时间段内主要工业产品产量，并对多种产品产量与上海的 AQI 值进行相关分析，比较多种产品对于 AQI 值的影响，最终得出显著影响上海 AQI 值的工业产品为原油、柴油和发电量，即通过空气质量数据的变化来大概地展示这些工业生产数据的实际情况。

最后，对模型进行灵敏度分析检验，通过改变组距，在 AQI=100 处的临界值落差依然存在且基本一致，与组数设定为 25 进行比较，变化不大，由此认为本文将 AQI=100 作为断点进行回归是合理的。

关键词： 断点回归 驱动变量 多元线性回归 相关性

一、问题重述

由于主客观原因，采集到的空气质量数据序列可能出现异常现象。题目在要求在大量数据资料的基础上解决以下问题：

1、搜集相关空气质量和气候数据，来查找并分析各个区域城市中的搜集到的异常数据。

2、在 1 的过程中，利用污染物之间的相关性、变化的连续性及指标在时间、空间等各层次上进行对比，以此来确定异常数据并讨论其严重性。

3、通过模型分析数据不真实的类型、原因，最终为环境保护和政策制定提供支撑。

4、进一步的讨论可以加入社会因素，例如分析空气质量与工业生产（例如钢产量）等数据之间的相关性，分析是否可以通过空气质量数据的变化来展示工业生产（例如钢产量）等数据的实际情况。

二、问题分析

2.1 问题一的分析

问题一要求根据附件中所给的三个区域的各项空气污染数据，建立一个指标来判断是否存在数据异常的现象。首先，在三个区域中选取典型城市作为后续的具体问题分析的城市，虽然数据存在缺失现象，但考虑到本题是在现有数据基础上进行真实性分析，若采取弥补数据方式可能导致最后的数据真实性误差更大，接着运用统计学中的断点回归分析，将典型城市的 AQI 指标作为回归的驱动变量，确定分组个数，依据空气质量指标在时间上的连续性假设，先绘制出典型城市 AQI 的频率分布直方图及概率密度曲线，由于驱动变量是连续的，若从图像上观察到出现断点值，则极可能是断点处的数据存在异常情况，将断点值在 Matlab 中进一步做断点回归分析，同时在 R 中进行进一步的验证，最终绘制出断点回归图像并且说明了所选取的典型城市的 AQI 指标的异常数据。

2.2 问题二的分析

问题二要求在问题一的求解过程中，考虑到除 AQI 指标外各个污染物浓度的相关性，各个污染物的连续性等问题，再次确定异常数据。此题从三个角度考虑，首先是空间即地理位置，由假设三个大区中的城市所处的总体环境及天气状况是一致的，绘制三个区域的天气状况变化并选取 AQI 指标，分别得到三个大区中的城市的 AQI 指标，在每个区域对其中的城市数据做方差分析，检验每个区域中的城市 AQI 指标是否认为一致，以达到判断数据是否异常的目的；第二种角度是从时间角度，类似于问题一的分析角度来进行的，类似于 AQI 指标，假设污染物浓度随着时间变化是连续的，首先将所选取的 4 个典型城市在 R 中作出 AQI 指标关于各个污染物浓度的多元回归方程，选取回归方程系数显著性

最大的对应的几个污染物再次做断点回归，得到污染物浓度的可能断点值并作为异常数据对待，在 Matlab 中进行断点回归的分析与验证。

2.3 问题三的分析

问题三要求将不真实数据做分类，根据问题一、二的异常数据结论，将异常数据做比较，同时根据异常数据发生的原因作为分类的标准，给出分类结果，并在最后为相应的环境保护政策制定提供相应的建议。

2.4 问题四的分析

对于问题四，本文通过查阅相关的资料，得到了上海在相应附件时间段内的工业产值的变化，并且运用相关分析，把工业生产值和 AQI 指标变化做比较，从相关系数大小最终确定可以通过空气质量数据的变化来展示工业生产的工业产品类型。

三、模型假设

1、假设正常情况下空气质量指数 AQI 及空气污染物浓度的分布在时间上是连续的。
2、假设监测站所检测到的空气污染物浓度在一天之内不发生变化。
3、假设正常情况下三个大区的空气污染环境状况近似一致。
4、假设污染物浓度的改变只与自身及天气状况有关。
5、不考虑极端环境对于空气污染状况的影响。

四、符号说明

符号	含义
f	驱动变量概率密度函数
C_p	污染物 P 的浓度
BP_{Hi}	与 C_p 相近的污染物浓度现值的高位置
BP_{Lo}	与 C_p 相近的污染物浓度现值的低位置
$IAQI_{Hi}$	与 BP_{Hi} 对应的空气质量指数
$IAQI_{Lo}$	与 BP_{Lo} 对应的空气质量指数
d	分组宽度

AQI	空气质量指数值

五、模型的建立和求解

5.1 建模准备

5.1.1 断点回归分析确定不真实数据

断点分析法[1]需要在潜在断点左右两侧分别进行局部回归；而局部回归多采用局部多项式回归，局部线性回归是局部多项式回归的特例，因为其算法较为简单且性能优越，本文采用局部线性回归。

断点分析法步骤[2]：

（1）在 100 左右两侧分别进行频率分直方图的建立；
（2）取直方图的中点作为自变量，对应的概率密度作为应变量；
（3）对左右两侧的数据分别进行局部线性回归。

多项式回归模型如下：

$$y_i = m(\mathrm{x}_i) + \varepsilon_i, i = 1, 2 \cdots n$$

其中 m(x) 表示回归函数，y_i 表示第个 i 采样点 x_i 处的采样值，ε_i 表示独立分布零均值噪声

在 m(x) 的形式未确定的情况下，假设是 N 阶局部平滑的，为了估计函数在给定数据下的任意点处的值，我们可以将函数在这一点局部展开。假设 x_i 是 X 附近的采样点，则有 N 阶泰勒展开式：

$$\begin{aligned} m(\mathrm{x}_i) &\approx \mathrm{m}(\mathrm{x}) + m'(\mathrm{x})(\mathrm{x}_i - \mathrm{x}) + \cdots + \frac{1}{N!} m^{(\mathrm{N})}(\mathrm{x})(\mathrm{x}_i - \mathrm{x})^N \\ &= \beta_0 + \beta_1(\mathrm{x}_i - \mathrm{x}) + \cdots + \beta_N(\mathrm{x}_i - \mathrm{x})^N \end{aligned}$$

首先定义权函数：

$$\omega_i(\mathrm{x}) = \frac{K_h[(\mathrm{x} - \mathrm{x}_i)]}{\sum_{j-1}^{n} K_h(\mathrm{x} - \mathrm{x}_j)}$$

其中 $K_h(\cdot)=K(\cdot/h)/h,\mathrm{K}(\cdot)$ 为核函数，它以估计点为中心，用来控制各个采样点的权重：距离 X 越近的点，权重越大，h 为带宽（平滑参数），用于控制核的尺度。核函数 $K_h(\cdot)$ 形式不固定，需满足关于 y 轴对称并在零点处取最大值，在这里我们使用高斯核

$$\mathrm{K(x)}=\frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})$$

选择 m(x) 来使得下面的局部加权平方和 Q 最小

$$Q=\sum_{i-1}^{n}\omega_i(\mathrm{x})(\mathrm{y}_i-\mathrm{m(x)})^2$$

估计 $\hat{\beta}=(\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_N)^T$ 依赖于目标值 X ，最终有

$$\hat{m}(\mathrm{x})=\hat{\beta}_0$$

当 N=1，为局部线性回归，因为其算法较为简单且性能优越，所以我们采用局部线性回归分别对临界值左右两侧的数据进行拟合。

局部线性参数的求解：

分别求 Q 对 $\beta_0,\beta_1,\cdots\beta_N$ 的偏导数，并让它们等于零，这里以线性回归进行局部的拟合，所以只需求 Q 对 β_0,β_1 的偏导数。

引入矩阵

$$X=\begin{pmatrix}1 & x_1\\ 1 & x_2\\ \vdots & \vdots\\ 1 & x_i\end{pmatrix},y=\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_i\end{pmatrix},\beta=\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}$$

整理关于 β_0,β_1 的线性方程组，使用矩阵表示如下：

$$(\mathrm{X}^T\,\omega\,\mathrm{X}\,\beta)=X^T\omega y$$

若 $(\mathrm{X}^T\,\omega\,\mathrm{X})^{-1}$ 存在，则有：

$$\hat{\beta}=\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}=(\mathrm{X}^T\ \omega\mathrm{X})^{-1}X^T\omega y$$

所以有 y 的估计值：

$$\hat{y}=\beta_0+\beta_1 x$$

直方图组距与核函数带宽的选取：

进行局部线性回归进行估计分两步：第一步需确保箱体大到包含足够多的样本使其样本点在临界值两边都比较平滑，但又要小到一定程度使得样本点在临界值处的跳跃能都明显的显现出来，这就需要选择合适的 b；第二步以直方图箱体的中点作为观测变量，以对应的概率密度作为结果变量，采用局部线性估计 y。但是潜在的不连续点不应包括在箱体中，分别在潜在不连续点左右两侧进行直方图的绘制，获得对应的样本点。

因所获选择不易过大，但是不宜过小。我们使用 $\hat{b}=\frac{2\hat{\sigma}}{\sqrt{n}}$，$\sigma$ 是 R_i 的标准差。带 h 宽控制着核函得数据较少，b 数权重变化的速率，即用来控制核的尺度，我们使用 McCrary[3]建议的带宽 $h=a\times b(a\in\{10,15,20\})$。

5.1.2 多元回归分析

回归分析是最常用的数据分析方法之一。它是根据已得的试验结果及往的经验来建立统计模型，并研究变量间的相关关系，建立起变量之间关系的近似表达式（即经验公式），并由此对相应的变量进行预测和控制。如果根据经验和知识判断与因变量有关联的自变量不止一个，就应该考虑用最小二乘准则建立多元线性回归模型设影响因变晕 y 的主要因素（自变量）有 m 个，记 $\mathrm{x}=(\mathrm{x}_1,\cdots\mathrm{x}_m)$，建立多元线性回归方程：

$$y=\beta_0+\beta_1\mathrm{x}_1+\cdots+\beta_m\mathrm{x}_m+\varepsilon,\varepsilon\sim N(0,\delta^2)$$

式中 β_0，β_1，$\ldots\beta_m$ 和 δ^2 都是与 $\mathrm{x}_1,\ldots\mathrm{x}_m$ 无关的未知参数，β_0，β_1，$\ldots\beta_m$ 为回归系数。

如果对变量 J 与自变量 $\mathrm{x}_1,\ldots\mathrm{x}_m$ 同时作 n 次观察（n＞m）得 n 组观察值，又采最小二乘估计求得回归方程

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m$$

5.1.3 变量相关性分析

相关性分析是考察两个变量之间线性关系的一种统计分析方式。更精确地说，当一个变量发生变化时，另一个变量如何变化。此时就需要通过计算相关系数来做深入的定量考察。要考察两个变量之间的线性关系，就要从两个重要的要来分析，一是相关的强度，二是相关的方向。皮尔逊积距相关系数，考察两个连续变量（或定距以上层次的变量）之间的相关关系

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

相关系数 r 的性质有：

（1）相关系数的取值范围为 $-1 \le r \le 1$;

（2）r 为正值时，两变量间为正相关；

（3）r 为负值时，两变量间为负相关；

（4）相关系数的绝对值 $|r|$ 愈大，两变量间相关程度愈密切。r=+1，为完全正相关；r=-1,完全负相关；r=0,两变量完全无关。

5.2 问题一模型的建立与求解

AQI 与 API 的关系

API 是空气污染指数（Air pollution Index，其分级计算评价的污染物为 SO2、NO2 和 PM10 3 项，每天发布一次。

AQI 是空气质量指数（Air Pollution Index），其分级计算评价的污染物在 API 的基础上又增加了细颗粒物(PM2.5)、臭氧(O3)、一氧化碳(CO)这三项，将雾霾的主因——PM2.5 并未纳入其中。

因此，AQI 采用的标准更严、污染物指标更多、发布频次更高，其评价结果也更加接近公众的真实感受。

API 和 AQI 的计算方法大致相同，我们以 AQI 的计算方法[1]为例：

$$IAQIp = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{Lo}} (C_p - BP_{Lo}) + IAQI_{Lo}$$

问题一给出了京津翼、长三角、珠三角三大区域的各个城市的空气环境状况，要分析每个城市的空气质量及污染物浓度数据是否存在异常，考虑到三个大区的空气环境状况基本一致，所以只选取三大区域中的两个典型城市进行数据异常的分析，其余城市的数据真实性分析所用到的方法与典型城市的方法一致。京津翼地区选取北京、天津，长三角地区选取上海、南京，珠三角地区选取广州和深圳。可以发现，所选取的城市中的数据存在以时间为单位的缺失现象，京津冀地区的北京、天津都缺少 2014 年 1 月 23 日、3 月 24 日、8 月 8 日、8 月 22 和

2015 年 1 月 1 日这五天的空气质量指数及污染物浓度数据；长三角地区的南京也是相同的 5 天的数据缺失；在北京等城市的基础上，上海还缺少 2014 年 12 月 6 日和 12 月 7 日的 数据，珠三角地区的广州与深圳这两个城市都缺少了与北京相同的五天的数据。

对于缺失的数据，虽然数据存在缺失现象，但考虑到本题是在现有数据基础上分析真实性，若采取补齐数据的方式可能导致最后的数据真实性误差更大，因此本题目的求解是建立在现有的数据基础上。

为找到异常数据，下面对这四个城市的 AQI 断点进行回归分析，城市的 AQI 值处于 100 的临界状态是作为此城市空气质量是否优良的临界标准，因此，认为 100 作为 AQI 数据的断点值是具有一定道理的，相关统计人士或领导或许为了提升城市的环境空气质量及相关利益，将 AQI 数据做相应的改动，AQI 指数在 100 作为评价临界值，受到改动的可能性较大，因此，在以后的 AQI 断点分析中都选取 AQI 数值在 100 的情况作为断点状况并且进行相关分析。在 Matlab 中绘制出了北京、天津、上海、广州四个城市的 AQI 频数分布直方图，并以 AQI 值 100 为中心，将直方图分为小于 100 及大于 100 的频数分布直方图显示如下。

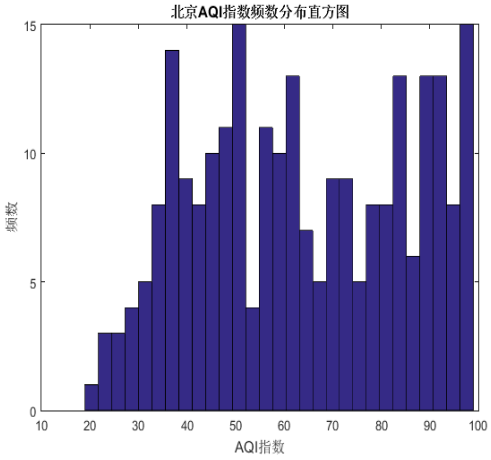


图 1 北京 AQI 频数(0-100)分布直方图

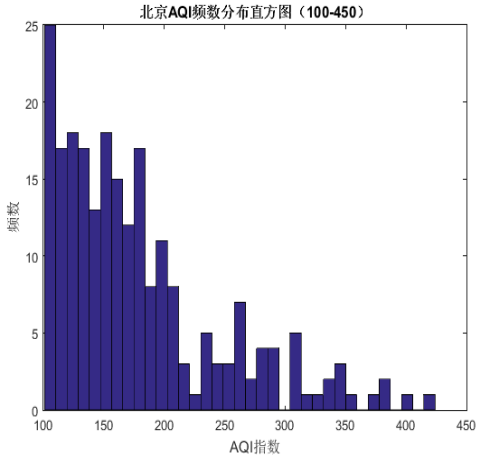
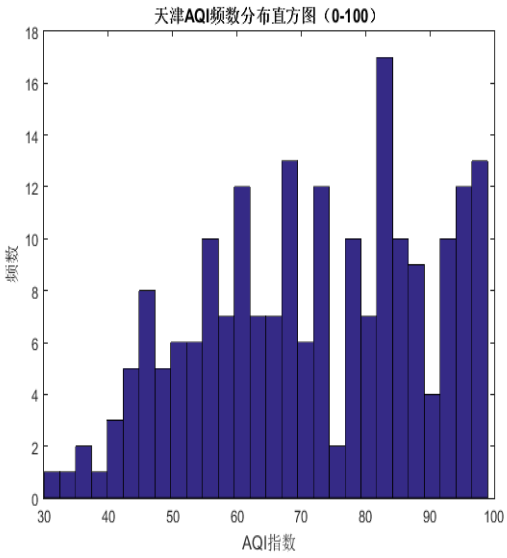


图 2 北京 AQI 频数(100-450)分布直方图



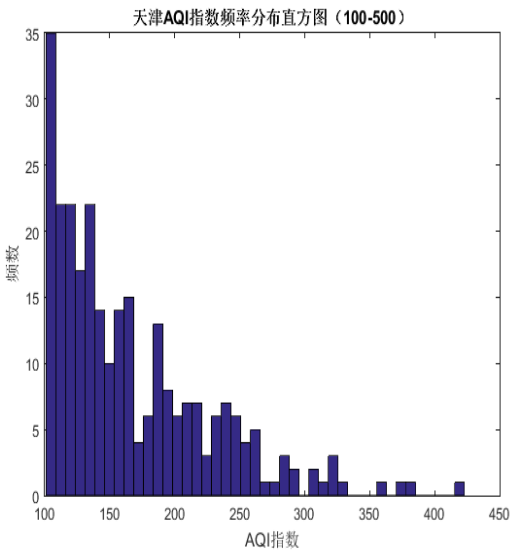


图 3 天津 AQI 频数(0-100)分布直方图

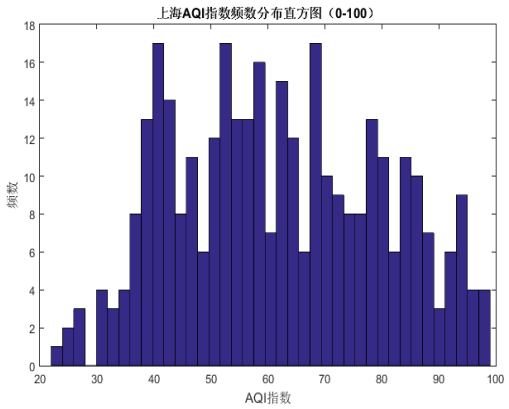


图 4 天津 AQI 频数(100-500)分布直方图

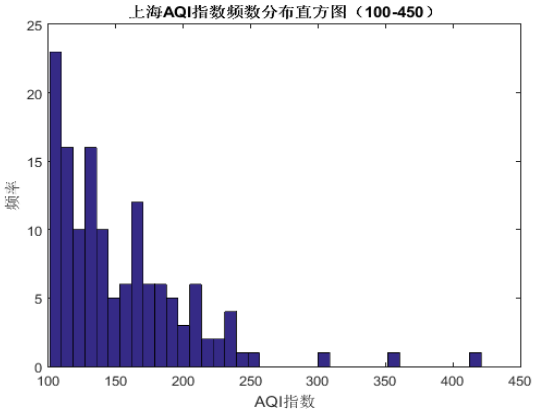
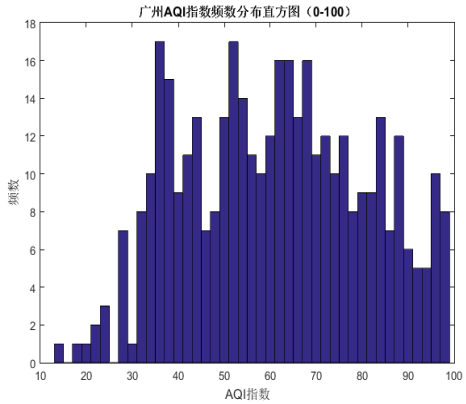


图 5 上海 AQI 频数(0-100)分布直方图

图 6 上海 AQI 频数(100-450)分布直方图



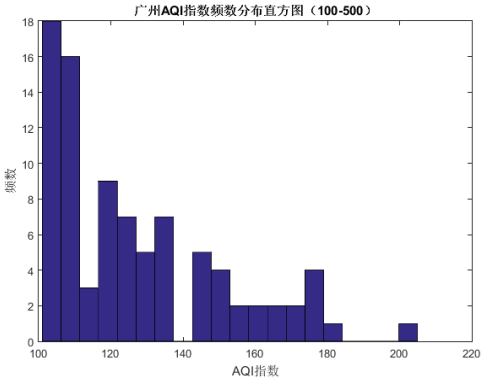
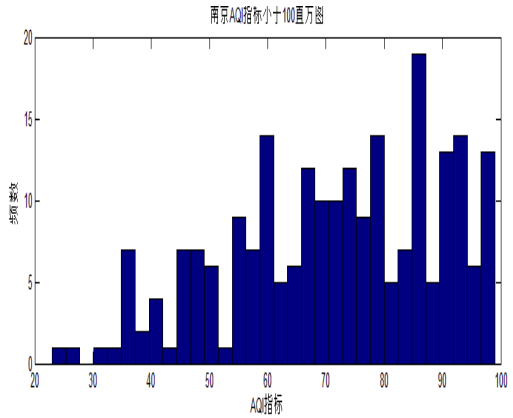


图 7 广州 AQI 频数(0-100)分布直方图

图 8 广州 AQI 频数(100-450)分布直方图



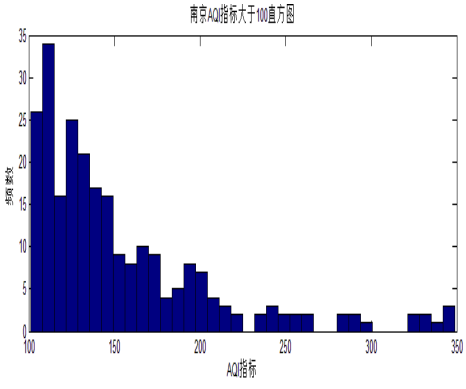


图 9 南京 AQI 频数(0-100)分布直方图

图 10 南京 AQI 频数(100-450)分布直方图

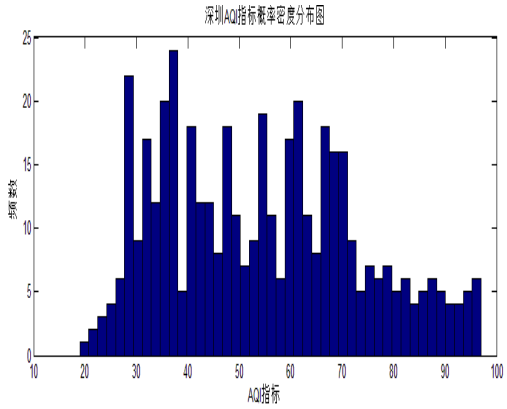


图 11 深圳 AQI 频数(0-100)分布直方图

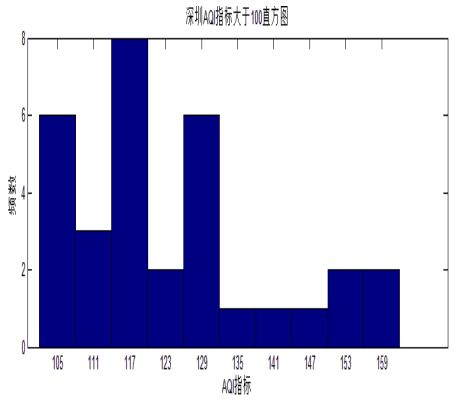


图 12 深圳 AQI 频数(100-450)分布直方图

通过观察个城市的以 AQI 指数为 100 为临界值的频数分布直方图可以知道，以 100 为中心的小于 100 的频数分布直方图与大于 100 的频数分布直方图的差距较为明显，断点处的左边与右边落差较大，说明以 AQI 值为 100 时候的临界值的时候左边与右边的极限概率函数极易出现不连续的情况。在 Matlab 中进而绘制出了以 AQI 指数 100 的断点回归图像，进行进一步的断点检查。

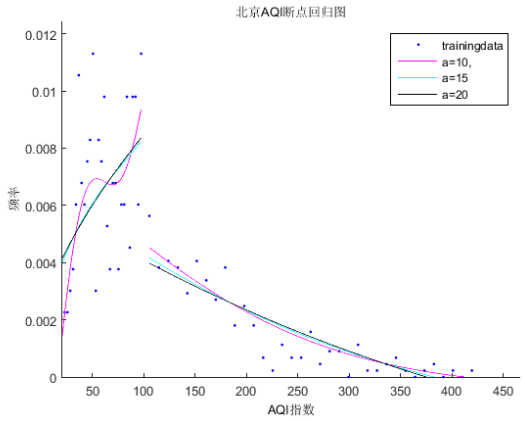


图 11 北京 AQI 断点回归图

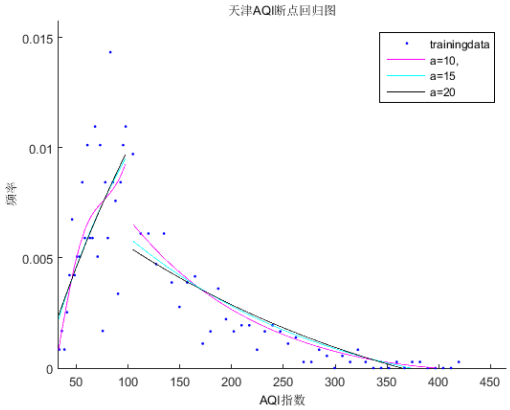


图 11 天津 AQI 断点回归图

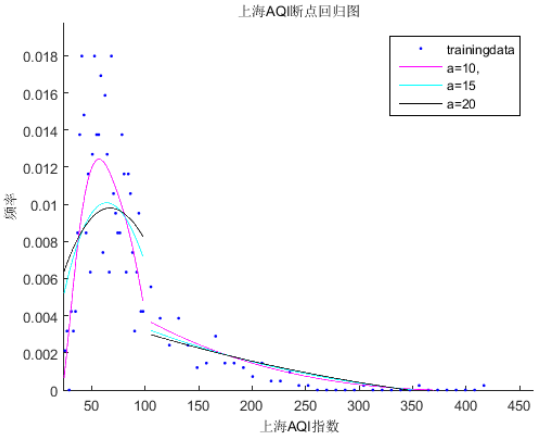


图 11 上海 AQI 断点回归图

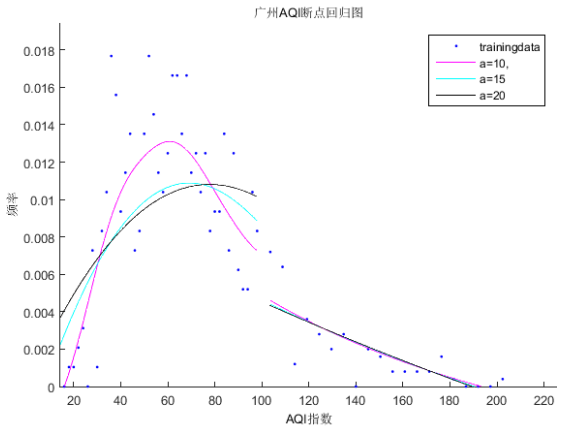


图 11 广州 AQI 断点回归图

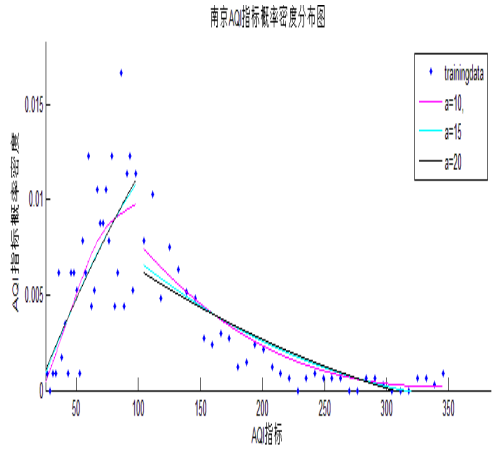


图 13 南京 AQI 断点回归图

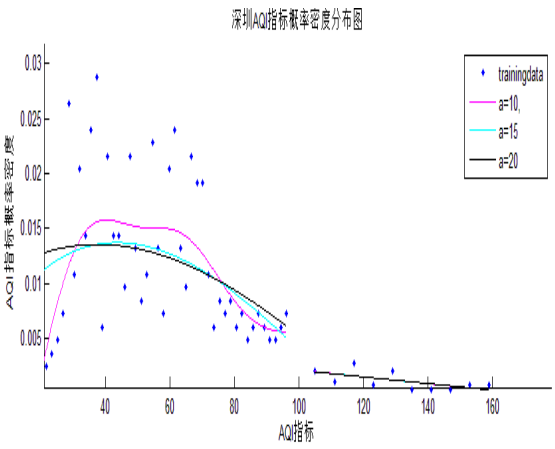


图 14 深圳 AQI 断点回归图

通过这 6 个城市的断点回归图，可以得到最终结论，以 AQI 指标 100 作为断点无论是从实际还是理论上都具有一定的依据以及意义，所以异常数据的发生点认为是在 AQI 数值为 100 的上下数值之间。同时，在 R 中对相同的城市进行断点回归分析，所得到的结果显示如下，从图像上也可以发现 AQI 值为 100 断点异常数据的合理性，并且也验证了 Matlab 结果的合理性。

下图显示了在 R 中的北京、天津、上海和广州的 AQI 时间变化散点图。

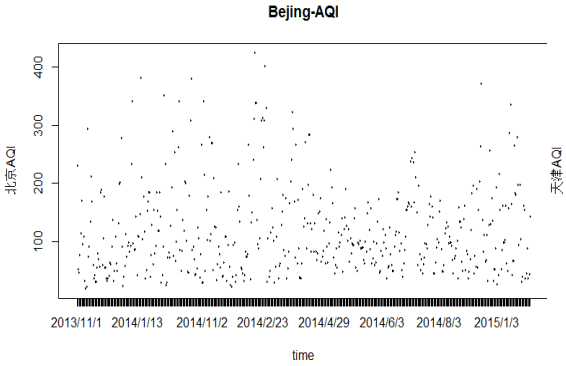


图 15 北京 AQI 指数变化

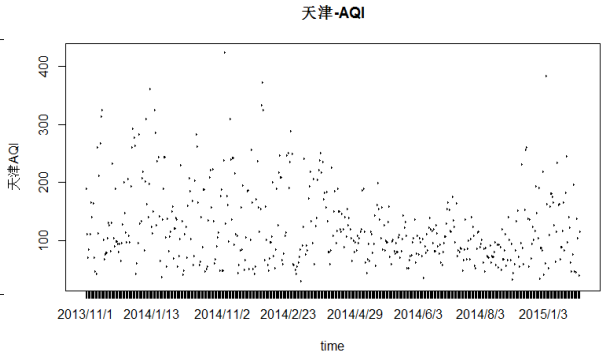


图 16 天津 AQI 指数变化

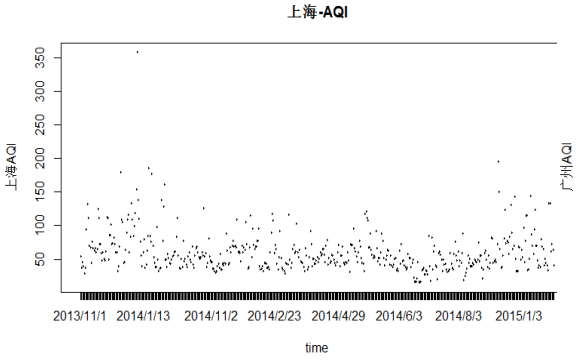


图 17 上海 AQI 指数变化

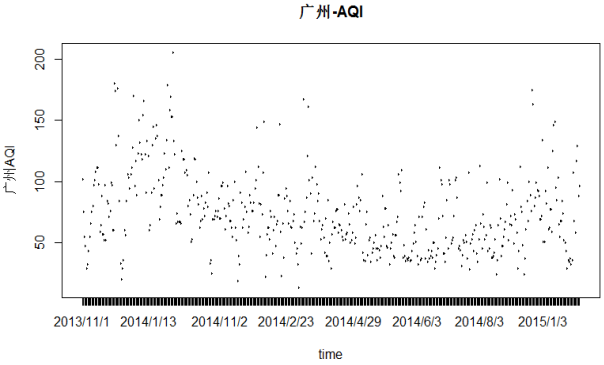


图 18 深圳 AQI 指数变化

下图显示了四个城市的频数分布直方图及概率密度曲线

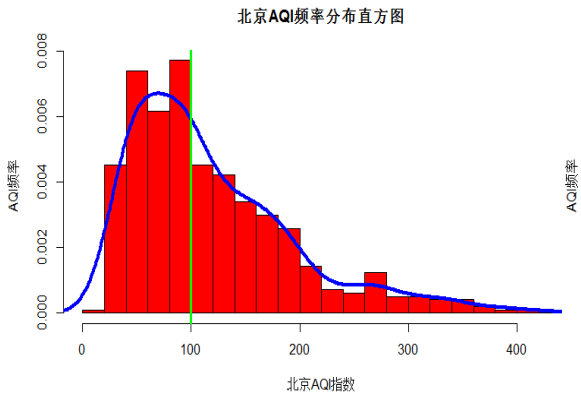


图 19 北京 AQI 频率分布及概率密度曲线

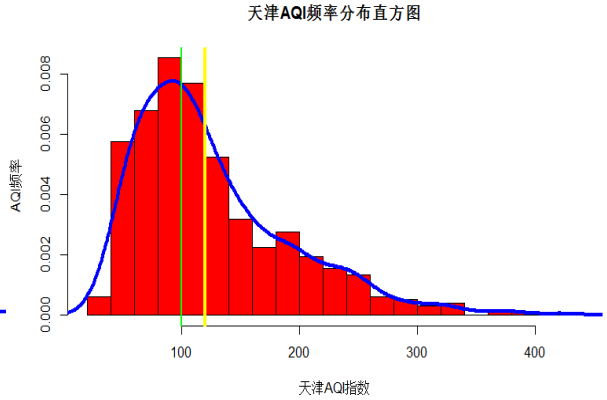


图 20 北京 AQI 频率分布及概率密度曲线

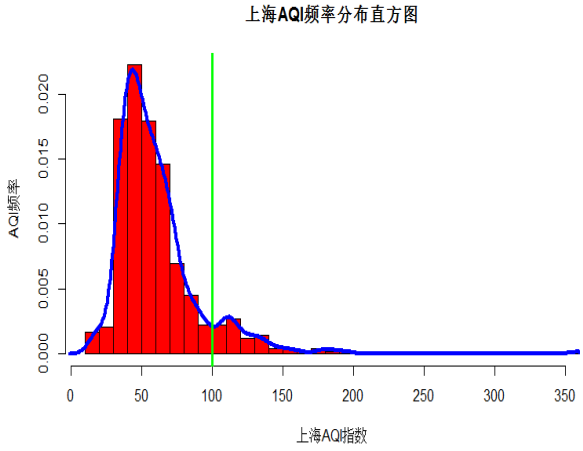


图 21 上海 AQI 频率分布及概率密度曲线

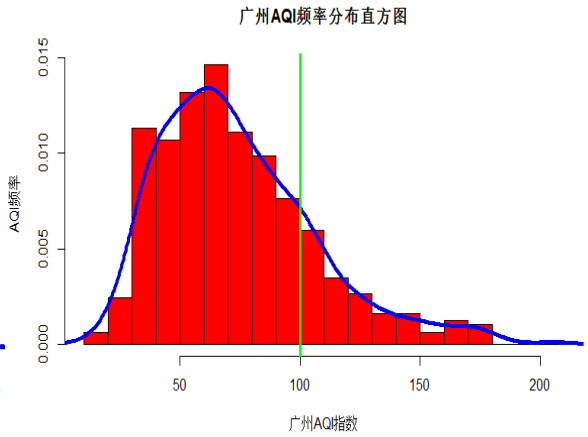


图 22 广州 AQI 频率分布及概率密度曲线

下图显示了四个城市的断点回归图像。

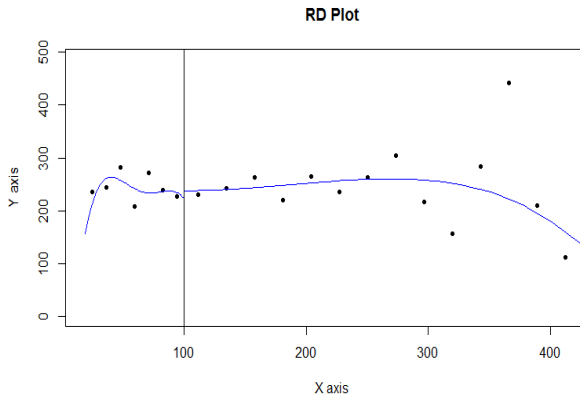


图 23 北京 AQI 断点回归图

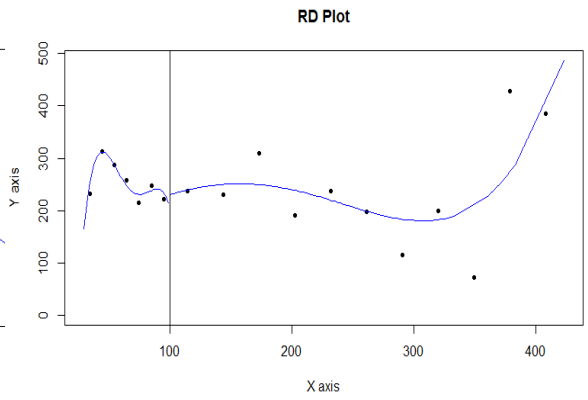


图 24 天津 AQI 断点回归图

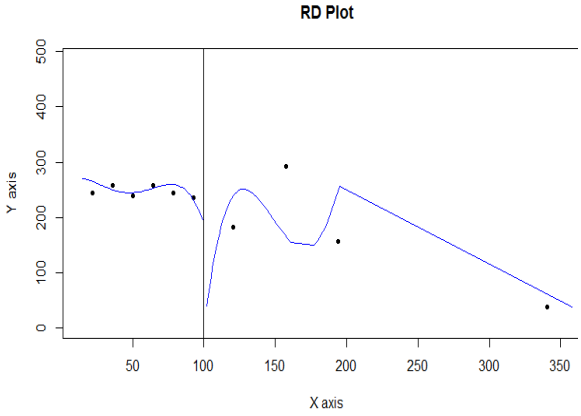


图 25 上海 AQI 断点回归图

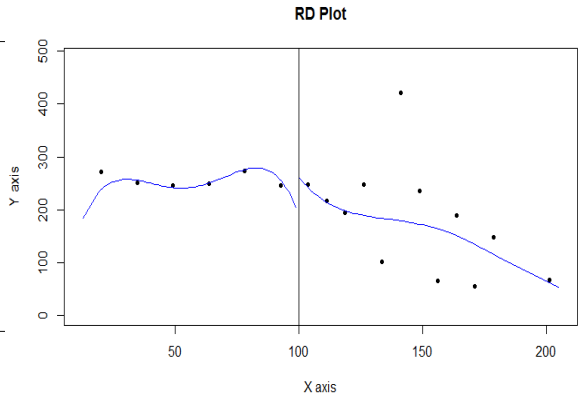


图 26 广州 AQI 断点回归图

通过 R 中的图像以及下表展示了 4 个城市的在 AQI 值为 100 处左右的频率，通过观察其落差可进一步说明异常数据发生在 AQI 值 100 左右的合理性。

表 1　4 个城市的 AQI 指数 100 左右频率落差显示

城市	AQI 指数 100 左频率	AQI 指数 100 右频率
北京	0.0077160494	0.0045267490
天津	0.0077160494	0.0052469136
上海	0.0146090535	0.0069958848
广州	0.0111111111	0.0098765432

通过观察及计算左右频率落差，可以发现在 AQI 值 100 附件落差都比较大，说明在 AQI 值为 100 时候，有很大的可能出现异常数据。

5.3 问题二求解

在问题一的基础上，首先查找相关的 4 个城市的天气状况，选取能够反映天气状况[4]的三项指标温度、相对湿度和风速，通过相关数据得到 4 个城市相同时间段下的温度湿度及风速，并在 R 中做出相应的图像，下图为北京和天津相应图示，其余城市的图像见附表。

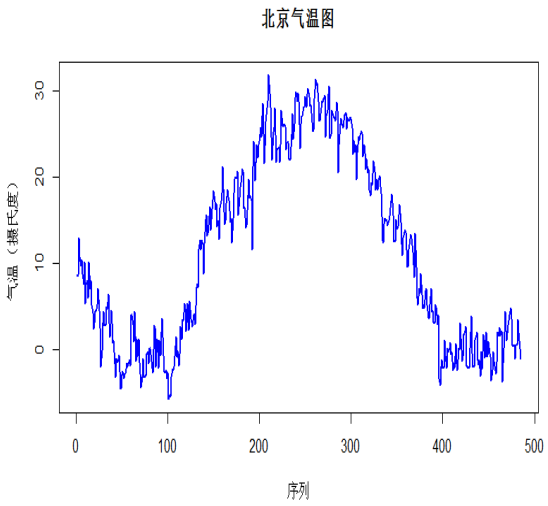


图 27 北京气温图

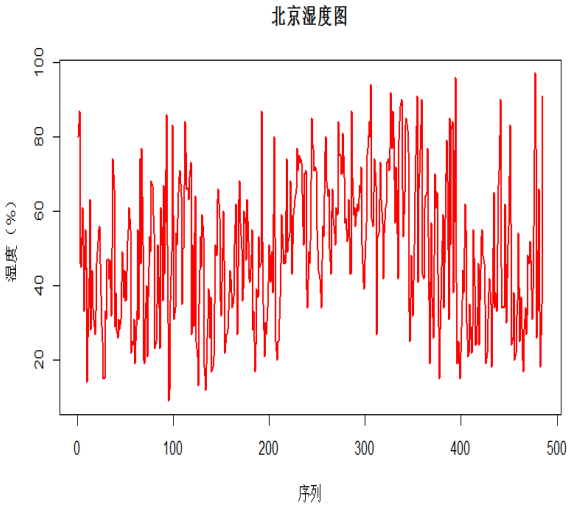


图 28 北京湿度图

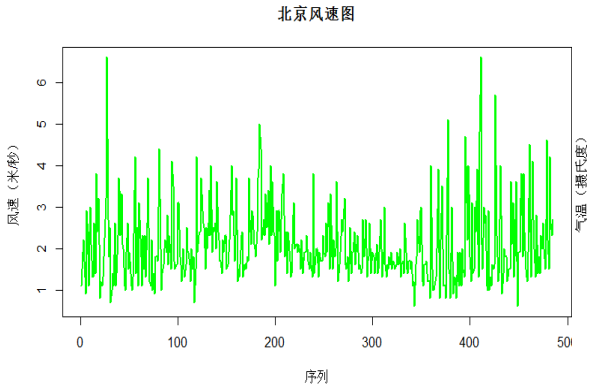


图 29 北京风速图

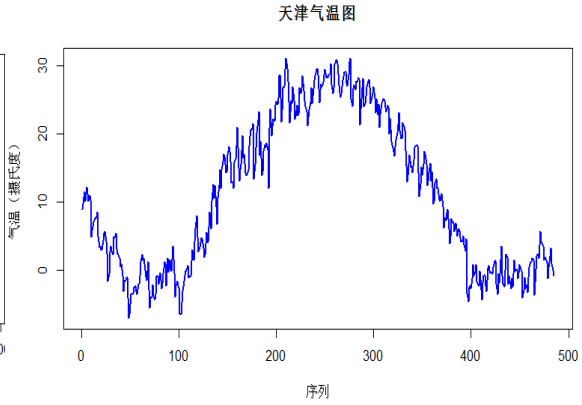


图 30 天津气温图

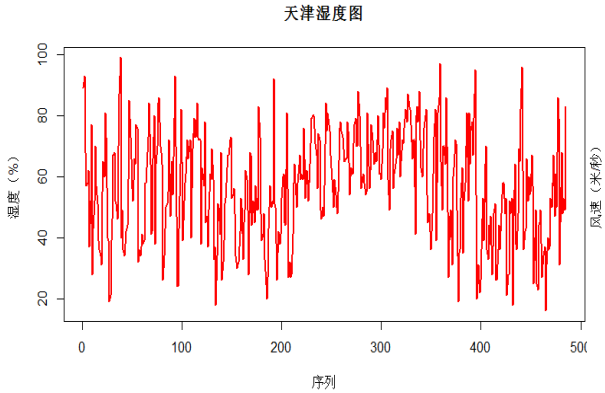


图 31 天津湿度图

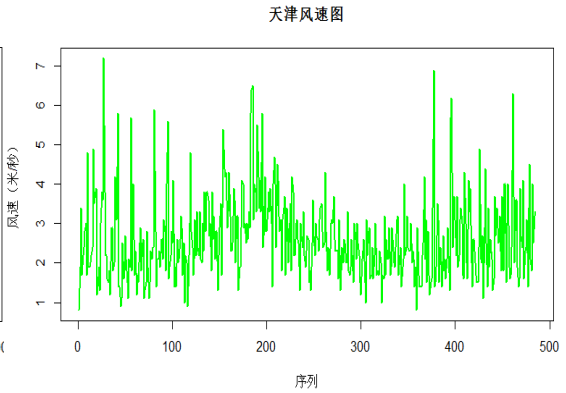


图 32 天津风速图

通过北京和天津的天气状况的三个指标的数据图像变化可以看出，天气的变化状况较为一致，在相同区域，天气状况几乎一致，所以认为在同一天气状况下，

相应的污染物浓度及 AQI 指标数值应该是趋于一致的，由于 AQI 数值是统一计算出来用于反映各个城市地区的空气质量的指标，所以可以假设其方差是齐性的，基于此，本文在 SPSS 中分别对这三个区域进行方差分析，并且将北京、天津、上海和广州作为参考城市。以北京天津为例，所得到的结果如下。

表 2 北京 AQI 与其他城市的方差分析显著性结果（显著性水平 0.05）

城市	显著性
天津	.174
石家庄	.000
唐山	.000
保定	.000
廊坊	.000
邢台	.000
张家口	.000
秦皇岛	.000
衡水	.000
邯郸	.000
承德	.000
沧州	.058

表 3 天津 AQI 与相同地区的方差分析显著性结果（显著性水平 0.05）

城市	显著性
北京	.174
石家庄	.000
唐山	.003
保定	.000

廊坊	.012
邢台	.000
张家口	.000
秦皇岛	.000
衡水	.000
邯郸	.000
承德	.000
沧州	.591

由表 2 及表 3 可以看出，北京及天津的 AQI 指标相对于其他城市差距很大，因此认为这两地的 AQI 数值出现了异常情况。上海与广州同理。

但从问题一的求解结果可以看出，城市的 AQI 指标存在异常及不真实的情况不只是 AQI 自身的问题，由于 AQI 指标会受到 PM2.5、PM10、CO、NO_2、SO_2 这些污染物浓度的影响，本文要进一步分析的是哪些具体的污染物浓度数值存在异常导致了 AQI 指数异常。首先，以 AQI 数值为因变量，5 个污染物浓度为自变量，对四个城市在 R 中进行多元线性回归分析，找出影响各个城市的 AQI 值最为显著的污染物。

4 个城市线性回归方程系数及其显著性如下：

表 4 城市回归系数及其显著性

回归方程	$Y(AQI)=\alpha X_1(PM_{2.5})+\lambda X_2(PM_{10})+\gamma X_3(CO)+\omega X_4(NO_2)+\eta X_5(SO_2)$									
回归系数	α	显著性	λ	显著性	γ	显著性	ω	显著性	η	显著性
北京	0.87551	<2e-16	0.33488	<2e-16	2.11828	0.370	-0.46760	2.6e-13	-0.08663	0.147
天津	0.74437	<2e-16	0.29732	<2e-16	4.95815	0.045285	-0.19498	0.01949	-0.14194	0.000183
上海	0.87167	<2e-16	0.17736	7.35e-07	5.31838	0.348	-0.03480	0.542	-0.04745	0.582
广州	1.12150	<2e-16	0.09387	0.019587	-2.18663	0.197748	-0.10931	0.000187	-0.02877	0.555776

选取显著性水平为 0.01，可以通过表 5 得到影响各个城市 AQI 指标的显著

污染物。下表显示所得到的 4 个城市的显著影响 AQI 的污染物。

表 5 影响城市 AQI 显著污染物

城市	显著污染物
北京	$PM_{2.5}$ PM_{10} NO_2
天津	$PM_{2.5}$ PM_{10} SO_2
上海	$PM_{2.5}$ PM_{10}
广州	$PM_{2.5}$ NO_2

根据表 4 并仿照问题一，分别对每个城市中的显著性污染物做断点回归分析，下表给出了各个污染物的临界指标，将临界指标作为断点（以北京为例）进一步在 Matlab 中做断点回归，绘制出断点回归图像。

表 6 各个显著污染物监测评价临界值

污染物	$PM_{2.5}$	PM_{10}	NO_2	SO_2
临界值	75	150	200	500

以北京天津为例，分别依据 $PM_{2.5}$ 和 PM_{10} 临界值仿照问题一绘制出断点回归的图像。

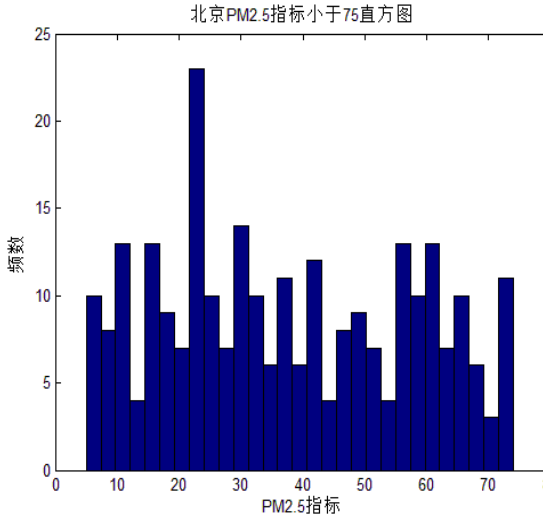


图 33 北京 PM2.5 频数(0-75)分布直方图

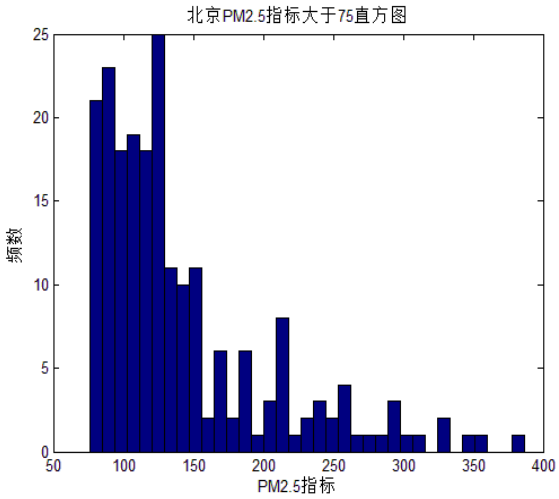


图 34 北京 PM2.5 频数(75-400)分布直方图

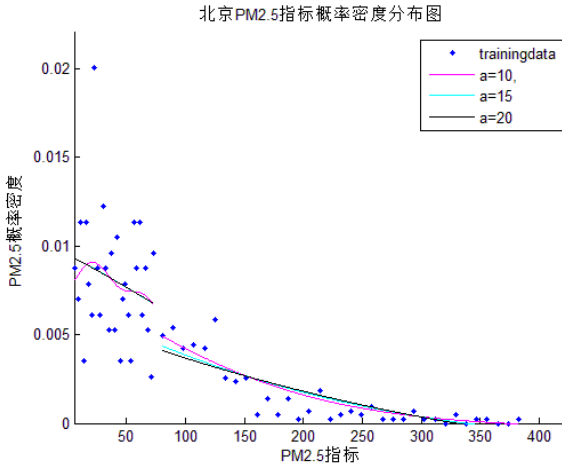


图 35 北京 PM2.5 断点回归图

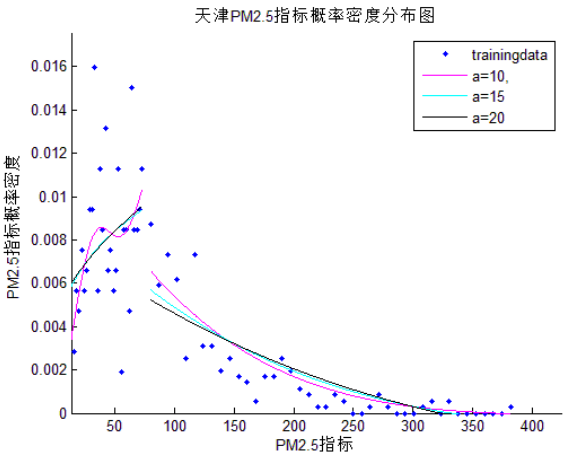


图 36 天津 PM2.5 断点回归图

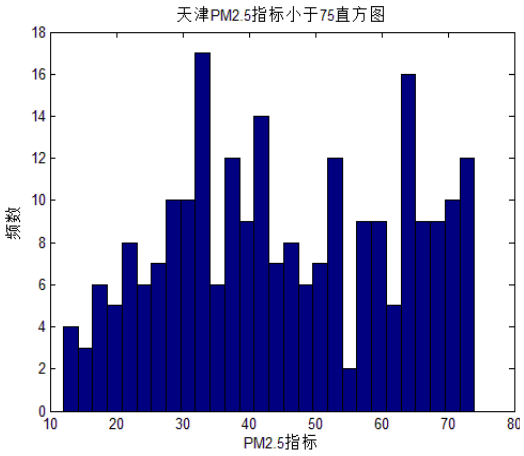


图 37 天津 PM2.5 频数(0-75)分布直方图

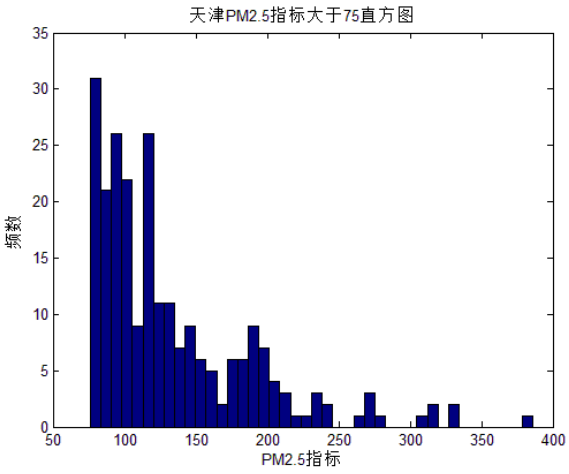


图 38 天津 PM2.5 频数(75-400)分布直方图

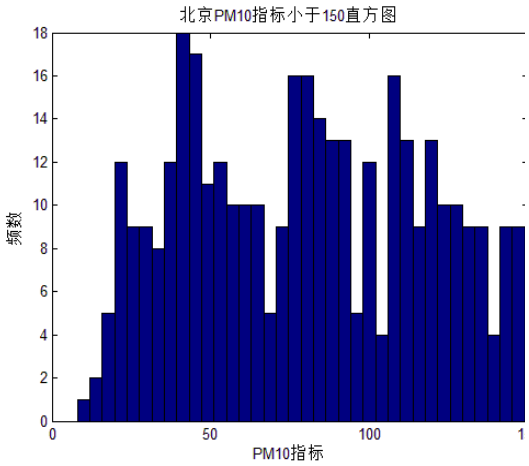


图 39 北京 PM10 频数(0-150)分布直方图

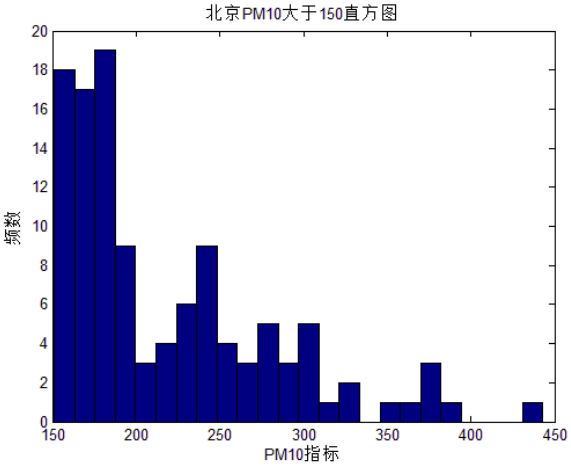


图 40 北京 PM10 频数(150-450)分布直方图

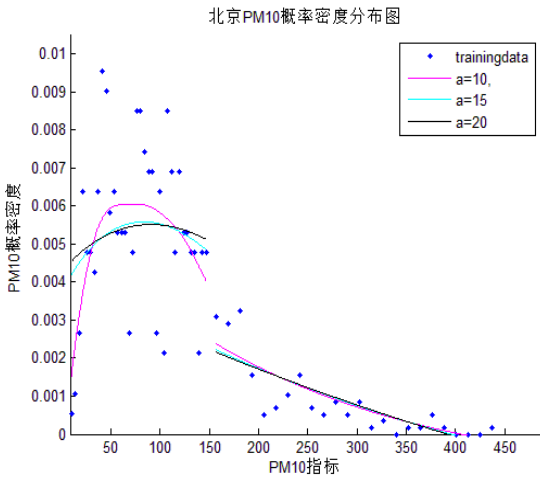


图 41 北京 PM10 断点回归图

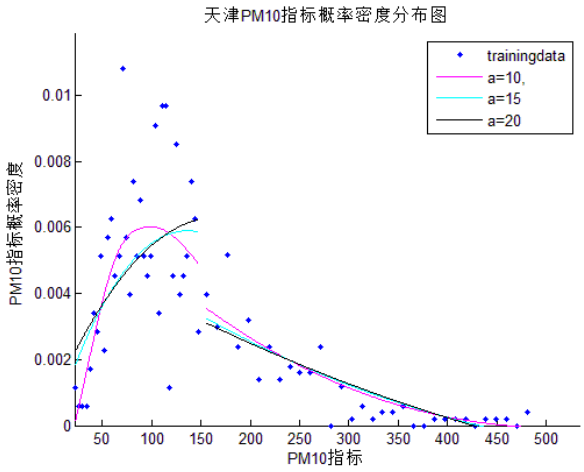


图 42 天津 PM10 断点回归图

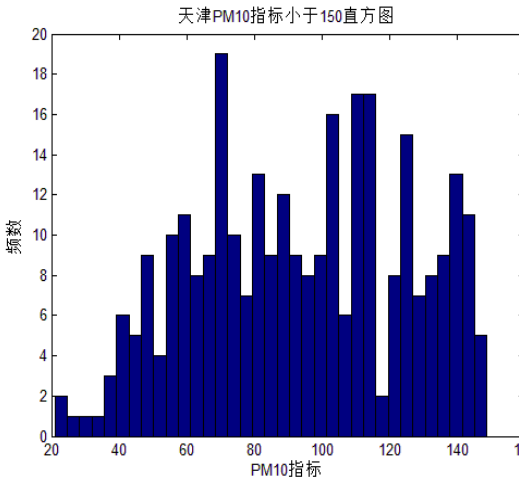


图 43 天津 PM10 频数(0-150)分布直方图

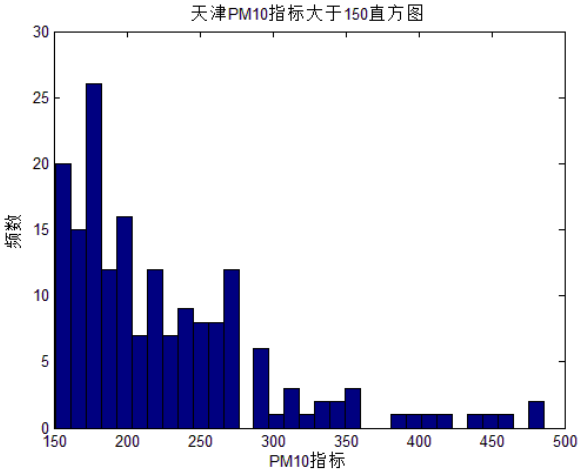


图 44 天津 PM10 频数(150-450)分布直方图

由图象可知污染物浓度异常数据很有可能出现在各个污染物浓度指标的临界值处。这也解释了 AQI 数值出现异常的原因。

5.4 问题三求解

根据问题一、二的模型分析，城市的 AQI 指标及污染物浓度数据异常及不真实的情况可以分为以下两种类型：

第一种是非主观性的数据统计错误，该类型的数据出错包括两种情况，一种是缺失型数据，这种类型的数据极为少见，但是由于仪器自身的检测问题或者是因为工作人员自身的疏忽大意造成的数据的丢失；还有一种情况是仪器自身的监测存在一定的误差，工作人员的统计时候也会或多或少受到其他因素的影响而导致的数据异常情况，这种错误在一定程度上属于系统误差，是较难避免的，这种数据的不真实性是不定向性的，数据可能会变大也可能会变小。对于缺失型的异常数据，只需要用多种统计方法进行拟合或回归等数学方法弥补，而对于第二种数据要看出现错误的数据是否是 AQI，还是具体某个污染物浓度的监测数据出现

了异常，这就可以仿照问题二中的多元回归分析的方法，通过确定显著污染物，再进一步确定污染物的临界值即异常数据。

第二种是主观性的数据统计错误，即有人为因素的干预使得数据朝着某一特定方向发生变化，例如政府为了使得整个城市的空气质量指标符合标准，强行修改部分污染物浓度的真实数据。对于这种异常数据，只需运用问题一中的断点回归分析，数据的异常在 AQI 的断点回归分析在图中会有非常直观的断层的体现，即在概率密度曲线上会在临界点出现较大的落差。如果要找到具体的污染物可以结合天气状况指标，污染物概率密度分布图中找到峰值异常的空气污染物数据。

统计数据质量的好坏，不仅影响以此为依据的政府决策的正确性与科学性，而且还直接影响着国家统计部门的形象和声誉。而目前出现的部分地方政府为了相关利益而对环境数据造假，不仅直接误导环境管理决策，而且严重损害环保部门和政府公信，数据质量是环境检测的保障，是保护环境的“红线”，对数据造假必须“零容忍”，对此必须：

1.加大执法力度，严厉打击环境监测数据造假行为，必须增加自上而下的数据抽检核对，并相应增强法律责任等层面的问责力度。环保部门不仅要关心企业的监测数据，更要确保数据的质量，对于监测数据的造假行为，必须及时发现并予以重罚。要让造假者因其造假行为付出的代价远远高于环保政绩可能带来的收益。

2.扩大环境监测的参与度，让新闻媒体和社会公众对环保监测进行全程监督，让环境监测数据造假行为面临更多障碍。不仅要使排污企业“不敢当，不敢为”还要推行“阳光排污口”行动，即将企业排污口置于公众监督下，人人都可随时监测检测，并与网上公布的数据相互印证，同时还要鼓励公众对企业数据举报、质疑、排查，形成强大的社会监督力量。

5.5 问题四求解

对于问题四，以上海为例，我们找到了上海在相应时间段内的月度工业生产总值，运用相关分析，找出 AQI 指标以及 5 个污染物对于工业生产总之的影响。首先绘制出空气质量与各个工业生产指标在相同对应时间下的月度数据如下表所示：

表 7 空气质量与部分用油工业生产指标数据

空气质量	植物油	服装	原油加工量	汽油	柴油
109.80	11.11	3680.58	217.85	43.53	72.29
164.48	11.26	4124.63	213.38	42.30	67.69
103.73	14.29	3817.82	203.75	40.47	68.68
73.71	4.97	2545.11	186.31	38.66	53.98
82.5	8.04	3278.82	206.49	41.81	61.14
76.93	6.62	3353.67	197.27	44.76	59.09
94.23	8.18	3538.04	215.35	49.84	64.63
66.37	10.43	4273.14	205.33	45.95	60.45
68.81	8,79	4910.28	151.71	35.98	46.08
60.07	9.43	4559.41	204.66	42.94	60.27
60.90	9.86	5005.46	140.31	27.31	44.0 7
68.52	7.63	4194.05	129.83	24.34	38.83

74.61	9.33	3996.00	186.63	35.27	57.94
100.24	12.04	4085.98	212.63	44.26	69.97
114.30	11.57	4029.78	208.43	44.30	65.07
89.86	6.42	2715.25	187.38	40.45	54.45

表 8 空气质量与部分工业生产指标数据

空气质量	乙烯	水泥	钢材	汽车	轿车
109.80	18.59	76.34	178.12	19.15	16.78
164.48	16.16	76.22	198.09	16.27	14.29
103.73	17.59	54.85	200.5	24.76	21.2
73.71	14.88	11.88	182.07	21.67	18.65
82.5	9.51	56.89	204.75	20.91	18.22
76.93	8.5	63.13	199.98	19.88	17.4
94.23	17.41	62.35	204.52	21.51	18.73
66.37	17.06	61.61	200.49	20.8	17.93
68.81	17.46	61.76	196.68	19.49	16.6
60.07	17.69	59.58	198.43	18.99	16.26
60.90	17.07	57.4	187.44	20.53	17.72
68.52	17.22	67.4	167.72	20.25	17.51
74.61	15.87	65.34	163.24	17.49	15.98
100.24	18.38	65.56	194.81	21.14	18.02
114.30	18.33	50.66	191.88	24.28	21.11
89.86	16.65	15.64	169.2	22.1	19.33

表 9 空气质量与部分家电工业生产指标数据

空气质量	家用电冰箱	房间空调	微机设备	集成电路	发电量
109.80	15.15	29.77	920.68	14.01	70.52
103.73	10.47	28.96	613	15.31	85.35
73.71	6.93	24.67	399.08	13.63	77.41
82.5	12.12	37.77	541.15	15.41	85.95
76.93	14.16	49.28	549.79	16.29	70.61
94.23	12.28	49.91	507.62	17.95	72.63
66.37	11.4	47.95	560.04	18.09	58.01
68.81	10.31	47.2	580.72	19.83	60.69
60.07	9.62	28.52	489.74	19.93	57.48
60.90	21.21	25.69	609.72	19.04	39.27
68.52	16.19	18.78	567.83	17.92	36.64
74.61	16.5	22.77	479.05	17.87	56.85
100.24	15.09	22.03	397.18	24.07	90.82
114.30	12.91	29.17	296.43	16.11	91.69
89.86	10.2	15.72	233.1	14.33	70.3

在 SPSS 中进行空气质量与 15 个工业生产指标做相关性分析，得到的相关系数如下所示：

表 10 生产指标与空气质量

指标	相关系数	指标	相关系数	指标	相关系数
植物油	0.46	柴油	0.64	轿车	-0.07
服装	-0.15	乙烯	0.15	家用电冰箱	0.06
原油	0.54	水泥	0.24	房间空气剂	-0.04
汽油	0.38	钢材	0.16	计算机设备	0.38
发电量	0.73	汽车	-0.09	集成电路	-0.36

从上表可以看出，影响上海城市空气质量的最为明显的工业生产指标依次为发电量、原油生产和柴油生产。由于原油及柴油的生产和加工都和一系列化学反应相关，而在工业中化学反应难免会生成污染物质及相关颗粒物，这对于当地空气质量的影响是巨大的。其次，电在生活中有着重要地位，所以发电是必不可少的，而现在电力来源基本上是火力发电和核能发电，在发电同时难免也会有污染物产生，所以发电量对空气质量的影响也很大。

六、灵敏度分析

对于四个典型城市，下面在设定不同组数的情况下，在 R 中绘制出相应的 AQI 数值频率分布直方图及概率密度曲线，通过观察可以发现，通过改变组距，在 AQI=100 处的临界值落差依然存在且趋于一致，与组数设定为 25 的情况下，变化不大，由此可以认为将 AQI=100 作为断点进行回归是合理的。

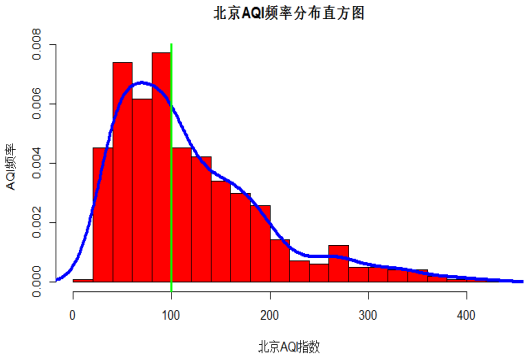


图 45 组数为 15 时的北京 AQI 分布图

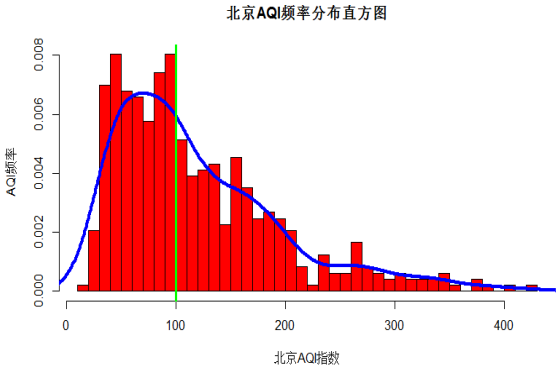


图 46 组数为 30 时的北京 AQI 分布图

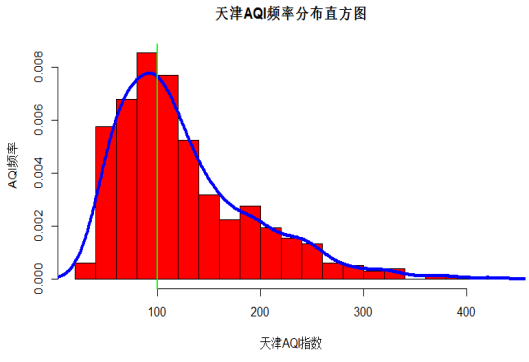


图 47 组数为 15 时的天津 AQI 分布图

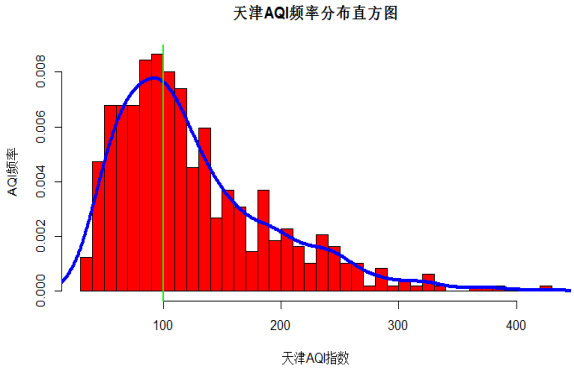


图 48 组数为 30 时的天津 AQI 分布图

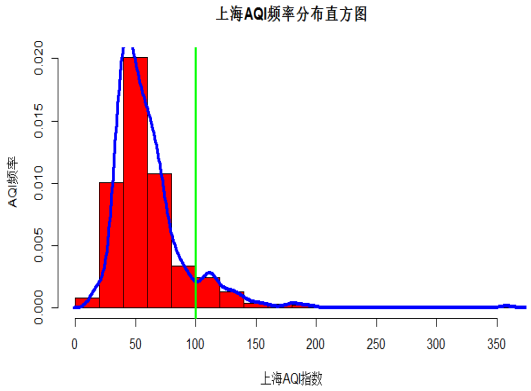


图 49 组数为 15 时的上海 AQI 分布图

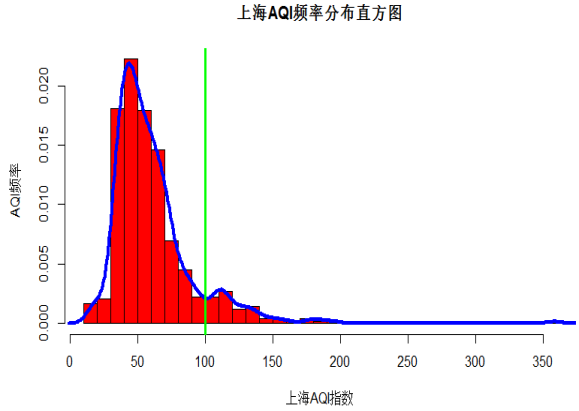


图 50 组数为 30 时的上海 AQI 分布图

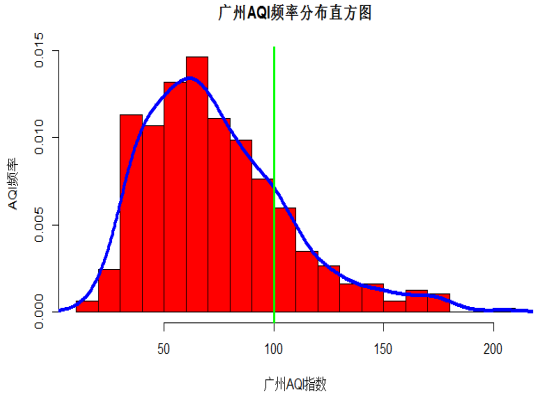


图 51 组数为 15 时的广州 AQI 分布图

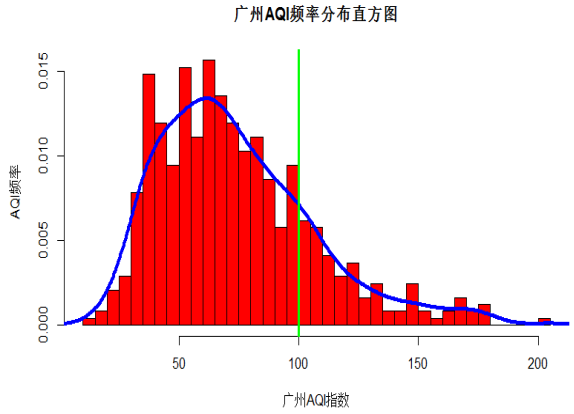


图 52 组数为 30 时的广州 AQI 分布图

七、 模型的评价和推广

7.1.模型优点：

（1）对于缺失值的处理方式考虑到了数据真实性的原则。

（2）采用断点回归分析，较为科学的解决了异常数据出现的状况，同时考虑了不同污染物浓度及天气状况的变化对于异常数据的影响，使得结果更加完善。

（3）通过多元线性回归中的系数显著性确定影响 AQI 的主要污染物，在具体的分析时达到断点回归的数据上的简化。

7.2.模型缺点：

（1）在做断点回归分析时大部分是从图像的观察上得出的断点的结论，缺少一定的定量分析。

（2）所选取的分析只是在四个城市上做的详细分析，在一定程度上不具有普遍意义，对于其他城市则需要依据相同方法再进行分析。

7.3.模型改进：

（1）三个区域之间也可以进行与气候条件相应的污染物浓度比较，运用相应时间序列方法作出一定预测，并查找相应预测时间内的真实值做对比，

从侧面反映原有数据的真实性。

（2）做断点回归分析时结合相应的理论，明确计算出左回归与右回归在临界值即断点值处的左右概率密度极限值，从数学角度分析断点的真实可靠性。

参考文献

[1] Ghanem, D., & Zhang, J. (2014). ‘Effortless perfection:’ Do Chinese cities manipulate air pollution data?. Journal of Environmental Economics and Management, 68(2), 203-225.

[2] Haddad M A. Increasing Environmental Performance in a Context of Low Governmental Enforcement: Evidence From China . Journal of Environment and Development, 2015, 24(1): 3-25.

[3]McCrary, Justin. 2008. “Manipulation of the runningvariable in the regression dis continuity design:A density test.” Journal of Econometrics 142 (2):698–714.

[4]周兆媛,张时煌,高庆先,李文杰,赵凌美,冯永恒,徐明洁,施蕾蕾. 京津冀地区气象要素对空气质量的影响及未来变化趋势分析[J]. 资源科学,2014,01:191-199.

[5]林艺滨. Excel 软件在计算空气质量指数的应用[J]. 科技资讯,2012,15:131+133.

[6]张锡颖,曲红伟. 城市空气污染数据的真实性判别及分析研究[J]. 科技经济导刊,2016,11:124+123.

[7]田成博文. 京津冀地区空气污染数据的真实性判别及分析[J]. 建材与装饰,2016,26:166-167.

[8]余静文,王春超. 新“拟随机实验”方法的兴起——断点回归及其在经济学中的应用[J]. 经济学动态,2011,02:125-131.

附录

问题一的 R 代码及结果显示

```
#北京
#北京
data1<-read.table("D:\\high\\R\\work\\建模实战\\1\\1_1.txt",header = TRUE)
#data1<-read.table("D:\\high\\R\\work\\建模实战\\1\\1_2.txt",header = TRUE)
data1
summary(data1)
plot(data1$time,data1$AQI1,xlab="time",ylab="北京AQI",main="Bejing-AQI",type="p")
h<-hist(data1$AQI1,breaks=25,freq=FALSE,col = "red",xlab = "北京 AQI 指数",ylab = "AQI 频率",main="北京 AQI 频率分布直方图")
lines(density(data1$AQI1),col="blue",lwd=4)
```

```
abline(v=100,lwd=3,col="green")
h$density
library(rdrobust)
a<-seq(from=1,to=486,by=1)
a
rdrobust(y=a,x=data1$AQI1,c=100)
rdplot(y=a,x=data1$AQI1,c=100)
结果显示：
> summary(data1)
                time                   AQI1
 2014/11/30:    2      Min.      : 19.0
 2013/11/1 :    1      1st Qu.: 62.0
 2013/11/10:    1      Median : 98.0
 2013/11/11:    1      Mean      :119.2
 2013/11/12:    1      3rd Qu.:158.0
 2013/11/13:    1      Max.      :424.0
 (Other)      :479
> h$density
 [1]  0.0001028807  0.0045267490  0.0074074074  0.0061728395  0.0077160494
0.0045267490
 [7]  0.0042181070  0.0033950617  0.0029835391  0.0025720165  0.0014403292
0.0007201646
[13]  0.0006172840  0.0012345679  0.0005144033  0.0005144033  0.0004115226
0.0004115226
[19] 0.0002057613 0.0001028807 0.0001028807 0.0001028807
> rdrobust(y=a,x=data1$AQI1,c=100)
Call:
rdrobust(y = a, x = data1$AQI1, c = 100)

Summary:

Number of Obs 486
NN Matches        3
BW Type           mserd
Kernel Type       Triangular

                            Left       Right
Number of Obs           249        237
Order Loc Poly (p) 1              1
Order Bias (q)          2              2
BW Loc Poly (h)       32.3641 32.3641
BW Bias (b)             50.4909 50.4909
rho (h/b)                0.6410    0.6410
bias                        2.2972    -8.9410
```

Estimates:

	Coef	Std. Err.	z	P>|z|	CI Lower	CI Upper
Conventional	27.6509	40.0135	0.6910	0.4895	-50.7741	106.0759
Robust				0.4048	-52.5986	130.3769

```
#天津
data2<-read.table("D:\\high\\R\\work\\建模实战\\2\\2_1.txt",header = TRUE)
#data2<-read.delim("clipboard",header = TRUE)
data2
plot(data2$time,data2$AQI2,xlab="time",ylab="天津AQI",main="天津-AQI",type="p")
h<-hist(data2$AQI2,breaks=20,freq=FALSE,col = "red",xlab = "天津 AQI 指数",ylab = "AQI 频率",main="天津 AQI 频率分布直方图")
lines(density(data2$AQI2),col="blue",lwd=4)
abline(v=100,lwd=2,col="green")
abline(v=120,lwd=4,col="yellow")
h$density

library(rdrobust)
a<-seq(from=1,to=486,by=1)
a
rdrobust(y=a,x=data2$AQI2,c=100)
rdplot(y=a,x=data2$AQI2,c=100)
```

结果显示：

```
> summary(data2)
          time              AQI2
 2014/11/30:  2   Min.   : 30.00
 2013/11/1 :  1   1st Qu.: 78.25
 2013/11/10:  1   Median :108.00
 2013/11/11:  1   Mean   :125.98
 2013/11/12:  1   3rd Qu.:159.00
 2013/11/13:  1   Max.   :423.00
 (Other)   :479
> h$density
 [1] 0.0006172840 0.0057613169 0.0067901235 0.0085390947 0.0077160494
0.0052469136
 [7] 0.0031893004 0.0022633745 0.0027777778 0.0019547325 0.0015432099
0.0013374486
[13] 0.0006172840 0.0005144033 0.0003086420 0.0004115226 0.0000000000
0.0002057613
[19] 0.0001028807 0.0000000000 0.0001028807
> rdrobust(y=a,x=data2$AQI2,c=100)
Call:
```

```
rdrobust(y = a, x = data2$AQI2, c = 100)

Summary:

Number of Obs 486
NN Matches        3
BW Type           mserd
Kernel Type       Triangular

                            Left       Right
Number of Obs           208         278
Order Loc Poly (p) 1            1
Order Bias (q)          2           2
BW Loc Poly (h)       15.1569 15.1569
BW Bias (b)             26.9750 26.9750
rho (h/b)               0.5619    0.5619
bias                    15.3683 -6.4482

Estimates:
                    Coef        Std. Err. z          P>|z|    CI Lower CI Upper
Conventional 76.4685 53.0197      1.4423 0.1492 -27.4481 180.3852
Robust                                          0.1169 -24.5725 221.1426
#上海
data3<-read.table("D:\\high\\R\\work\\建模实战\\3\\3_1.txt",header = TRUE)
#data3<-read.delim("clipboard",header = TRUE)
data3
summary(data3)
plot(data3$time,data3$AQI3,xlab="time",ylab="  上  海  AQI",main="  上  海
-AQI",type="p")
h<-hist(data3$AQI3,breaks=25,freq=FALSE,col = "red",xlab = "上海 AQI 指数",ylab
= "AQI 频率",main="上海 AQI 频率分布直方图")
lines(density(data3$AQI3),col="blue",lwd=4)
abline(v=100,lwd=3,col="green")
h$density

library(rdrobust)
a<-seq(from=1,to=486,by=1)
a
rdrobust(y=a,x=data3$AQI3,c=100)
rdplot(y=a,x=data3$AQI3,c=100)
结果显示：
> summary(data3)
              time                  AQI3
 2014/11/30:   2      Min.      : 15.00
```

```
 2013/11/1 :    1      1st Qu.: 42.00
 2013/11/10:    1      Median : 53.00
 2013/11/11:    1      Mean      : 60.91
 2013/11/12:    1      3rd Qu.: 70.00
 2013/11/13:    1      Max.      :358.00
 (Other)     :479
> h$density
 [1] 0.0016460905 0.0020576132 0.0181069959 0.0222222222 0.0179012346
0.0146090535
 [7] 0.0069958848 0.0045267490 0.0022633745 0.0022633745 0.0026748971
0.0012345679
[13] 0.0014403292 0.0004115226 0.0004115226 0.0002057613 0.0004115226
0.0002057613
[19] 0.0002057613 0.0000000000 0.0000000000 0.0000000000 0.0000000000
0.0000000000
[25] 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000
0.0000000000
[31] 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0002057613
> rdrobust(y=a,x=data3$AQI3,c=100)
Call:
rdrobust(y = a, x = data3$AQI3, c = 100)

Summary:

Number of Obs 486
NN Matches        3
BW Type           mserd
Kernel Type       Triangular

                            Left       Right
Number of Obs           439        47
Order Loc Poly (p) 1              1
Order Bias (q)          2              2
BW Loc Poly (h)        22.1042 22.1042
BW Bias (b)              37.1528 37.1528
rho (h/b)                0.5950    0.5950
bias                          -1.0601 -5.9477

Estimates:
                    Coef          Std. Err. z             P>|z|    CI Lower    CI Upper
Conventional -109.1019 112.1650    -0.9727 0.3307 -328.9412 110.7373
Robust                                                  0.4463 -372.3951 163.9664
#广州
data4<-read.table("D:\\high\\R\\work\\建模实战\\4\\4_1.txt",header = TRUE)
```

```
#data4<-read.delim("clipboard",header = TRUE)
data4
plot(data4$time,data4$AQI4,xlab="time",ylab="广州AQI",main="广州-AQI",type="p")
h<-hist(data4$AQI4,breaks=25,freq=FALSE,col = "red",xlab = "广州AQI指数",ylab = "AQI频率",main="广州AQI频率分布直方图")
lines(density(data4$AQI4),col="blue",lwd=4)
abline(v=100,lwd=3,col="green")
h$density

library(rdrobust)
a<-seq(from=1,to=486,by=1)
a
rdrobust(y=a,x=data4$AQI4,c=100)
rdplot(y=a,x=data4$AQI4,c=100)
```

结果显示：

```
> summary(data4)
              time                AQI4
 2014/11/30:   2     Min.      : 13.00
 2013/11/1 :   1     1st Qu.: 50.25
 2013/11/10:   1     Median : 68.00
 2013/11/11:   1     Mean      : 74.05
 2013/11/12:   1     3rd Qu.: 92.00
 2013/11/13:   1     Max.      :205.00
 (Other)     :479
> h$density
 [1] 0.0006172840 0.0024691358 0.0113168724 0.0106995885 0.0131687243 0.0146090535
 [7] 0.0111111111 0.0098765432 0.0076131687 0.0059670782 0.0034979424 0.0026748971
[13] 0.0016460905 0.0016460905 0.0006172840 0.0012345679 0.0010288066 0.0000000000
[19] 0.0000000000 0.0002057613
> rdrobust(y=a,x=data4$AQI4,c=100)
Call:
rdrobust(y = a, x = data4$AQI4, c = 100)

Summary:

Number of Obs 486
NN Matches       3
BW Type          mserd
Kernel Type      Triangular
```

```
                            Left       Right
Number of Obs               393        93
Order Loc Poly (p) 1               1
Order Bias (q)          2               2
BW Loc Poly (h)         23.2854 23.2854
BW Bias (b)             33.8278 33.8278
rho (h/b)               0.6883    0.6883
bias                    6.8955    2.4459

Estimates:
                    Coef        Std. Err. z           P>|z|    CI Lower CI Upper
Conventional 27.7116 48.1134      0.5760 0.5646 -66.5890 122.0121
Robust                                          0.5727 -79.5860 143.9083
```

问题二 R 中的天气状况绘制

```
#北京
data11<-read.table("D:\\high\\R\\work\\建模实战\\天气\\北京.txt",header = TRUE)
data11[,8]<-seq(from=1,to=485,by=1)
colnames(data11)[8]<-"sequence"
data11
summary(data11)
plot(data11$sequence,(data11$temperature)/10,'l',col="blue",lwd=2,main="北京气温图",xlab="序列",ylab="气温（摄氏度）")
plot(data11$sequence,(data11$dampness),'l',col="red",lwd=2,main=" 北 京 湿 度 图 ",xlab="序列",ylab="湿度（%）")
plot(data11$sequence,(data11$wind)/10,'l',col="green",lwd=2,main=" 北 京 风 速 图 ",xlab="序列",ylab="风速（米/秒）")
#天津
data22<-read.table("D:\\high\\R\\work\\建模实战\\天气\\天津.txt",header = TRUE)
data22[,8]<-seq(from=1,to=485,by=1)
colnames(data22)[8]<-"sequence"
data22
summary(data22)
plot(data22$sequence,(data22$temperature)/10,'l',col="blue",lwd=2,main="天津气温图",xlab="序列",ylab="气温（摄氏度）")
plot(data22$sequence,(data22$dampness),'l',col="red",lwd=2,main=" 天 津 湿 度 图 ",xlab="序列",ylab="湿度（%）")
plot(data22$sequence,(data22$wind)/10,'l',col="green",lwd=2,main=" 天 津 风 速 图 ",xlab="序列",ylab="风速（米/秒）")
#上海
data33<-read.table("D:\\high\\R\\work\\建模实战\\天气\\上海.txt",header = TRUE)
data33[,8]<-seq(from=1,to=485,by=1)
colnames(data33)[8]<-"sequence"
data33
summary(data33)
```

```
plot(data33$sequence,(data33$temperature)/10,'l',col="blue",lwd=2,main="上海气温图",xlab="序列",ylab="气温（摄氏度）")
plot(data33$sequence,(data33$dampness),'l',col="red",lwd=2,main="上海湿度图",xlab="序列",ylab="湿度（%）")
plot(data33$sequence,(data33$wind)/10,'l',col="green",lwd=2,main="上海风速图",xlab="序列",ylab="风速（米/秒）")
#广州
data44<-read.table("D:\\high\\R\\work\\建模实战\\天气\\广州.txt",header = TRUE)
data44[,8]<-seq(from=1,to=485,by=1)
colnames(data44)[8]<-"sequence"
data44
summary(data44)
plot(data44$sequence,(data44$temperature)/10,'l',col="blue",lwd=2,main="广州气温图",xlab="序列",ylab="气温（摄氏度）")
plot(data44$sequence,(data44$dampness),'l',col="red",lwd=2,main="广州湿度图",xlab="序列",ylab="湿度（%）")
plot(data44$sequence,(data44$wind)/10,'l',col="green",lwd=2,main="广州风速图",xlab="序列",ylab="风速（米/秒）")
```

问题二中的 R 回归代码及结果

```
#北京
d111<-read.table("D:\\high\\R\\work\\建模实战\\全\\111.txt",header = TRUE)
d111
summary(d111)
d111.lm<-lm(d111$AQI1~d111$PM2.5+d111$PM10+d111$CO+d111$NO2+d111$SO2)
summary(d111.lm)
d111.lm1<-lm(d111$AQI1~d111$PM2.5+d111$PM10+d111$NO2)
summary(d111.lm1)
#回归分析得到，只有 PM2.5，PM10，NO2,下面绘制出显著性污染的图像
plot(d111$sequence,d111$PM2.5,'l',col='red',xlab="序列",ylab="PM2.5",main="北京 PM2.5 变化图")
plot(d111$sequence,d111$PM10,'l',col='blue',xlab="序列",ylab="PM10",main="北京 PM10 变化图")
plot(d111$sequence,d111$NO2,'l',col='green',xlab="序列",ylab="NO2",main="北京 NO2 变化图")
结果显示：
>
d111.lm<-lm(d111$AQI1~d111$PM2.5+d111$PM10+d111$CO+d111$NO2+d111$SO2)
> summary(d111.lm)

Call:
lm(formula = d111$AQI1 ~ d111$PM2.5 + d111$PM10 + d111$CO + d111$NO2 +
    d111$SO2)
```

```
Residuals:
      Min          1Q    Median          3Q         Max
-71.357    -7.252    -1.457     6.164 113.822

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.47125       1.98081    16.393    < 2e-16 ***
d111$PM2.5      0.87551       0.02913    30.055    < 2e-16 ***
d111$PM10       0.33488       0.02377    14.087    < 2e-16 ***
d111$CO          2.11828       2.35841     0.898       0.370
d111$NO2       -0.46760       0.06213    -7.527    2.6e-13 ***
d111$SO2       -0.08663       0.05969    -1.451       0.147
---
Signif. codes:    0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.44 on 480 degrees of freedom
Multiple R-squared:    0.9535, Adjusted R-squared:    0.953
F-statistic:    1970 on 5 and 480 DF,    p-value: < 2.2e-16

> d111.lm1<-lm(d111$AQI1~d111$PM2.5+d111$PM10+d111$NO2)
> summary(d111.lm1)

Call:
lm(formula = d111$AQI1 ~ d111$PM2.5 + d111$PM10 + d111$NO2)

Residuals:
      Min          1Q    Median          3Q         Max
-69.660    -7.375    -1.542     5.932 114.337

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.02469       1.91520    17.243     <2e-16 ***
d111$PM2.5      0.88478       0.02541    34.815     <2e-16 ***
d111$PM10       0.33389       0.02356    14.175     <2e-16 ***
d111$NO2       -0.47288       0.04873    -9.704     <2e-16 ***
---
Signif. codes:    0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.44 on 482 degrees of freedom
Multiple R-squared:    0.9533, Adjusted R-squared:    0.953
F-statistic:    3282 on 3 and 482 DF,    p-value: < 2.2e-16

#天津
```

```
d222<-read.table("D:\\high\\R\\work\\建模实战\\全\\222.txt",header = TRUE)
d222
summary(d222)
d222.lm<-lm(d222$AQI2~d222$PM2.5+d222$PM10+d222$CO+d222$NO2+d222$SO2)
summary(d222.lm)
d222.lm1<-lm(d222$AQI2~d222$PM2.5+d222$PM10+d222$SO2)
summary(d222.lm1)
#回归分析得到，只有 PM2.5，PM10，SO2,下面绘制出显著性污染的图像
plot(d222$sequence,d222$PM2.5,'l',col='red',xlab=" 序 列 ",ylab="PM2.5",main=" 天津 PM2.5 变化图")
plot(d222$sequence,d222$PM10,'l',col='blue',xlab=" 序 列 ",ylab="PM10",main=" 天津 PM10 变化图")
plot(d222$sequence,d222$SO2,'l',col='green',xlab="序列",ylab="NO2",main="天津 SO2 变化图")
结果显示：
>
d222.lm<-lm(d222$AQI2~d222$PM2.5+d222$PM10+d222$CO+d222$NO2+d222$SO2)
> summary(d222.lm)

Call:
lm(formula = d222$AQI2 ~ d222$PM2.5 + d222$PM10 + d222$CO + d222$NO2 +
    d222$SO2)

Residuals:
     Min       1Q   Median       3Q      Max
-105.617   -7.416   -2.107    4.915  223.372

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.86244    2.96217   9.069  < 2e-16 ***
d222$PM2.5   0.74437    0.04809  15.479  < 2e-16 ***
d222$PM10    0.29732    0.03401   8.741  < 2e-16 ***
d222$CO      4.95815    2.47015   2.007 0.045285 *
d222$NO2    -0.19498    0.08319  -2.344 0.019498 *
d222$SO2    -0.14194    0.03764  -3.771 0.000183 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.39 on 480 degrees of freedom
Multiple R-squared:  0.9069, Adjusted R-squared:  0.9059
F-statistic: 934.9 on 5 and 480 DF,  p-value: < 2.2e-16
```

```
> d222.lm1<-lm(d222$AQI2~d222$PM2.5+d222$PM10+d222$SO2)
> summary(d222.lm1)

Call:
lm(formula = d222$AQI2 ~ d222$PM2.5 + d222$PM10 + d222$SO2)

Residuals:
      Min       1Q   Median       3Q      Max
-102.201   -7.820   -2.227    5.190  224.143

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.48168    1.97736  12.887  < 2e-16 ***
d222$PM2.5   0.74963    0.04410  16.998  < 2e-16 ***
d222$PM10    0.28886    0.03403   8.489 2.59e-16 ***
d222$SO2    -0.14290    0.02713  -5.267 2.10e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.52 on 482 degrees of freedom
Multiple R-squared:  0.9053, Adjusted R-squared:  0.9047
F-statistic:  1537 on 3 and 482 DF,  p-value: < 2.2e-16

#上海
d333<-read.table("D:\\high\\R\\work\\建模实战\\全\\333.txt",header = TRUE)
d333
summary(d333)
d333.lm<-lm(d333$AQI3~d333$PM2.5+d333$PM10+d333$CO+d333$NO2+d333$
SO2)
summary(d333.lm)
d333.lm1<-lm(d333$AQI3~d333$PM2.5+d333$PM10)
summary(d333.lm1)
#回归分析得到，只有 PM2.5,PM10,下面绘制出显著性污染的图像
plot(d333$sequence,d333$PM2.5,'l',col='red',xlab=" 序 列 ",ylab="PM2.5",main=" 上
海 PM2.5 变化图")
plot(d333$sequence,d333$PM10,'l',col='blue',xlab=" 序 列 ",ylab="PM10",main=" 上
海 PM10 变化图")
结果分析：
>
d333.lm<-lm(d333$AQI3~d333$PM2.5+d333$PM10+d333$CO+d333$NO2+d333$
SO2)
> summary(d333.lm)

Call:
```

```
lm(formula = d333$AQI3 ~ d333$PM2.5 + d333$PM10 + d333$CO + d333$NO2 +
    d333$SO2)

Residuals:
     Min       1Q   Median       3Q      Max
-91.797  -5.150  -1.488   3.277 157.676

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.96434    2.73000   6.580 1.23e-10 ***
d333$PM2.5   0.87167    0.05019  17.366  < 2e-16 ***
d333$PM10    0.17736    0.03534   5.018 7.35e-07 ***
d333$CO      5.31838    5.66476   0.939    0.348
d333$NO2    -0.03480    0.05698  -0.611    0.542
d333$SO2    -0.04745    0.08606  -0.551    0.582
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.35 on 480 degrees of freedom
Multiple R-squared:  0.9116, Adjusted R-squared:  0.9107
F-statistic: 990.2 on 5 and 480 DF,  p-value: < 2.2e-16

> d333.lm1<-lm(d333$AQI3~d333$PM2.5+d333$PM10)
> summary(d333.lm1)

Call:
lm(formula = d333$AQI3 ~ d333$PM2.5 + d333$PM10)

Residuals:
     Min       1Q   Median       3Q      Max
-90.468  -5.068  -1.702   3.666 158.632

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.49307    1.28202  15.205  < 2e-16 ***
d333$PM2.5   0.89544    0.03893  23.004  < 2e-16 ***
d333$PM10    0.16669    0.03273   5.092 5.08e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.32 on 483 degrees of freedom
Multiple R-squared:  0.9114, Adjusted R-squared:  0.911
F-statistic:  2484 on 2 and 483 DF,  p-value: < 2.2e-16
```

```
#广州
d444<-read.table("D:\\high\\R\\work\\建模实战\\全\\444.txt",header = TRUE)
d444
summary(d444)
d444.lm<-lm(d444$AQI4~d444$PM2.5+d444$PM10+d444$CO+d444$NO2+d444$
SO2)
summary(d444.lm)
d444.lm1<-lm(d444$AQI4~d444$PM2.5+d444$NO2)
summary(d444.lm1)
#回归分析得到，只有 PM2.5,NO2,下面绘制出显著性污染的图像
plot(d333$sequence,d333$PM2.5,'l',col='red',xlab=" 序 列 ",ylab="PM2.5",main=" 广
州 PM2.5 变化图")
plot(d333$sequence,d333$NO2,'l',col='blue',xlab=" 序 列 ",ylab="NO2",main=" 广 州
NO2 变化图")
结果分析：
>
d444.lm<-lm(d444$AQI4~d444$PM2.5+d444$PM10+d444$CO+d444$NO2+d444$
SO2)
> summary(d444.lm)

Call:
lm(formula = d444$AQI4 ~ d444$PM2.5 + d444$PM10 + d444$CO + d444$NO2 +
      d444$SO2)

Residuals:
      Min          1Q    Median          3Q         Max
-49.385    -3.044    -0.634     2.330    32.997

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.35847       1.32194    13.888    < 2e-16 ***
d444$PM2.5      1.12150       0.05041    22.248    < 2e-16 ***
d444$PM10       0.09387       0.04008     2.342 0.019587 *
d444$CO        -2.18663       1.69535    -1.290 0.197748
d444$NO2       -0.10931       0.02903    -3.765 0.000187 ***
d444$SO2       -0.02877       0.04880    -0.590 0.555776
---
Signif. codes:    0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.325 on 479 degrees of freedom
Multiple R-squared:    0.9609, Adjusted R-squared:    0.9605
F-statistic:    2353 on 5 and 479 DF,    p-value: < 2.2e-16

> d444.lm1<-lm(d444$AQI4~d444$PM2.5+d444$NO2)
```

```
> summary(d444.lm1)

Call:
lm(formula = d444$AQI4 ~ d444$PM2.5 + d444$NO2)

Residuals:
      Min        1Q    Median        3Q       Max
-48.551   -3.183   -0.779    2.652   30.772

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.94310       0.77274   21.926   < 2e-16 ***
d444$PM2.5   1.22773       0.01757   69.862   < 2e-16 ***
d444$NO2    -0.11056       0.02299   -4.809 2.03e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.368 on 482 degrees of freedom
Multiple R-squared:  0.9602, Adjusted R-squared:    0.96
F-statistic:  5810 on 2 and 482 DF,  p-value: < 2.2e-16
```

Matlab 中的断点回归分析代码

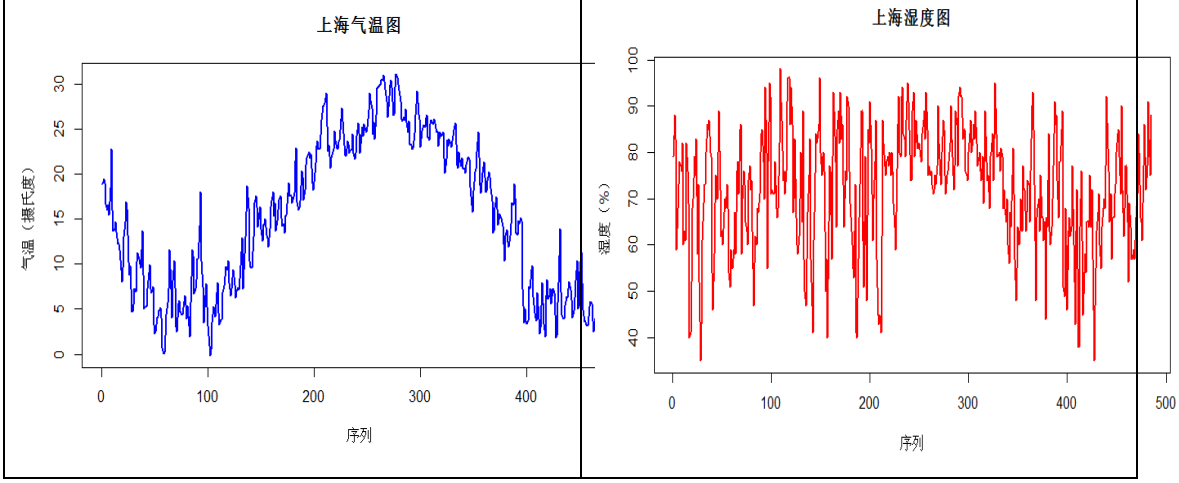
```
function k = gaussian_kernel(x,Weight_speed)
% Weight_speed  越大，权值变化越慢

k = (1/(sqrt(2*pi)))*exp(-(x^2)/(2*Weight_speed^2));
function [fit_x,fit_y] = local_linear_fit(x,y,h_bandwidth)
% local_linear_fit:local linear regression
%     [fit_x,fit_y] = local_linear_fit(x,y,h_bandwidth),x: independent
%     variable,y:dependent variable,h_bandwidth:bandwidth
%
% ncf,July,2016
% Email:1476879092@qq.com
% log:
% 2016-7-12:Complete

% column matrix
test_sample_x = x;
test_sample_y = y;

test_sample_X = [ones(length(test_sample_x),1) test_sample_x];

num_x = length(test_sample_x);
% Weight_speed
Weight_speed = 1;
```

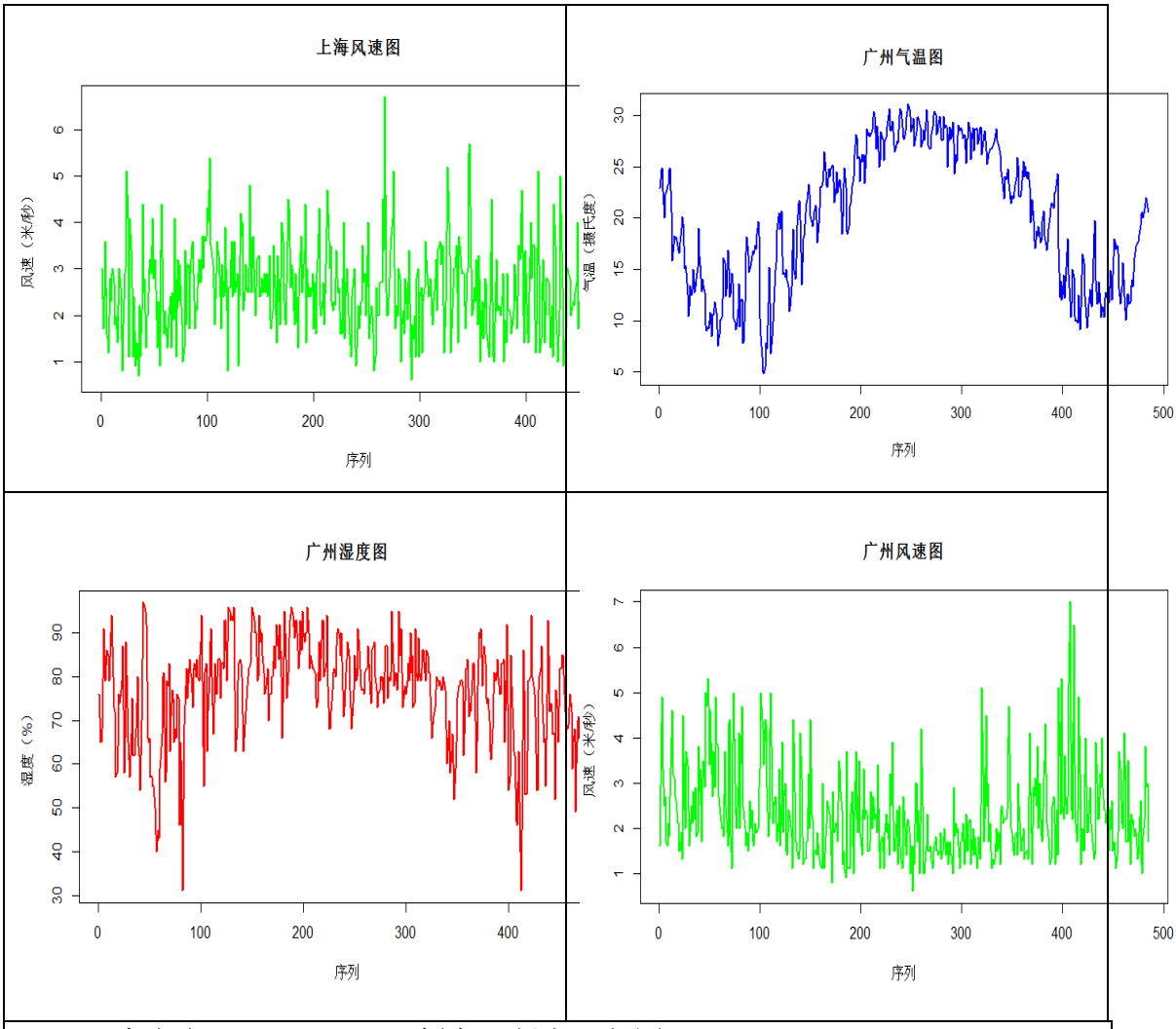
```
%% linear fit
% line_fit_beta = (test_sample_X'*test_sample_X)\(test_sample_X'*test_sample_y);
% line_fit_y = line_fit_beta(1) + line_fit_beta(2)*test_sample_x;
% plot(test_sample_x,line_fit_y,'b')
%% local linear fit
% fit_x and fit_y
fit_x = min(test_sample_x):0.3:max(test_sample_x);
fit_y = zeros(length(test_sample_x),length(h_bandwidth));

% colors = ['m' 'c' 'k' 'r' 'g'];

for k_bandwidth = 1:length(h_bandwidth)
      h = h_bandwidth(k_bandwidth);
      for k_fit_y = 1:length(fit_x)
            w = zeros(num_x,num_x);
            K_h_all = zeros(num_x,1);
            % compute K_h
            for k_w = 1:num_x
                  K_h_all(k_w)                                                    =
gaussian_kernel((fit_x(k_fit_y)-test_sample_x(k_w))/h,Weight_speed)/h;
            end
            sum_K_h_all = sum(K_h_all);
            % compute w
            for k_w = 1:num_x
                w(k_w,k_w) = K_h_all(k_w)./sum_K_h_all;
            end
            local_beta                                                        =
(test_sample_X'*w*test_sample_X)\(test_sample_X'*w*test_sample_y);
            fit_y(k_fit_y,k_bandwidth) = local_beta(1)+local_beta(2)*fit_x(k_fit_y);
      end
%         plot(fit_x,fit_y,colors(k_bandwidth))
end
% legend('trainingdata','linear','r=.1','r=.3','r=.8,','r=2','r=10');
fit_x = [fit_x',fit_x',fit_x'];
clc
clear
load('C:\Users\lrw\Desktop\数据.mat');

%% 局部线性拟合
AQI=C{1}(:,1);
data = AQI(1:481);

cut_off =100;
```

```
x_lim_all = [min(AQI),max(AQI)];

% 获取左侧的样本点
data_low = data(data<cut_off);
bin_width_low = 2*std(data_low)/sqrt(length(data_low));%std 计算标准差，默认
n-1,默认按列
bin_num_low = (cut_off-min(data_low))/bin_width_low;
[y_low,x_low] = hist(data_low,bin_num_low);
figure(1);
hist(data_low,bin_num_low);%频率直方图
y_low=y_low/length(AQI)/mean(diff(x_low)); %概率密度

% 获取 右侧的样本点
data_up = data(data>cut_off);
bin_width_up = 2*std(data_up)/sqrt(length(data_up));
bin_num_up = (max(data_up)-cut_off)/bin_width_up;
[y_up,x_up] = hist(data_up,bin_num_up);
figure(2);
hist(data_up,bin_num_up);
y_up=y_up/length(AQI)/mean(diff(x_up));

y_lim_all = [min([min(y_low) min(y_up)]) max([max(y_low) max(y_up)])];

% 拟合
% 绘制样本点
colors = ['m' 'c' 'k' 'r' 'g'];
figure;
hold on;
plot([x_low x_up],[y_low y_up],'.b');%散点图

% 局部线性拟合
% 左侧拟合
h_bandwidth = [5*bin_width_low 15*bin_width_low 20*bin_width_low]; %
bandwidth
[x_low_fit,y_low_fit] = local_linear_fit(x_low',y_low',h_bandwidth);
for k = 1:length(h_bandwidth)
    plot(x_low_fit(:,k),y_low_fit(:,k),colors(k))
end
%  右侧拟合
h_bandwidth = [10*bin_width_up 15*bin_width_up 20*bin_width_up]; % bandwidth
[x_up_fit,y_up_fit] = local_linear_fit(x_up',y_up',h_bandwidth);
for k = 1:length(h_bandwidth)
    plot(x_up_fit(:,k),y_up_fit(:,k),colors(k))
end
```

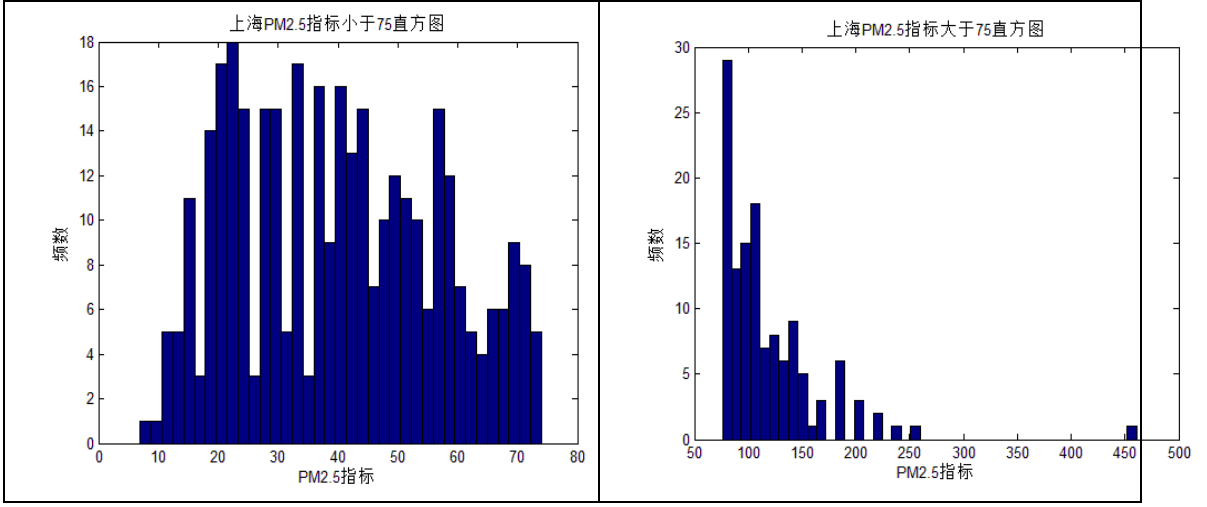
```
legend('trainingdata','a=10,','a=15','a=20')
axis([x_lim_all*1.1 y_lim_all*1.1])
% %% 鲁棒性带宽过大过小的比较
% figure;
% hold on;
% plot([x_low x_up],[y_low y_up],'.b');
% % 局部线性拟合
% h_bandwidth = [1*bin_width_low 15*bin_width_low 30*bin_width_low]; % bandwidth
% local_linear_fit(x_low',y_low',h_bandwidth)
%
% h_bandwidth = [1*bin_width_up 15*bin_width_up 30*bin_width_up]; % bandwidth
% local_linear_fit(x_up',y_up',h_bandwidth)
%
% legend('trainingdata','a=1,','a=15','a=30')
% axis([x_lim_all*1.1 y_lim_all*1.1])
%%%
%%%
%%%
%%%
%%%
%%%
%%%
%%%
%%%
%%%
```

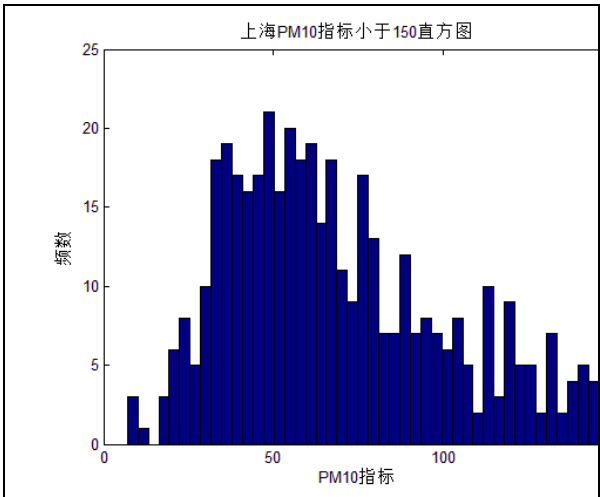
R 中的广州与上海气候图

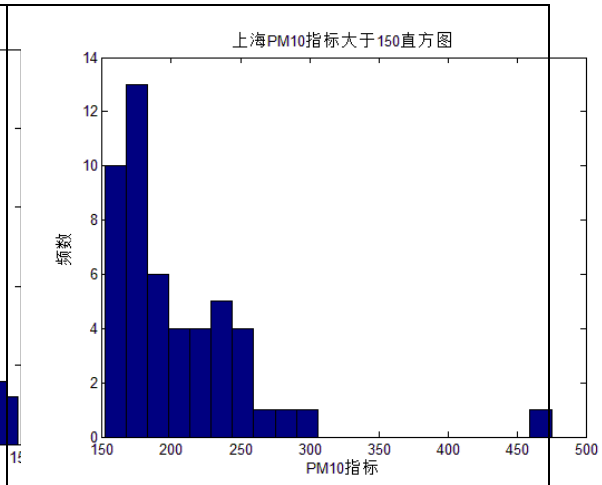


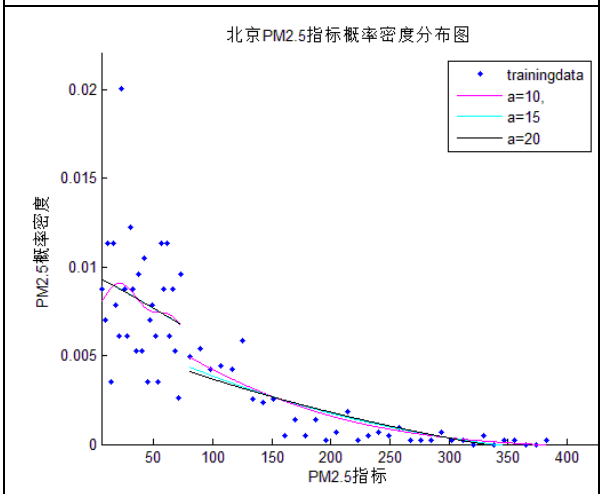


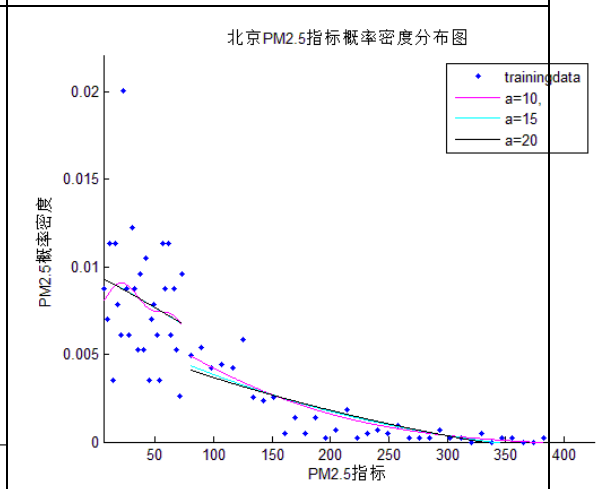
Matlab 中上海 PM2.5、PM10 频率及断点回归图



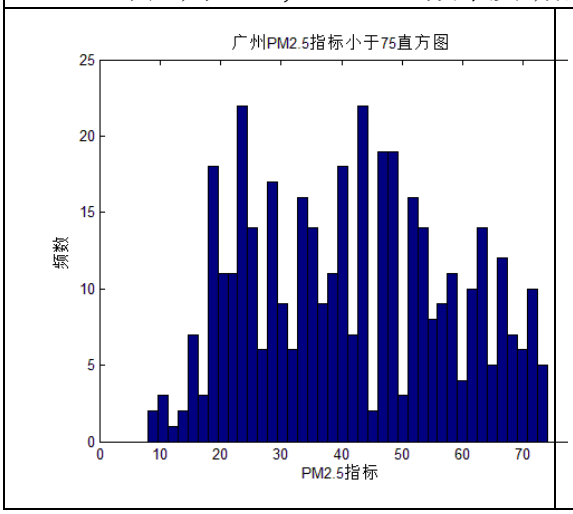


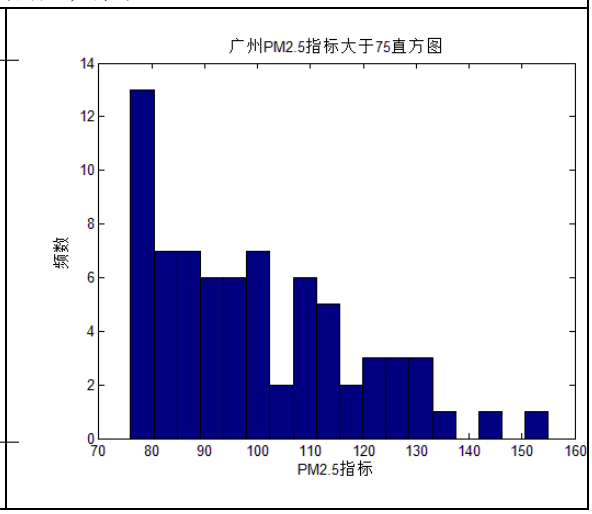


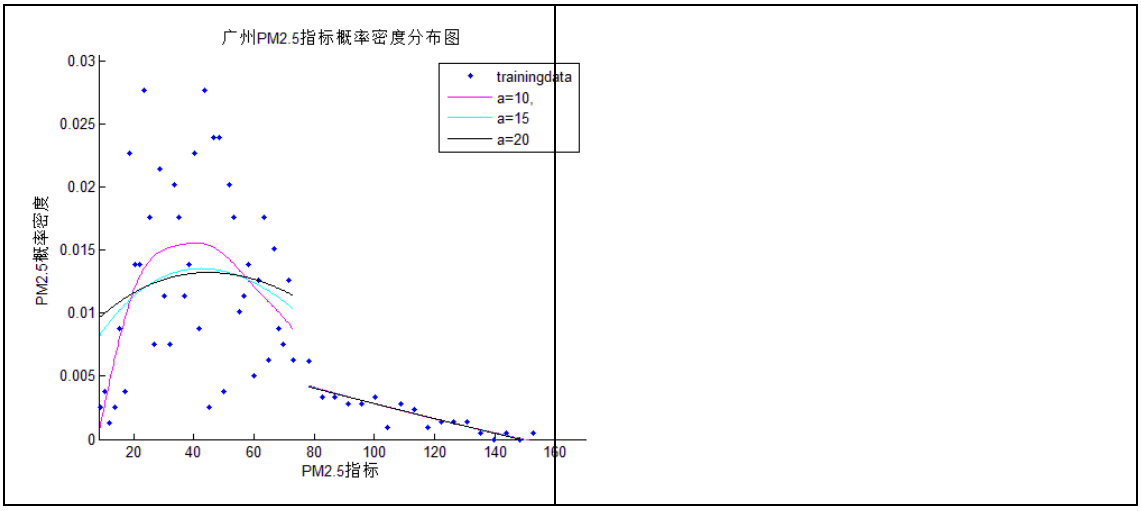




Matlab 中广州 PM2,5、PM10 频率及断点回归图






广州PM2.5指标概率密度分布图
trainingdata
a=10,
a=15
a=20
PM2.5概率密度
PM2.5指标
0.03
0.025
0.02
0.015
0.01
0.005
0
20
40
60
80
100
120
140
160