

# 空气污染与R软件

张兵

Email: zhangbing4502431@outlook.com

Homepage: <http://spatial-r.github.io/>

浙江省疾病预防控制中心

2014 年 12 月 6 日

# 目录

## 空气污染-简介

空气污染

监测指标

空气质量指数

## 空气污染-数据获取

## 空气污染-描述性分析

## 空气污染-健康效应

## 联系作者

空气污染与R软件

张兵

### 2 空气污染-简介

空气污染

监测指标

空气质量指数

### 空气污染-数据获取

数据来源

获取方式

### 空气污染-描述性分析

时间维度

空间维度

### 空气污染-健康效应

研究方法

分布非线性滞后模型

模型选择

具体示例

### 联系作者

# 空气污染

空气污染与R软件

张兵

空气污染-简介

3

空气污染

监测指标

空气质量指数

空气污染-数据获取

数据来源

获取方式

空气污染-描述性分析

时间维度

空间维度

空气污染-健康效应

研究方法

分布非线性滞后模型

模型选择

具体示例

联系作者

Zhejiang Provincial Center  
for Disease Control and  
Prevention

- ▶ 空气污染(Air pollution)通常是指由于人类活动或自然过程引起某些物质进入大气中，呈现出足够的浓度，达到足够的时间，并因此危害了人类的舒适、健康和福利或环境的现象。
- ▶ 空气污染的主要来源是工业、锅炉、交通运输等。



(a) 雾中有太极



(b) 长跑须防毒

# 监测指标

空气质量标准(GB: 3095-2012), 监测指标及其浓度限值:

污染物项目	平均时间	浓度限值		单位
		一级	二级	
二氧化硫 ( $\text{SO}_2$ )	年平均	20	60	$\mu\text{g}/\text{m}^3$
	24 小时平均	50	150	
	1 小时平均	150	500	
	年平均	40	40	
二氧化氮 ( $\text{NO}_2$ )	24 小时平均	80	80	$\mu\text{g}/\text{m}^3$
	1 小时平均	200	200	
	24 小时平均	4	4	
	1 小时平均	10	10	
臭氧 ( $\text{O}_3$ )	日最大 8 小时平均	100	160	$\mu\text{g}/\text{m}^3$
	1 小时平均	160	200	
颗粒物 (粒径小于等于 $10 \mu\text{m}$ )	年平均	40	70	$\mu\text{g}/\text{m}^3$
	24 小时平均	50	150	
颗粒物 (粒径小于等于 $2.5 \mu\text{m}$ )	年平均	15	35	$\mu\text{g}/\text{m}^3$
	24 小时平均	35	75	

- ▶ 一级标准适合于自然保护区、风景名胜区和其他需要特殊保护的区域。
- ▶ 二级标准适合与居民区、商业交通居民混合区、文化区、工业区和农村地区。

空气污染与R软件

张兵

空气污染-简介

空气污染

监测指标

空气质量指数

空气污染-数据获取

数据来源

获取方式

空气污染-描述性分析

时间维度

空间维度

空气污染-健康效应

研究方法

分布非线性滞后模型

模型选择

具体示例

联系作者

Zhejiang Provincial Center  
for Disease Control and  
Prevention

# 空气质量指数

空气污染与R软件

张兵

空气污染-简介

空气污染

监测指标

空气质量指数

5

空气污染-数据获取

数据来源

获取方式

空气污染-描述性分析

时间维度

空间维度

空气污染-健康效应

研究方法

分布非线性滞后模型

模型选择

具体示例

联系作者

- ▶ 空气质量指数(Air Quality Index, AQI)是定量描述空气质量状况的非线性无量纲指数，数值越大，说明空气污染状况越严重。其计算方式为：

$$AQI = \max\{IAQI_1, IAQI_2, IAQI_3, \dots, IAQI_n\} \quad (1)$$

式中：IAQI为空气质量分指数；n为污染物项目。

- ▶ 污染物项目P的空气质量分指数( $IAQI_p$ )的计算方式如下：

$$IAQI_p = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{Lo}}(C_p - BP_{Lo}) + IAQI_{Lo} \quad (2)$$

式中： $C_p$ 为污染物P的浓度； $BP_{Hi}$ 和 $BP_{Lo}$ 分别为国家环境保护标准HJ633-2012中与 $C_p$ 相近的污染物浓度的高位值和低位值； $IAQI_{Hi}$ 和 $IAQI_{Lo}$ 分别是与 $BP_{Hi}$ 和 $BP_{Lo}$ 相对应的空气质量分指数。

# 目录

## 空气污染-简介

## 空气污染-数据获取

### 数据来源

### 获取方式

## 空气污染-描述性分析

## 空气污染-健康效应

## 联系作者

空气污染与R软件

张兵

### 空气污染-简介

空气污染

监测指标

空气质量指数

### 6 空气污染-数据获取

数据来源

获取方式

### 空气污染-描述性分析

时间维度

空间维度

### 空气污染-健康效应

研究方法

分布非线性滞后模型

模型选择

具体示例

### 联系作者

# 数据来源

空气污染与R软件

张兵

空气污染-简介

空气污染

监测指标

空气质量指数

空气污染-数据获取

数据来源

获取方式

空气污染-描述性分析

时间维度

空间维度

空气污染-健康效应

研究方法

分布非线性滞后模型

模型选择

具体示例

联系作者

Zhejiang Provincial Center  
for Disease Control and  
Prevention

7

- ▶ 国家环保局数据中心: 城市逐日空气质量指数数据(起始时间为2014年1月1日)
- ▶ 各省及各市的环境保护局, 如武汉市环保局, 10个国控点各个污染物分指数数据。
- ▶ 数据存放在相应GIS平台上, 如广东省环境信息综合发布平台及亚洲空气质量公布平台。
- ▶ 某些网站会无偿公开空气污染数据的API接口, 如PM25.in.
- ▶ 公共信息平台上, 如微博AQI及微博PM<sub>2.5</sub>。

22

# 数据获取-1

空气污染与R软件

张兵

空气污染-简介

空气污染

监测指标

空气质量指数

空气污染-数据获取

数据来源

获取方式

空气污染-描述性分析

时间维度

空间维度

空气污染-健康效应

研究方法

分布非线性滞后模型

模型选择

具体示例

联系作者

Zhejiang Provincial Center  
for Disease Control and  
Prevention

总的原则：利用RCurl或(和)XML解析网页，提取数据。

网页中有历史数据，直接用代码读取；

实时更新的网页，在R软件中设置循环，整点读取数据。

▶ 国家环保局数据中心，代码如下：

```
library(XML) +  
  
url<-"http://datacenter.mep.gov.cn/report/air_daily/air_dairy.jsp?"+  
city=北京市 &startdate=2014-01-01&enddate=2014-06-05&page=1" +  
dat<-readHTMLTable(readLines(url, warn=FALSE)) +  
dat.final<-dat[[4]][-c(1,32,33,34,35),-c(1,7)] +
```

▶ 改变city、startdate、enddate和page参数值，进而获取不同城市在不同时间段空气质量指数数据。

8

22

# 数据获取-2

空气污染与R软件

张兵

- ▶ 广东省环境信息综合发布平台，代码<sup>1</sup>如下：

```
require(RCurl); require(XML); require(rjson)
t = as.numeric(as.POSIXct("2012-10-22 15:34:56", "%Y-%m-%d %H:%M:%S"))
t = as.character(621355968000000000 + (t * 1000 * 10000))
x = getURI(paste("http://www-app.gdepb.gov.cn/silvergis2/pm25.mvc/GetAirPointDataByItemId?itemId=6&_t=' , t, sep = ""))
y = gsub(pattern = "\\\\"EndMoniTime\\\\" : new Date\\\[\\\"]*\\\"\\\",\\r\\n", x = x, replacement = "")
d = fromJSON(y)$data
area<-sapply(d,function(data){as.character(data[2][1])}) ##### 获取城市信息
site<-sapply(d,function(data){as.character(data[3][1])}) ##### 获取站点信息
long<-sapply(d,function(data){as.character(data[7][1])}) ##### 站点经度信息
lat<-sapply(d,function(data){as.character(data[8][1])}) ##### 站点纬度信息
dat.lasthour<-as.character(sapply(d,function(data){unlist(data$LastHourData[5]$Data)})) ##### 最近1小时浓度值
dat.last24<-as.character(sapply(d,function(data){unlist(data$Last24HourData[4]$Data)})) ##### 最近24小时浓度值
dat.final<-data.frame(area,site,long,lat,dat.lasthour,dat.last24) ##### 最终结果
```

- ▶ 改变itemId参数的值，即可获得其他空气污染物数据。

itemId	1	2	3	4	5	6	7	8
指标	SO <sub>2</sub>	NO <sub>2</sub>	O <sub>3</sub> 1h	O <sub>3</sub> 8h	PM <sub>10</sub>	PM <sub>2.5</sub>	AQI	CO

<sup>1</sup>感谢肖楠大神

空气污染-简介

空气污染

监测指标

空气质量指数

空气污染-数据获取

数据来源

获取方式

空气污染-描述性分析

时间维度

空间维度

空气污染-健康效应

研究方法

分布非线性滞后模型

模型选择

具体示例

联系作者

# 数据获取-3

空气污染与R软件

张兵

空气污染-简介

空气污染

监测指标

空气质量指数

空气污染-数据获取

数据来源

获取方式

空气污染-描述性分析

时间维度

空间维度

空气污染-健康效应

研究方法

分布非线性滞后模型

模型选择

具体示例

联系作者

Zhejiang Provincial Center  
for Disease Control and  
Prevention

10

```
library(XML); library(plyr)  
url="http://www.pm25.in/api/querys/all_cities.json?token=5j1znBVAsnSf5xQyNQyq"  
pm=getURL(url); pm2<-fromJSON(pm)  
pm3=data.frame(do.call("rbind",pm2)); pm4=apply(pm3,2,as.character)  
dat.final=gsub("-", "", substr(pm4[1,21],1,13))
```

获取全国所有站点空气质量数据的示例代码如下：

需注意两点：

- ▶ 需要申请密钥，上述代码中的密钥是公钥。
- ▶ API调用次数限制：1.10和1.11每小时15次、1.12每小时5次、1.13每小时15次，其余每小时500次。

22

# 目录

空气污染-简介

空气污染-数据获取

空气污染-描述性分析

时间维度

空间维度

空气污染-健康效应

联系作者

空气污染与R软件

张兵

空气污染-简介

空气污染

监测指标

空气质量指数

空气污染-数据获取

数据来源

获取方式

11 空气污染-描述性分析

时间维度

空间维度

空气污染-健康效应

研究方法

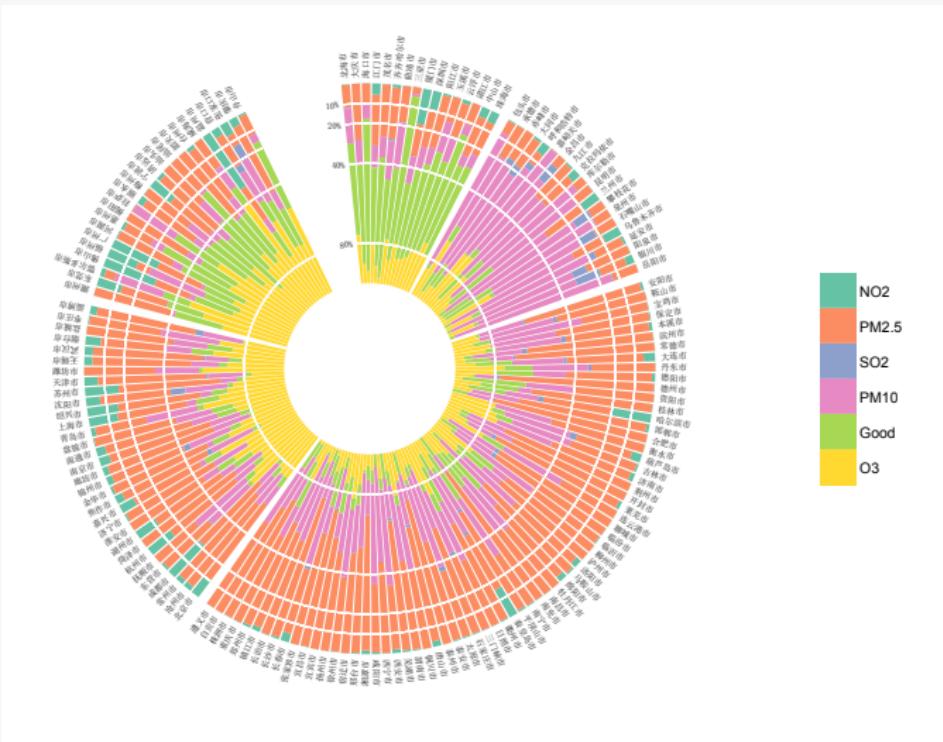
分布非线性滞后模型

模型选择

具体示例

联系作者

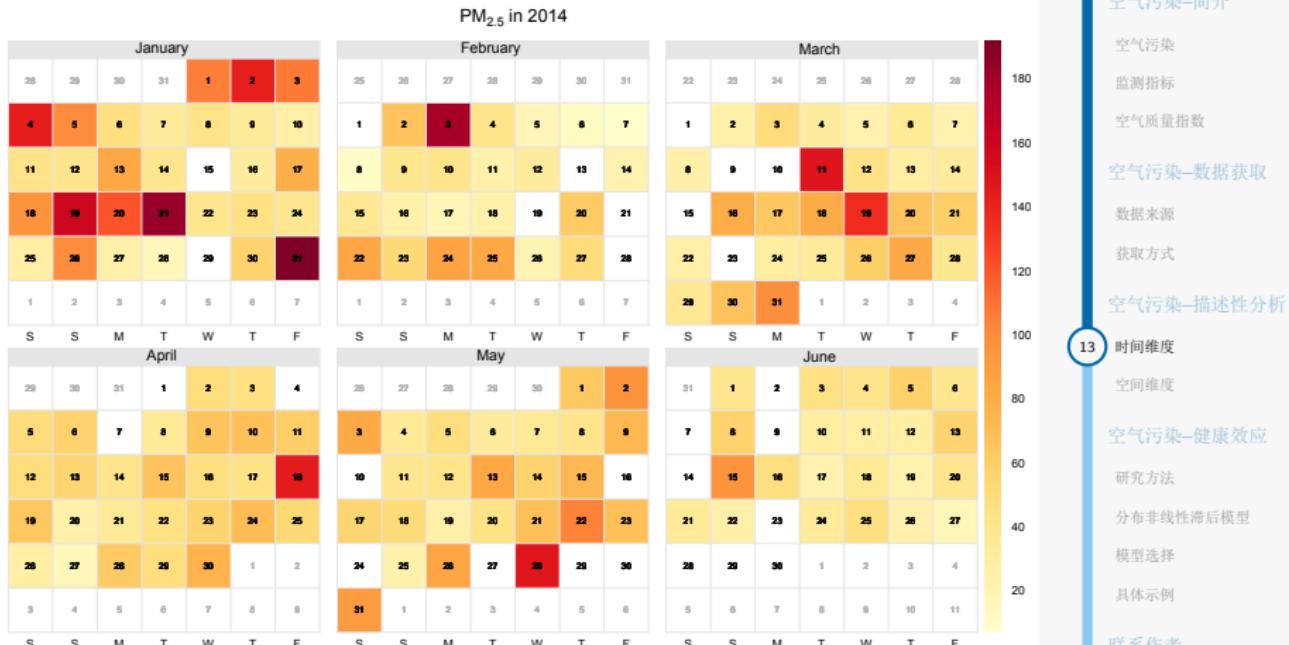
22 Zhejiang Provincial Center  
for Disease Control and  
Prevention



全国空气质量一览，PM<sub>10</sub>、PM<sub>2.5</sub>和臭氧是主要的污染物！

# 日历图: openair

## 上海市PM<sub>2.5</sub>日均浓度



空气污染与R软件

张兵

空气污染-简介

空气污染

监测指标

空气质量指数

空气污染-数据获取

数据来源

获取方式

空气污染-描述性分析

时间维度

空间维度

空气污染-健康效应

研究方法

分布非线性滞后模型

模型选择

具体示例

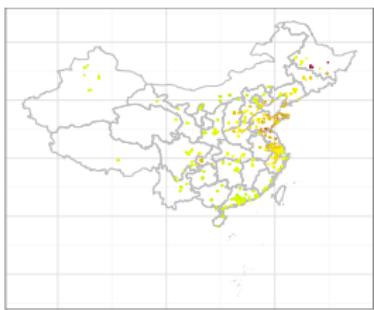
联系作者

Zhejiang Provincial Center  
for Disease Control and  
Prevention

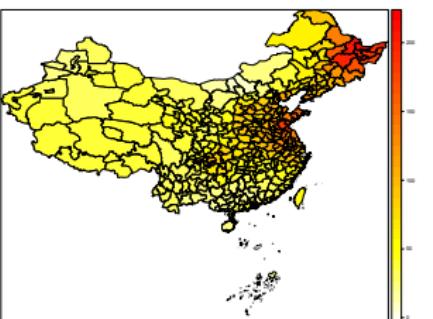
22

# 空间维度: ggplot2和automap

2014年11月11日PM<sub>2.5</sub>日均浓度:



(c) 日均浓度



(d) 空间插值

近45.6%的监测站点PM<sub>2.5</sub>日均浓度超过了空气质量二级标准限值，主要集中在山东和江苏地区。光棍相亲需慎重，既面临被拒的风险，也面临着空气污染的毒害！

# 目录

空气污染-简介

空气污染与R软件

张兵

空气污染-数据获取

空气污染-简介

空气污染

监测指标

空气质量指数

空气污染-描述性分析

空气污染-数据获取

数据来源

获取方式

空气污染-健康效应

空气污染-描述性分析

时间维度

空间维度

研究方法

15

空气污染-健康效应

研究方法

分布非线性滞后模型

分布非线性滞后模型

模型选择

模型选择

具体示例

具体示例

联系作者

联系作者

空气污染所致健康危害主要体现在两个方面：急性危害和慢性危害。

- ▶ 急性危害的研究方法主要有时间序列研究、病例交叉研究和面板研究。
- ▶ 慢性危害的研究方法主要是面板研究和队列研究。

其中，时间序列和病例交叉两类研究被广泛应用于探讨空气污染对人群的健康危害。原因如下：

- ▶ 对数据的要求相对而言要宽松。空气污染物日均浓度数据和健康危害的终点指标(日累计死亡或患病人数)。
- ▶ 分析方法相对成熟且更易操作。广义线性或广义相加模型结合滞后项，如分布非线性滞后模型。

空气污染所致健康危害主要体现在两个方面：急性危害和慢性危害。

- ▶ 急性危害的研究方法主要有时间序列研究、病例交叉研究和面板研究。
- ▶ 慢性危害的研究方法主要是面板研究和队列研究。

其中，时间序列和病例交叉两类研究被广泛应用于探讨空气污染对人群的健康危害。原因如下：

- ▶ 对数据的要求相对而言要宽松。空气污染物日均浓度数据和健康危害的终点指标(日累计死亡或患病人数)。
- ▶ 分析方法相对成熟且更易操作。广义线性或广义相加模型结合滞后项，如分布非线性滞后模型。

16

研究方法  
分布非线性滞后模型  
模型选择  
具体示例  
联系作者

▶ 分布非线性滞后模型 (distributed lag non-linear model, DLNM)在2006年首次应用于探讨气温的健康效应，随后Gasparrini(主页在这)等在广义相加模型的基础上，利用交叉基过程，重新构建了DLNM的理论和框架，并开发了dlnm包。

▶ DLNM的基本结构：

$$g(\mu_t) = \alpha + \sum_{j=1}^J f_j(x_{ij}; \beta_j) + \sum_{k=1}^K \gamma_k u_{tk} \quad (3)$$

$f_j$ 是表示自变量 $x_j$ 的基函数，常用的基函数为正交函数、线性阈值函数和样条函数。

▶ 分布非线性滞后模型既考虑了暴露与反应之间的非线性关系，又能分析暴露因素对反应的滞后效应。

17

空气污染-简介

空气污染

监测指标

空气质量指数

空气污染-数据获取

数据来源

获取方式

空气污染-描述性分析

时间维度

空间维度

空气污染-健康效应

研究方法

分布非线性滞后模型

模型选择

具体示例

联系作者



- ▶ 数据来源： dlnm包自带的chicagoNMMAPS数据集，包含了1987-2000年芝加哥日总死亡人数、日均温度、日均PM<sub>10</sub>浓度等数据。
- ▶ 研究目的：探讨PM<sub>10</sub>日均浓度对日总死亡人数的影响。
- ▶ 示例代码如下：

```
library(dlnm); library(splines)#
data(chicagoNMMAPS); chicagoNMMAPS$pm10<-na.locf(chicagoNMMAPS$pm10)#
varknots<-equalknots(chicagoNMMAPS$pm10,fun="ns",df=4,degree=2)#
lagknots<-logknots(27,3)#
cb3.pm<-crossbasis(chicagoNMMAPS$pm10,lag=27,argvar=list(fun="thr",#
    th=50),arglag=list(knots=lagknots)) #####假设阈值为 50#
model3<-glm(death~cb3.pm+ns(temp,3)+ns(time,7*14)+dow,#
    family=quasipoisson()),chicagoNMMAPS) #####时间序列方法#
strata30<-floor((chicagoNMMAPS$time-min(chicagoNMMAPS$time))/30) #####建立分层#
model3<-glm(death~cb3.pm+ns(temp,3)+dow+as.factor(strata30),#
    family=quasipoisson()),chicagoNMMAPS) #####病例交叉方法#
pred3.temp<-crosspred(cb3.pm,model3,at=c(0:180))#
plot(pred3.temp,xlab="Concentration",zlab="RR",theta=200,phi=40,lphi=30,main="")#
plot(pred3.temp,"contour",xlab="Concentration ",key.title=title("RR"),#
    plot.title=title("",xlab="Concentration ",ylab="Lag"))#
plot.title(title("",xlab="Concentration ",ylab="Lag"))#
```

19 具体示例

联系作者

22

空气污染-简介

空气污染

监测指标

空气质量指数

空气污染-数据获取

数据来源

获取方式

空气污染-描述性分析

时间维度

空间维度

空气污染-健康效应

研究方法

分布非线性滞后模型

模型选择

# 示例结果

空气污染与R软件

张兵

空气污染-简介

空气污染

监测指标

空气质量指数

空气污染-数据获取

数据来源

获取方式

空气污染-描述性分析

时间维度

空间维度

空气污染-健康效应

研究方法

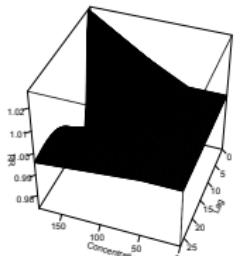
分布非线性滞后模型

模型选择

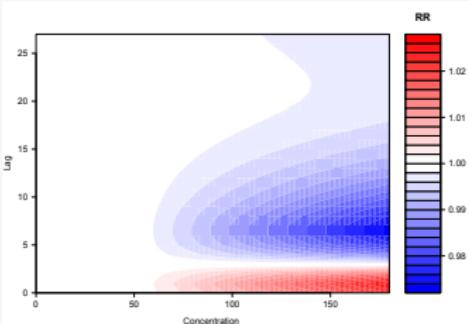
具体示例

联系作者

Zhejiang Provincial Center  
for Disease Control and  
Prevention



(e) 3D plot



(f) Contour plot

总的说来， $PM_{10}$ 浓度超过 $50\mu g/m^3$ 时，相对危险度(RR)会随着 $PM_{10}$ 浓度的升高而增大，也就是说由空气污染引发的死亡风险会增大。

20

22

# 目录

## 空气污染-简介

空气污染与R软件

张兵

## 空气污染-数据获取

空气污染-简介

空气污染

监测指标

空气质量指数

## 空气污染-描述性分析

空气污染-数据获取

数据来源

获取方式

## 空气污染-健康效应

空气污染-描述性分析

时间维度

空间维度

## 联系作者

空气污染-健康效应

研究方法

分布非线性滞后模型

模型选择

具体示例

21 联系作者

Zhejiang Provincial Center  
for Disease Control and  
Prevention

22

# 联系作者

空气污染与R软件

张兵

空气污染-简介

空气污染

监测指标

空气质量指数

空气污染-数据获取

数据来源

获取方式

空气污染-描述性分析

时间维度

空间维度

空气污染-健康效应

研究方法

分布非线性滞后模型

模型选择

具体示例

22 联系作者