北京市气象数据可视化
实验报告

A 000 B 001

July 1, 2015

1 实验目标

自2013年北京第一次出现雾霾以来，北京的空气质量备受世人关注，同时也是关乎每位老师同学身心健康的大事。近年来，北京不断添加空气质量监测设备、不断完善数据公开制度，让每位市民都能实时了解到北京天气如何。

时间已经过去了将近两年，数据也已经积累了很多，但是大多数人也还是停留在翻看一下今天的PM2.5是多少的习惯上，很少见到有人对历史数据进行研究整理。所以想借本课程的机会，研究一下北京的空气质量数据。

北京空气质量由北京市环境保护监测中心负责，通过北京各区县设立的35个监测站进行实时测量。目前收集的空气质量数据包括：$PM2.5$、$PM10$、AQI、SO_2、NO_2、O_3、CO 每小时一次的实际值以及它们各自在过去24小时的平均值，共计14个指标。本方案就是针对这些数据进行可视化分析。

我们的研究目标整体上分三层：
1、对官方数据进行清洗并存储；
2、任意数据的精确查询，并实现任意时点数据的可视化、任意时间段数据的动态展示；
3、在此基础上进行空气质量变化规律的分析。

2 数据介绍

北京空气质量可以实时地通过http://zx.bjmemc.com.cn/网站查看，但是该网站并不负责提供历史数据，所以需要自行抓取数据。网站提供的数据包括从20131205至今的每小时测量一次的污染物浓度数据。已经拿到的数据格式是：

```
date,     hour, type,      东四, 天坛, 官园, 万寿西宫,...
20131205, 1,    PM2.5,     93,  93,   63,  79,...
20131205, 1,    PM2.5_24h, 93,  108,  99,  123,...
20131205, 1,    PM10,      103, 124,  81,  107,...
20131205, 1,    PM10_24h,  97,  130,  122, 141,...
20131205, 1,    AQI,       123, 141,  130, 161,...
```

Figure 1: 已获取的数据格式

数据含义是：

Type	数据类型	单位
PM2.5	PM2.5实时浓度	(微克/立方米)
PM2.5_24h	PM2.5 24小时均值	(微克/立方米)
PM10	PM10实时浓度	(微克/立方米)
PM10_24h	PM10 24小时均值	(微克/立方米)
AQI	AQI实时值	N/A
SO2	SO2实时浓度	(微克/立方米)
SO2_24h	SO2 24小时均值	(微克/立方米)
NO2	NO2实时浓度	(微克/立方米)
NO2_24h	NO2 24小时均值	(微克/立方米)
O3	O3实时浓度	(微克/立方米)
CO	CO实时浓度	(毫克/立方米)
CO_24h	CO 24小时均值	(毫克/立方米)

北京市共35个监测站，涵盖所有的16个区县，每个区县最多3个，最少1个。35个监测站的地理位置与详细信息是：

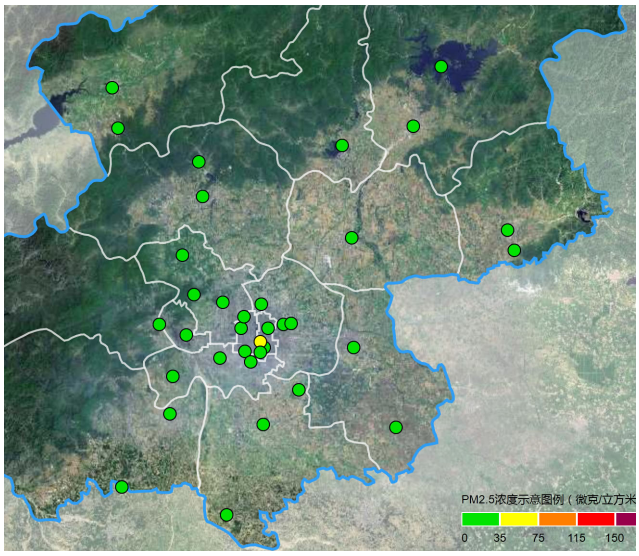


Figure 2: *35*个监测站的位置

城区环境评价点12个			
监测点	**监测点全称**	**经度**	**纬度**
东四	东城东四	116.417	39.929
天坛	东城天坛	116.407	39.886
官园	西城官园	116.339	39.929
万寿西宫	西城万寿西宫	116.352	39.878
奥体中心	朝阳奥体中心	116.397	39.982
农展馆	朝阳农展馆	116.461	39.937
万柳	海淀万柳	116.287	39.987
北部新区	海淀北部新区	116.174	40.09
植物园	海淀北京植物园	116.207	40.002
丰台花园	丰台花园	116.279	39.863
云岗	丰台云岗	116.146	39.824
古城	石景山古城	116.184	39.914

郊区环境评价点11个			
监测点	**监测点全称**	**经度**	**纬度**
房山	房山良乡	116.136	39.742
大兴	大兴黄村镇	116.404	39.718
亦庄	亦庄开发区	116.506	39.795
通州	通州新城	116.663	39.886
顺义	顺义新城	116.655	40.127
昌平	昌平镇	116.23	40.217
门头沟	门头沟龙泉镇	116.106	39.937
平谷	平谷镇	117.1	40.143
怀柔	怀柔镇	116.628	40.328
密云	密云镇	116.832	40.37
延庆	延庆镇	115.972	40.453

对照点及区域点7个			
监测点	**监测点全称**	**经度**	**纬度**
定陵	昌平定陵	116.22	40.292
八达岭	京西北八达岭	115.988	40.365
密云水库	京东北密云水库	116.911	40.499
东高村	京东东高村	117.12	40.10
永乐店	京东南永乐店	116.783	39.712
榆垡	京南榆垡	116.30	39.52
琉璃河	京西南琉璃河	116.00	39.58

交通污染监控点5个			
监测点	**监测点全称**	**经度**	**纬度**
前门	前门东大街	116.395	39.899
永定门内	永定门内大街	116.394	39.876
西直门北	西直门北大街	116.349	39.954
南三环	南三环西路	116.368	39.856
东四环	东四环北路	116.483	39.939

3 可视化方案

使用百度地图API展示数据，最终以Web形式作为交互界面。Web包含四个页面：

3.1 页面一：各区县污染物平均浓度展示

各区县内观测站数据求平均作为整个区县的数据，划分污染等级后用图层透明度展示污染物浓度。每个污染物根据《中国环境空气质量标准GB3095-1996》划分为6个等级，等级越低空气越好。具体如下：

污染物	等级1上限	等级2上限	等级3上限	等级4上限	等级5上限
PM2.5	35	75	115	150	250
PM10	50	150	250	350	420
AQI	50	100	150	200	300
SO2	150	500	650	800	1600
NO2	100	200	700	1200	2340
O3	160	200	300	400	800
CO	5	10	35	60	90

这个可视化视图下可以进行确定时间点的展示，也可以选择时间区间展示动画，效果如下：

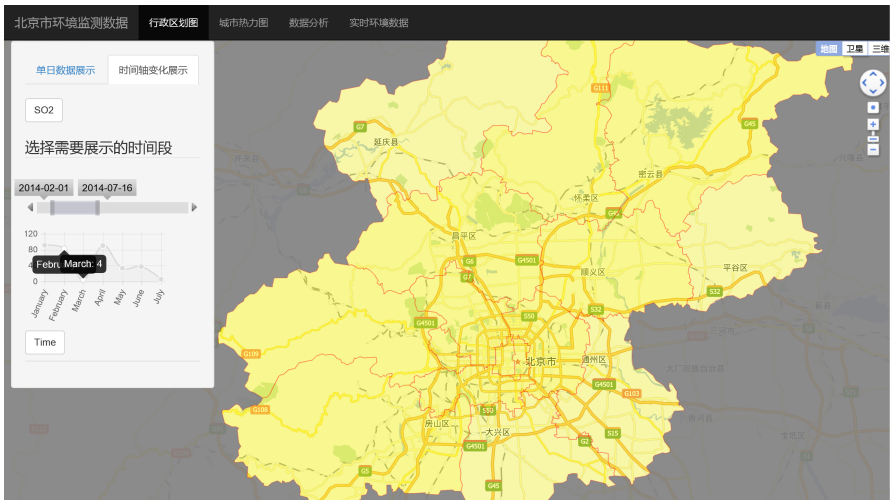


Figure 3: 页面一

3.2 页面二：各观测点污染物浓度热力图展示

各观测站数据直接做热力点，用点的大小和颜色一同展示浓度。该可视化视图下也可以进行确定时间点的展示，以及选择时间区间展示动画，效果如下：

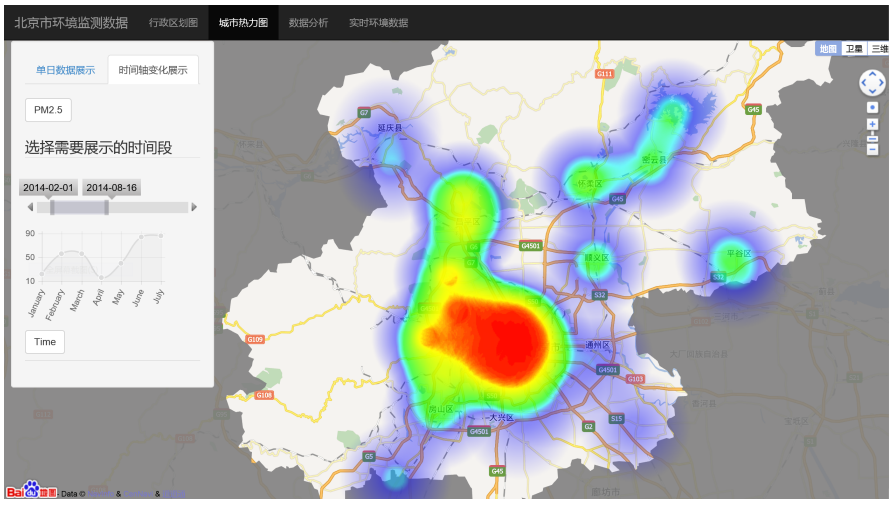


Figure 4: 页面二

3.3 页面三：统计分析

进行两种数据分析：单日24小时的7种污染物浓度曲线变化、任意时间段的某种确定污染物的浓度曲线变化。单日分析中需要指定确定的日期、确定的观测站；时间段的分析需要制定时间段、确定的污染物类型，效果如下：

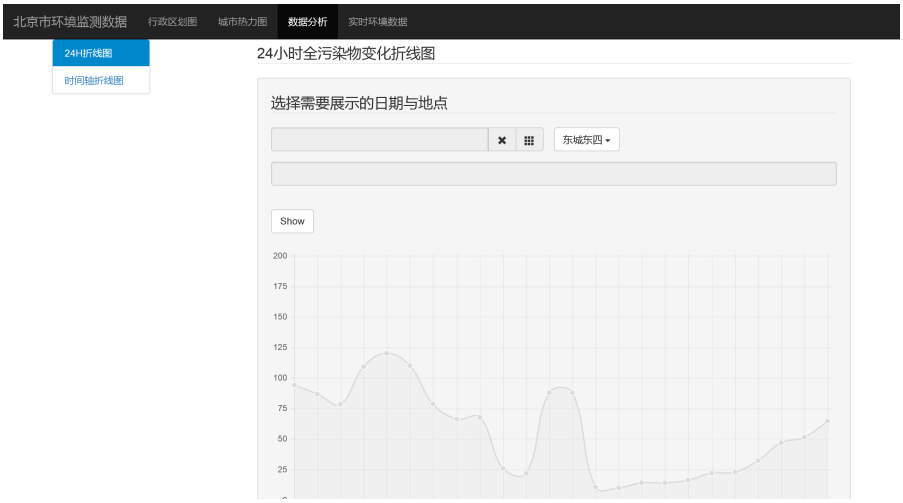


Figure 5: 页面三

3.4 页面四：精确数据查询

任意时间点、观测站、污染物类型的精确数据查询，除时间点外，均可以多选，效果如下：

Figure 6: 页面四

4 可视化技术

4.1 数据清洗

项目建立的基础是北京市环境监测网站提供的数据文件，数据以.csv的形式保存。其格式为：

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	date	hour	type	涓滃洓	澶╁潧	瀹樺洯	涓囧 瑗	濂ヤ綋涓	鍐滃睍棣	鍖楅儴鏂	妞嶇墿鍥	涓拌彴	浜ф棴	鍦垮北	澶у叴	浜︿簞	閫氬窞
2	20131205	0	PM2.5														
3	20131205	0	PM2.5_24h														
4	20131205	0	PM10														
5	20131205	0	PM10_24h														
6	20131205	0	AQI	126	145	134	167	113	123	155	147	72	219	143	145	216	284
7	20131205	1	PM2.5	93	93	63	79	59	108	85	169	49	72	90	109	80	128
8	20131205	1	PM2.5_24h	93	108	99	123	82	92	115	113	53	163	109	113	164	231

数据的前3列分别存放对应日期、时间和污染物类型，之后的35列分别对应35个观测站。鉴于原始数据本身存在相当程度的缺失等异常情况，并存在一定程度的不合理数据，对于这些数据进行数据清洗工作存在其必要性。

首先是确定对于缺失数据和错误数据的修复策略。我们的选择是建立平均值矩阵，对应每一个观测站，每一种污染物，每一个时间段，建立对应的平均值矩阵，作为需要填充时的依据。

矩阵的部分结果为：

```
public final static double[][] aves = {{87.84,128.1,121.85,17.84,53.79,58.37,1.29},
    {86.06,122.01,118.15,13.96,52.28,62.37,1.21},
    {86.3,124.71,118.38,17.07,60.15,59.01,1.25},
    {88.25,138.21,121.71,15.88,58.56,62.08,1.3},
    {86.88,139.94,120.28,15.49,61.55,57.02,1.13},
    {90.03,131.42,123.84,18.57,60.84,63.39,1.26},
    {88.31,141.14,122.78,18.32,68.88,45.03,1.24},
    {87.62,124.02,119.04,17.28,46.96,48.36,1.44},
    {78.21,98.75,108.77,13.61,37.44,75.67,1.02},
    {99.26,152.69,133.72,17.91,62.21,52.57,1.37},
```

对于异常数据，去除了数值异常大，明显不合理的数据，例如2013年底，50000数值的PM2.5指标等。

4.2 数据库建设

本身采用.csv文件存储的环境监测数据集大约有21M左右，但是考虑到总共469天的检测区间，7中污染物，35个监测站，24小时的环境数据，一共会有24*35*7*469，大约2757720行的环境数据，单纯使用内存的方法，例如HashMap等难以满足复杂的查询需求，所以建立了数据库后台，用于支撑数据的相关操作。数据库使用MySQL进行建立，使用两张表来进行数据的存储。

imformation.bp_position：存放监测站的地理位置等信息

#	Field	Schema	Table	Type	Character Set	Display Size	Precision	Scale
1	Dist_ID	imformation	bp_position	VARCHAR	utf8	45	2	0
2	Dist_Name	imformation	bp_position	VARCHAR	utf8	45	12	0
3	Dist_All	imformation	bp_position	VARCHAR	utf8	45	21	0
4	Nation_Name	imformation	bp_position	VARCHAR	utf8	45	15	0
5	longitude	imformation	bp_position	VARCHAR	utf8	45	7	0
6	latitude	imformation	bp_position	VARCHAR	utf8	45	6	0

imformation.bp_data：存放具体的监测数据信息

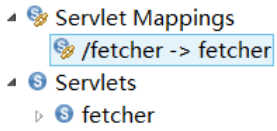
#	Field	Schema	Table	Type	Character Set	Display Size	Precision	Scale
1	dict_id	imformation	bp_data	VARCHAR	utf8	45	2	0
2	date	imformation	bp_data	VARCHAR	utf8	45	8	0
3	hour	imformation	bp_data	VARCHAR	utf8	45	1	0
4	type	imformation	bp_data	VARCHAR	utf8	45	9	0
5	rate	imformation	bp_data	VARCHAR	utf8	45	6	0

使用这两张表将原始数据完成清洗之后的结果加以存储，完成数据的支撑工作。

4.3 服务器搭建

鉴于我们的UI展示会采用网站的形式，需要完成服务器的搭建工作。服务器的构架是较为常用的Serverlet服务形式，使用Java语言完成后台处理代码的编写，并使用Jsp作为网页的编码形式，同时采用Tomcat作为网站的发布载体，完成网站服务器的搭建。

Serverlet结构截图：



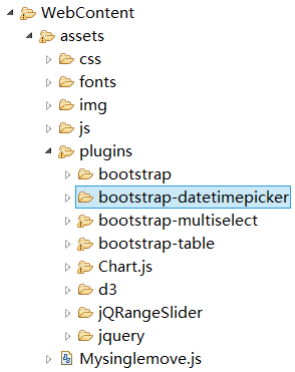
Tomcat服务器运行截图：

```
Tomcat v8.0 Server at localhost [Apache Tomcat] C:\Program Files\Java\jdk1.7.0_51\bin\javaw.exe (2015年6月30日 上午10:21:36)
六月 30, 2015 10:21:42 上午 org.apache.tomcat.util.digester.SetPropertiesRule begin
警告: [SetPropertiesRule]{Server/Service/Engine/Host/Context} Setting property 'source'
六月 30, 2015 10:21:42 上午 org.apache.catalina.startup.VersionLoggerListener log
信息: Server version:        Apache Tomcat/8.0.17
六月 30, 2015 10:21:42 上午 org.apache.catalina.startup.VersionLoggerListener log
信息: Server built:          Jan 9 2015 15:58:59 UTC
六月 30, 2015 10:21:42 上午 org.apache.catalina.startup.VersionLoggerListener log
信息: Server number:         8.0.17.0
```

4.4 网站UI设计与实现

网页的UI采用了常用的Bootstrap项目结构，除了常规的Bootstrap和Jquery的相关内容之外，鉴于网站的内容，还采用了以下插件：
bootstrap-datetimepicker：提供日期的选择；
bootstrap-multiselect：提供条件复选；
bootstrap-table：提供表格的展示；
Chart.js：提供图标的展示；
jQRangeSlider：提供日期段的选取。

这些插件都存在于asset文件夹目录之下：



4.5 百度地图API

百度地图JavaScript API是一套由JavaScript语言编写的应用程序接口，可在网站中构建功能丰富、交互性强的地图应用，支持PC端和移动端基于浏览器的地图应用开发，且支持HTML5特性的地图开发。也是本次可视化中地图部分主要依赖的工具。

百度地图在中国有很好的链接速度，而且API完全免费、不限制使用次数，只需在developer.baidu.com上注册后，获得密钥即可。具体可见http://developer.baidu.com/map/index.php?title=jspopular。

在页面一区县数据可视化中，实际上是在北京市各区县范围上分别添加多边形覆盖物BMap.Polygon，并将覆盖物颜色设为yellow，透明度设为污染物浓度等级。动画则是重复这一过程，并在每次重复前清空之前的覆盖物。

在页面二热力图中，直接使用百度地图API的扩展库中的BMapLib.HeatmapOverlay。

页面三、页面四中未使用百度地图API。

5 数据分析

利用页面三进行了多种数据分析，其中一些成果包括：

5.1 PM_{10}、$PM_{2.5}$污染物的全年浓度变化

从下图中可以看到，两种污染物虽然属于一个家族，但是变化规律也完全不一样。PM_{10}全年在4月份有一个特别高的峰值，但是其余时间浓度平稳；PM2.5则波动巨大，基本上以2个月为周期大起大落。

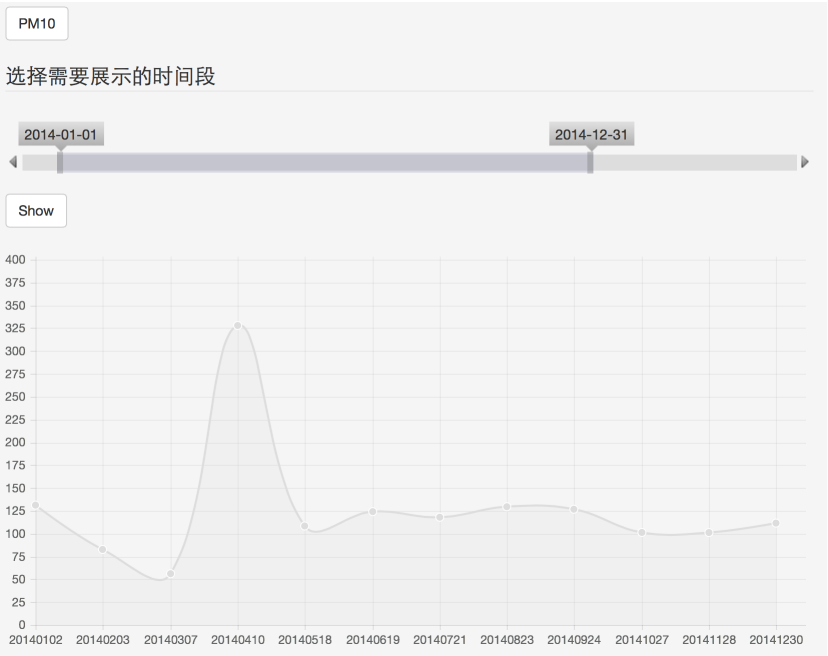


Figure 7: *PM10*全年浓度变化

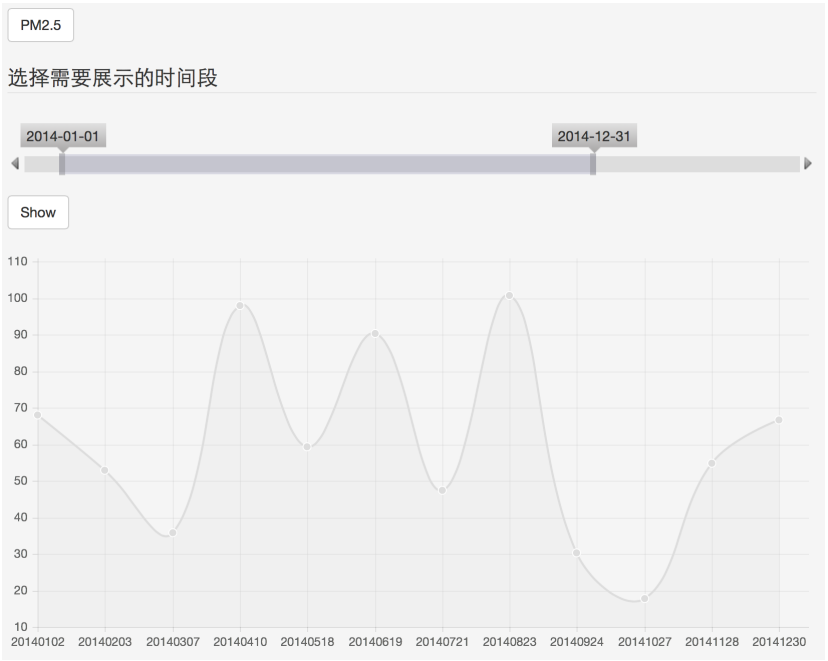


Figure 8: *PM2.5*全年浓度变化

5.2 APEC会议期间空气质量

APEC会议时间是2014年11月5日至11日，华北地区全面减排是从11月3号开始的，这个月北京的PM2.5浓度变化如下。除了11月9日浓度稍有上升以外，整个会议期间空气质量都是很好的。而会议结束后空气质量的明显变差也提醒了我们，空气质量也是可防可控的。

5.3 冬季供暖的影响

北京供暖期是每年的11月15日至次年的3月15日，一共4个月，为了进行对比，把这个时间段扩展一倍，即前延到9月15日，后延到5月15日，考察这段时间的$PM_{2.5}$ 浓度。可以看到，除了APEC会议期间，供暖期的空气质量是明显糟于其余时段的。

5.4 年度之间的变化

受数据时间范围的限制（仅拥有2013年12月至2015年6月的数据），所以只对2013年12月和2014年12月的空气质量进行了对比。注意到2013年图中的纵轴坐标与2014年中的不同，整体曲线形状也证明了2014年空气质量优于2013年，这也是一个可喜的结果。

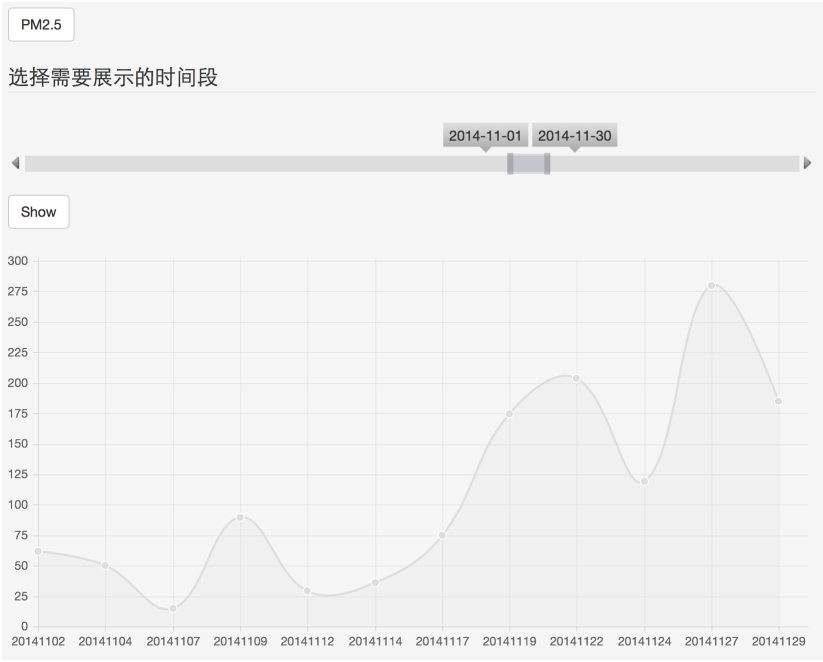


Figure 9: *APEC*会议期间空气质量

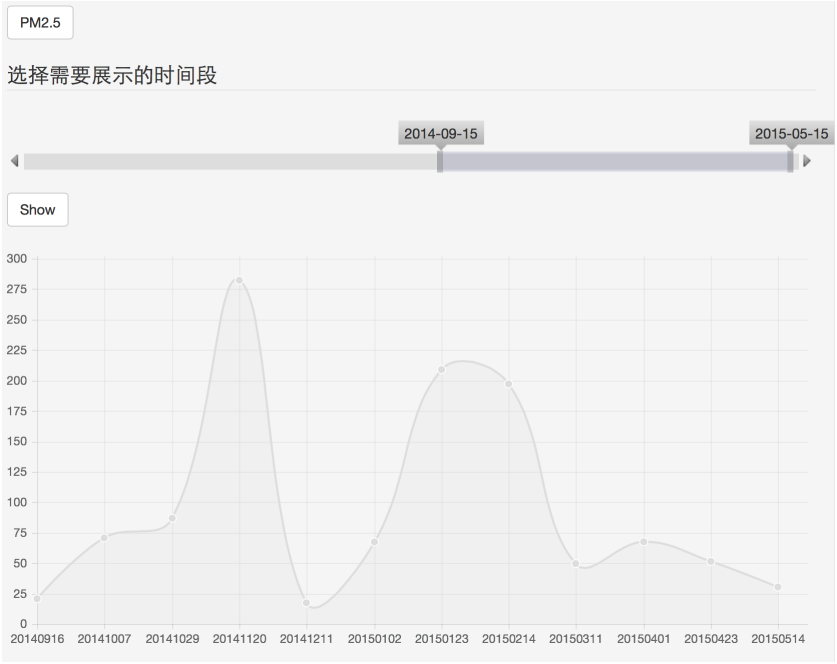


Figure 10: 冬季供暖对空气质量的影响

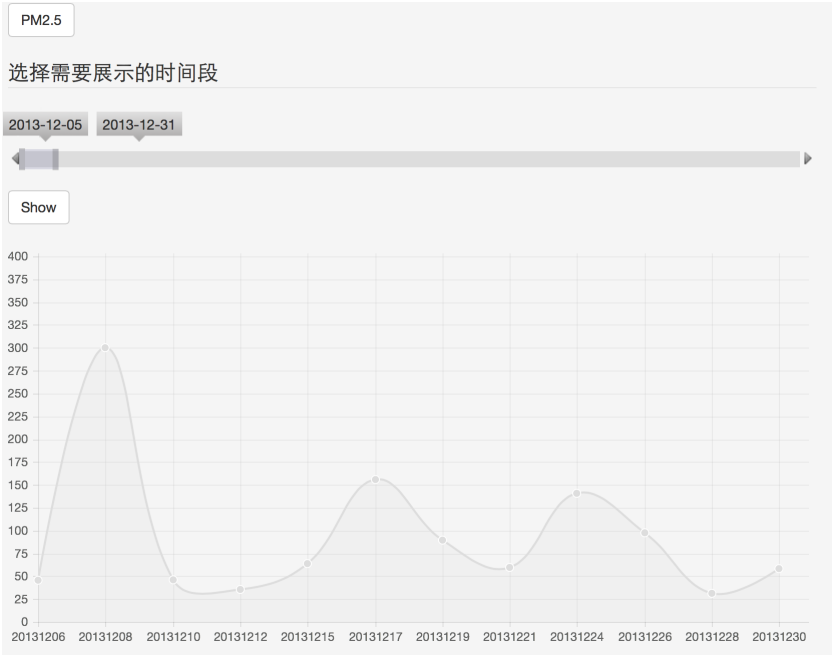


Figure 11: *2013*年*12*月空气质量

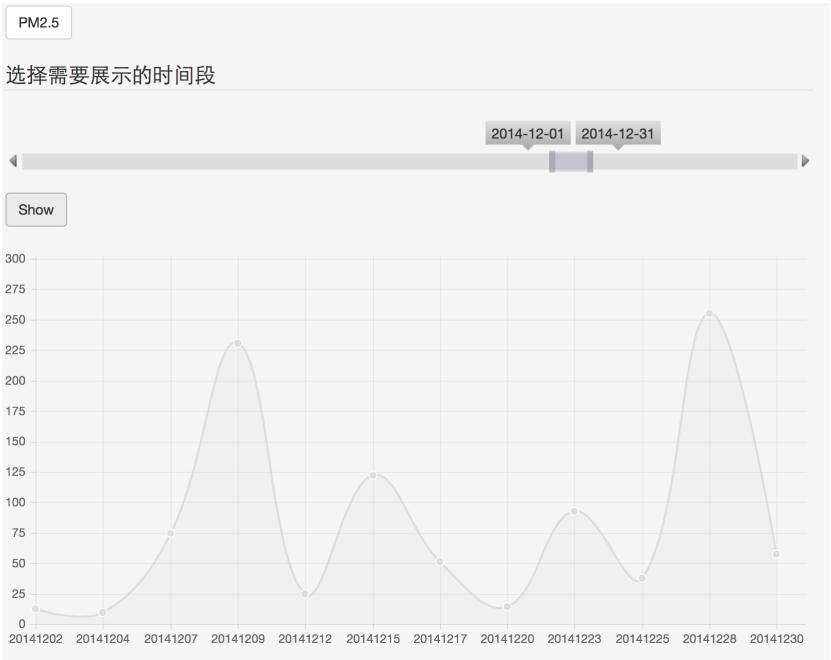


Figure 12: *2014*年*12*月空气质量