**2018年河北省首届研究生数学建模竞赛**

题 目 河北省中南部空气质量预报模型研究

摘 要：

随着我国工业化和城市化进程的蓬勃发展，区域性大范围的大气污染日趋严重，制约了社会经济的可持续发展，对人类的生态健康和生态环境安全也造成了极大威胁。京津冀地区空气污染问题尤为突出，河北省中南部空气污染最为严重。空气污染防控已成为我国政府关注的热点之一。如何获取与污染物相关性较好的气象因子以采取相应地防控措施成为环境保护关注的重点问题之一。本文考虑了影响河北省中南部不同城市不同季节空气污染的气象因子，建立了合理的相关性气象因子选择模型、空气污染物相邻两日浓度差值的预报模型、不同季节的污染物预报模型等。对上述模型进行验证，并给出合理化的建议。

问题一，选取与污染物浓度相关性较好的气象因子。首先进行数据预处理，通过分析数据集以及题目内容，创新性地增加三个属性值，分别表示：1）温度持续上升；2）相邻两日湿度维持在50%以上；3）相邻两日风速差值。按污染物浓度采集时间为主键集成气象数据，静稳指数以及其他因子数据；通过计算欧氏距离对缺失值插补；采用分段均值填充缺失记录。然后采用2014-2017的数据进行问题分析。由于不同地域不同季节污染物浓度与气象因子的相关关系不同，故按照不同城市不同季节分别分析以获取其相关性较好的气象因子。最后采用皮尔森系数来验证与已有相关研究的一致性，并确定污染物与气象因子之间是否是线性相关。同时结合信息增益（IG）和互信息对数据进行特征选择，分析确定属性贡献度以获取与污染物相关性较好的气象因子。

问题二，构建空气污染物相邻两日浓度差值的预报模型。首先构建两个模型：浓度差值预报模型和浓度预报模型，后者根据预测值再计算差值。通过对比BaseLine实验结果选择构建各城市浓度预报模型进行优化。然后使用数据预处理后的新增属性的数据集，构建CART（Classification and Regression Tree）决策树及随机森林（Random Forest）模型进行预报。最后通过GridSearch自动搜寻模型最优参数，模型在PM2.5的预测误差值为44*ug/m3*。

问题三，分不同季节建立污染物模型。首先通过数据统计分析发现，不同城市各污染物浓度在四个季度上平均值的变化趋势表现一致（呈现春冬高，夏秋低的趋势），然后通过对比两个城市建立各自的BaseLine模型以及两个城市共同建立的BaseLine模型预报结果发现，后者的预报效果更加理想，因此我们选择后者构建CART和RF模型，同样使用GridSearch自动搜索最优参数。实验结果显示夏秋两季模型比春冬两季模型效果要好。

最后，本文采用回归模型常用的评估指标R2，MAE，MSE和RMSE对模型进行了合理性评估，并对其缺点部分提出改进，同时提出了模型优化的构想。

**关键词** 皮尔森系数 互信息 信息增益 CART决策树 随机森林

目录

[一、问题重述 - 3 -](#_Toc514708163)

[1.1 问题背景 - 3 -](#_Toc514708164)

[1.2 需要解决的问题 - 3 -](#_Toc514708165)

[二、问题分析 - 3 -](#_Toc514708166)

[2.1 对问题一的分析 - 3 -](#_Toc514708167)

[2.2 对问题二的分析 - 4 -](#_Toc514708168)

[2.3 对问题三的分析 - 4 -](#_Toc514708169)

[三、模型假设 - 4 -](#_Toc514708170)

[四、符号说明 - 5 -](#_Toc514708171)

[五、模型的建立与求解 - 5 -](#_Toc514708172)

[5.1 数据预处理 - 5 -](#_Toc514708173)

[5.1.1 数据集成 - 5 -](#_Toc514708174)

[5.1.2 缺失值处理 - 6 -](#_Toc514708175)

[5.1.3 缺失记录处理 - 6 -](#_Toc514708176)

[5.1.4 连续属性离散化 - 7 -](#_Toc514708177)

[5.1.5 特征提取及数据集描述 - 8 -](#_Toc514708178)

[5.2 模型评价 - 8 -](#_Toc514708179)

[5.2.1 评价指标 - 8 -](#_Toc514708180)

[5.2.2 评价方式——交叉验证法 - 9 -](#_Toc514708181)

[5.3 问题一相关性因子的选取 - 9 -](#_Toc514708182)

[5.3.1 相关概念 - 9 -](#_Toc514708183)

[5.3.2 基于皮尔森系数的相关因子分析 - 9 -](#_Toc514708184)

[5.3.3 基于信息增益的相关因子分析 - 10 -](#_Toc514708185)

[5.3.4 基于互信息的相关因子分析 - 13 -](#_Toc514708186)

[5.3.5 相关因子选择模型的检验与评价 - 16 -](#_Toc514708187)

[5.4 问题二相邻两日空气污染物浓度差预报模型 - 17 -](#_Toc514708188)

[5.4.1 相关概念 - 17 -](#_Toc514708189)

[5.4.2 基于CART决策树+随机森林构建预报模型 - 18 -](#_Toc514708190)

[5.4.3 预报模型结果及分析 - 20 -](#_Toc514708191)

[5.5 问题三不同季节污染物模型构建 - 21 -](#_Toc514708192)

[5.5.1 按季节预报模型描述与构建 - 21 -](#_Toc514708193)

[5.5.2 预报模型结果及分析 - 23 -](#_Toc514708194)

[六、模型的评价与优化 - 23 -](#_Toc514708195)

[6.1 模型优点 - 23 -](#_Toc514708196)

[6.2 模型缺点 - 24 -](#_Toc514708197)

[6.3 模型优化 - 24 -](#_Toc514708198)

[七、参考文献 - 24 -](#_Toc514708199)

[八、附录 - 26 -](#_Toc514708200)

[附表1——信息增益值 - 26 -](#_Toc514708201)

[附表2——互信息系数值 - 29 -](#_Toc514708202)

# 一、问题重述

## 1.1 问题背景

近年来，随着我国工业化和城市化进程的蓬勃发展，大气污染日趋严重。一方面传统的SO2、NO2、PM10污染尚未解决，O3和PM2.5为主的二次污染问题亦日益突出，呈现区域性大气复合型污染状况；另一方面大气环境质量的变化与气象条件改变之间的相互作用越来越复杂，污染防控的压力越来越大。区域性大范围的重污染天气制约着社会经济的可持续发展，也极大威胁人类的生态健康和生态环境的安全。污染防控工作也成为我国政府关注的热点重点问题之一，已成为我国政府“十三五”规划重点防控工作内容并上升到国家战略高度。最新统计资料显示：京津冀为近3年来国内重污染区域之一，既有区域影响因素，也有自身城市的发展，人口增多、车辆增加，污染物排放总量不断增加等不利因素的叠加。京津冀地区的经济快速发展，使其环境问题日趋严重，空气污染问题尤为突出。河北省中南部也是全国空气污染最为严重的地区之一，其空气污染特征具有一定的代表性。

目前，国内学者针对气象因素对大气污染物的影响开展了大量研究，并取得很多有意义的结果。研究表明：气象要素与污染物的聚集、传输、扩散、干湿沉降等密切相关[1, 2]，在污染源相对稳定的条件下，气象条件成为影响城市空气污染的主导因素。因此，研究气象条件与大气污染物的相关关系显得至关重要，也成为有效开展污染防控工作的关键环节。

## 1.2 需要解决的问题

我国工业化和城市化进程的发展使得大气污染问题日趋严重，研究气象条件与污染物的关系成为污染防控的关键问题。通过初步分析，合理选择城市，建立数学模型，查找相关数据，解决以下问题：

1. 请用适当的方法，挑选出与污染物相关性较好的气象因子，由于各个城市各季影响因子不同，故不同城市不同季节不同污染物的气象因子也不同。（以石家庄和邢台为例分析，数据见附件1）
2. 采用适当的方法构建空气污染物相邻两日浓度差值的预报模型，对模型的构建过程进行详细阐述，选取空气污染物对预报模型进行设计。（合理选择城市分部情况，附件2给出其中部分城市的参考数据）
3. 利用适当的方法，分不同季节建立的污染物的数学模型，采用不同的指标对预报结果进行评价分析？

# 二、问题分析

## 2.1 对问题一的分析

对于问题一，选取与污染物浓度相关性较好的气象因子。首先进行数据预处理，按污染物浓度采集时间为主键集成气象数据，静稳指数以及其他因子数据；通过计算欧氏距离对缺失值插补；采用分段均值填充缺失记录，采用C4.5决策树算法中采用的二分法对连续属性进行离散化处理。然后采用2014-2017的数据进行问题分析。影响因子包含气象数据（每日的气温四次平均值、最高气温、最低气温、降水量、气压四次平均值、相对湿度四次平均值、最小相对湿度、十分钟风速四次平均值、日最大风风速、日照时数）、混合层高度、地表通风系数以及静稳指数（当日四次静稳指数平均值），相关污染物包括PM2.5、PM10、SO2、NO2、CO、O3。

由于不同地域不同季节污染物浓度与气象因子的相关关系不同，故按照不同城市不同季节分别分析以获取其相关性较好的气象因子。利用python将训练集分为八组，分别为：石家庄春，石家庄夏，石家庄秋，石家庄冬，邢台春，邢台夏，邢台秋，邢台冬；即前四组为石家庄地区四个季节分别对应的气象数据，后四组为邢台地区四个季节分别对应的气象数据；按照污染物的不同，分别进行相应地处理。

最后采用皮尔森系数来验证与已有相关研究的一致性，并确定污染物与气象因子之间是否是线性相关。除此之外，与其他方法比较以确定污染物与气象因子之间是否具有线性相关性。同时结合信息增益（IG）和互信息对训练数据进行特征选择，分析确定属性贡献度以获取与污染物相关性较好的气象因子。

## 2.2 对问题二的分析

在数据预处理完成的数据集DS基础上，通过分析数据集以及题目内容，创新性地增加三个属性值，分别表示：1）温度持续上升；2）相邻两日湿度维持在50%以上；3）相邻两日风速差值。并据此构建数据集DS+，然后构建两个模型：浓度差值预报模型和浓度预报模型，后者根据预测值再计算差值。通过对比BaseLine实验结果选择较优模型进行优化。然后使用数据预处理后的新增属性的数据集，构建CART（Classification and Regression Tree）决策树及随机森林（Random Forest）模型进行预报。最后通过GridSearch自动搜寻模型最优参数。

## 2.3 对问题三的分析

分不同季节建立污染物模型。首先通过进行数据统计分析不同城市各污染物浓度在四个季度上平均值的变化趋势，然后建立两个城市各自的BaseLine模型以及两个城市共同的BaseLine模型，根据模型预报结果选择较优方式构建CART和RF模型，同样使用GridSearch自动搜索最优参数。最后采用回归模型常用的评估指标R2，平均绝对误差MAE（mean absolute error），均方误差MSE（Mean Squared Error）和均方根误差RMSE（Root Mean Squard Error）进行模型评估。

# 三、模型假设

由于构建的预报模型是基于附件中的数据进行构建的，为了评估我们的模型，做出以下假设：

1. 假设影响浓度的因子只有数据表中列出的气象数据、混合层高度、地表通风系数以及静稳指数；
2. 假设表中提供的相关数据是在无人为干预的情况下采集得到的；
3. 假设表中提供的相关数据是真实准确的；
4. 假设2014-2017这几年期间没有自然灾害事件等不可抗力影响数据稳定性；
5. 假设2014-2017这几年期间没有重大污染事件、政府出台环保政策或人工降雨等人为因素影响数据；
6. 假设插补的缺失值接近于实际值。

# 四、符号说明

本文使用了一些统一的符号，其符号说明如下表4-1所示：

表4-1 符号说明表

|  |  |
| --- | --- |
| 符号 | 说明 |
|  | 属性集合 |
|  | 属性集中的第个属性 |
|  | 20-20降水量缺失值及记录集合 |
|  | 第条20-20降水量缺失值的记录 |
|  | 无缺失值记录的第条记录的第个属性值 |
|  | 第条缺失值记录的第个属性值 |
|  | 无缺失值记录的第条记录与第条缺失值记录的距离 |
|  | 无缺失值记录总数 |
|  | 有缺失值记录总数 |
|  | 样本集 |
|  | 连续属性 |
|  | 当前样本集中第类样本所占的比例 |
|  | 第个分支结点中中所有在属性上取值为的样本 |
|  | 采用二分法连续属性离散化的划分点集合 |

# 五、模型的建立与求解

## 5.1 数据预处理

由于附件二中7个地市的数据不完整（只有气象数据，没有污染物浓度数据），故无法对模型训练结果进行测试评估。所以采用附件一中2014-2017的数据作为全部数据，使用随机打乱拆分训练集和测试集，保证训练集和测试集服从相同的数据分布，利于模型学习泛化能力。通过分析数据，需要对数据的缺失值进行处理，对数据的缺失记录进行处理，另外还需要对连续属性进行离散化和特征缩放。

### 5.1.1 数据集成

由于影响污染物浓度的影响因子不在一个表中，所以需要对所有的影响因子进行集成处理融合在一个数据表中，即污染物浓度、气象因子、混合层高度、通风系数和静稳指数。采用python对数据进行集成处理。以污染物浓度采集时间为主键对各个表中的数据进行集成。并将集成处理后的数据集命名为DS。通过集成处理后得到的属性描述如表5-1所示：

表5-1 属性集合各属性与气象因子对应关系表

|  |  |
| --- | --- |
| 属性 | 属性名称 |
| *a*1 | 通风系数（Coefficient\_of\_ventilatin） |
| *a*2 | 混合层高度（Mixing\_layer\_height） |
| *a*3 | 相对湿度4次平均（Average\_relative\_humidity） |
| *a*4 | 最高气温（Maximum\_temperature） |
| *a*5 | 日最大风风速（Maximum\_wind\_spee） |
| *a*6 | 最小相对湿度（Minimum\_relative\_humidity） |
| *a*7 | 最低气温（Minimum\_temperature） |
| *a*8 | 本站气压4次平均（Average\_station\_pressure） |
| *a*9 | 静稳指数（Stable\_weather\_index(SWI)） |
| *a*10 | 日照时数合计（Sunshine\_hour） |
| *a*11 | 气温4次平均（Average\_temperatur） |
| *a*12 | 20-20降水量（Precipitation） |
| *a*13 | 10分钟风速4次平均（Average\_wind\_speed） |

### 5.1.2 缺失值处理

通过分析表中的数据发现2014-2016气象数据表中20-20降水量和10分钟风速4次平均中有缺失值。查阅相关资料发现，降水量与气温，湿度等是有关系的。故对于20-20降水量的缺失值采用回归方法进行插值。根据已有的20-20降水量值和与其相关的气温4次平均、最高气温、最低气温、相对湿度4次平均、最小相对湿度和日照时数合计属性值建立拟合模型来预测缺失的20-20降水量属性值。假定（，等于属性的个数，为属性集合）为影响20-20降水量属性值的第个因子, 为无缺失值记录总数。20-20降水量缺失值的记录（，等于缺失值的记录总数，为20-20降水量缺失值集合），通过计算缺失值记录与已有记录之间的欧式距离（为无缺失值记录的第条记录的第个属性值，为第条缺失值记录的第个属性值，有缺失值记录总数为）找到与其距离最小的记录所对应的属性值进行插补。其计算公式为：

 （1）

通过计算各无缺失值记录与缺失值记录之间的距离找到与该有缺失值记录距离最小的记录所对应的属性值进行插补。

对于10分钟风速4次平均值的缺失值也采用同样的方法进行插补。

### 5.1.3 缺失记录处理

在数据集成过程中发现有某些数据缺失，即当天只有污染物浓度数据记录而没有气象因子等数据，也就是说出现了缺失记录的问题。2014年7月3日-2014年7月9日静稳指数是缺失的。面对这个问题，我们的解决办法是利用2014年7月2日和2014年7月10静稳指数进行增补。每日的静稳指数为其当日静稳指数记录值的平均值，而且静稳指数的变化趋势是连续的，则缺失的这几日的数据值按首项是2日的静稳指数值，末项为10日的静稳指数值得等差数列进行插补。

### 5.1.4 连续属性离散化

由于连续属性的可取值数目不再是有限的，因此，不能直接根据连续属性的可取值来对节点进行划分。此时，需要对连续属性进行离散化处理以便进行信息增益及决策树构建。我们采用C4.5决策树算法中采用的二分法对连续属性进行处理。

在这里通过计算信息熵和信息增益来确定，信息熵的计算公式为：

 （2）

其中，是当前样本集中第类样本所占的比例。

假定离散属性有个可能的取值，若使用来对样本集进行划分。则会产生个分支结点，其中第个分支结点包含了中所有在属性上取值为的样本，记为。考虑到不同分支结点包含的样本数不同，分支节点赋予权重，即样本数越多的分支结点影响越大，则用属性对样本集进行划分获得的信息增益为：

 （3）

信息增益越大，则意味着使用属性来进行划分所获得的“纯度提升”越大。

给定样本集和连续属性，假定在上出现了个不同的取值，将这些值从小到大进行排序，记为。基于划分点*t*可将分为子集和，其中包含那些属性上取值不大于*t*的样本。显然，对相邻的属性取值与来说，*t*在区间中取任意值所产生的划分结果相同。因此，对连续属性，我们可考察n-1个元素的候选划分点集合

 （4）

即把区间的中位点作为候选划分点。然后，我们就可离散属性值一样来考虑这些划分点，选取最优的划分点进行样本集合的划分。即对公式（2）稍加改造：

 （5）

其中是样本集基于划分点二分后的信息增益，于是，我们就可选择使最大化的划分点。

### 5.1.5 特征提取及数据集描述

考虑到杜传耀等研究结论：1）温度持续上升，相对湿度维持在50%以上，风速基本在2 m•s-1以下；2）逆温层的持续存在，大气对流减弱，阻止了颗粒物向高空扩散，并且数据中提供了温度，湿度以及风速维度信息，因此，我们在数据集成处理得到的数据集（DS）基础上选择增加三个维度来表示温度的持续上升，湿度的持续维持50%以上，以及风速的差值。第一个维度，当日温度与昨日温度都大于设置的阈值，且当日温度大于昨日温度则置为1，否则为0。第二个维度，当日湿度与昨日湿度都大于设置的50%则置为1，否则为0。第三个维度为当日平均风速与上日的平均风速差值，两日最大的风速差值小于10。

得到新的数据集记为DS+，其具体的描述如下表5-2所示：

表5-2 数据集说明

|  |  |
| --- | --- |
| 数据集 | 属性说明 |
| DS | 数据集成处理之后的所有属性（即原始属性，13个） |
| DS+ | 数据集成基础上新增三个属性（即原始属性+3个） |

通过在上述两个数据集上的模型训练，可以进一步评估影响污染物浓度的因子，同时也可以形成对比实验以评估模型。

## 5.2 模型评价

### 5.2.1 评价指标

我们采用采用回归模型常用的评估指标R2，MAE，MSE和RMSE进行模型评估。

**R2** 模型预测的所有误差与平均值之间的关系，最大值为1，越大表示模型能力越好，其计算公式为：

 （5）

**MAE** 用真实值-预测值，然后取绝对值之后求和取得的平均值，其计算公式为：

 （6）

**MSE** 用真实值-预测值，然后平方之后求和取得的平均值，其计算公式为：

 （7）

**RMSE** 它是MSE的平均值，其计算公式为：

 （8）

### 5.2.2 评价方式——交叉验证法

我们采用交叉验证法对该模型进行评估。“折交叉法”先将数据集划分为个互斥子集，即。每个子集都尽可能保持数据分布的一致性，即从中通过分层采样得到。然后，每次用个子集的并集作为训练集，余下的那个子集作为测试集；这样就可获得组训练/测试集，从而进行次训练和测试，最终返回得是这个测试结果的均值。显然，交叉验证评估结果的稳定性和保真性在很大程度上取决于的取值，在这里，我们取的值为10，即进行10折交叉验证。

## 5.3 问题一相关性因子的选取

### 5.3.1 相关概念

**皮尔森系数** 皮尔森相关系数[3]用来衡量线性关联性的程度,两个连续变量的Pearson相关性系数等于它们之间的协方差除以它们各自标准差的乘积。系数的取值总是在-1.0到1.0之间，接近0的变量被成为无相关性，接近1或者-1被称为具有强相关性。

**信息增益** 信息增益[4]在决策树算法中是用来选择特征的指标，信息增益越大，则这个特征的选择性越好。在信息增益中，重要性的衡量标准就是看特征能够为分类系统带来多少信息，带来的信息越多，该特征越重要。

**互信息** 互信息[5]是信息论里一种有用的信息度量，它可以看成是一个随机变量中包含的关于另一个随机变量的信息量，或者说是一个随机变量由于已知另一个随机变量而减少的不确定性，因此常用来表示两个变量X与Y是否有关系，以及关系的强弱。

### 5.3.2 基于皮尔森系数的相关因子分析

为了挑选出不同城市不同季节与不同污染物相关性较好的气象因子，本文使用三种特征选择方法结合的方式对题目一进行求解。

首先，使用了皮尔森相关系数法，皮尔森系数也称皮尔森积距相关系数，是一种线性相关系数，是最常用的一种相关系数，记为r,用来反映两个变量X和Y的线性相关程度，r的取值范围为（-1，1），r的绝对值越大表明相关性越强，当r>0时表明两个变量正相关，即一个变量值越大则另一个变量值也会越大；当r<0时表明两个变量负相关，即一个变量越大则另一个变量反而越小；当r=0时，表明两个变量是非线性相关的；当r=1或r=-1时，表明两个变量可以很好的由直线方程来描述，所有样本点都很好的落在一条直线上。总体相关系数p定义为两个变量X、Y之间的协方差和标准差的比值，如下：

 （9）

估算样本的协方差和标准差，可得到样本的皮尔森相关系数：

 （10）

*r*还可以由样本点的标准分数均值估计得到与上式等价的表达式：

 （11）

其中，为样本的标准分数；为样本均值；为样本标准差；n为样本数量。

对于虽然皮尔森相关系数的绝对值越大表示相关性越高，但是皮尔森系数有一个明显的缺点，即它接近于1的程度与数据组数n相关，这容易给人一种假象。因为，当n较小时，相关系数的波动较大，对有些样本相关系数的绝对值易接近于1；当n较大时，相关系数的绝对值容易偏小。特别是当n=2时，相关系数的绝对值总为1。因此在样本容量n较小时，仅凭相关系数较大就判定变量x与y之间有密切的线性关系是不妥当的，因此本文只用该方法来验证题目已知的相关内容(风速的增加，污染物浓度呈下降趋势；污染的严重程度主要取决于风速大小和强风持续时间；连续污染与稳定的垂直层结及地面较弱的气压场有关等)以及污染物浓度与气象因子之间的非线性关系。石家庄和邢台污染程度与温度、湿度、风速的皮尔森系数如表5-3所示：

表5-3 两个城市的皮尔森系数分析

|  |  |  |  |
| --- | --- | --- | --- |
|  | 温度 | 湿度 | 风速 |
| 石家庄 | -0.42630914 | 0.26078098 | -0.40308647 |
| 邢台 | -0.41082751 | 0.18659954 | -0.32469533 |

从表中数据可以看出，温度与污染程度呈负相关关系，结合实际情况分析考虑，冬季温度低，烧煤产生大量污染物，造成污染加重，而夏季气温高，会降低烧煤量，并且会有大量植物生长吸收净化空气中的部分污染物使得污染物浓度降低。湿度与污染程度呈现正相关关系，风速与污染程度呈现负相关关系，这与已有研究得出的结论（郑美琴[6]对日照市的空气污染与气象要素进行分析，指出随风速的增加，污染物浓度呈下降趋势。李国翠[7]对石家庄大气污染与沙尘天气的关系进行分析，得出污染的严重程度主要取决于风速大小和强风持续时间）是一致的。

### 5.3.3 基于信息增益的相关因子分析

选择2014年到2017年的数据，根据其气象因子及对应的污染物浓度数据，通过计算其“信息熵”和信息增益值度量与空气污染相关性较好的气象因子。假定当前样本集合*D*中第*k*类样本所占的比例为*pk（1，2，…，|Y|）* ,则*D*的信息熵计算公式同公式（2）的值越小，则*D*的纯度越高。

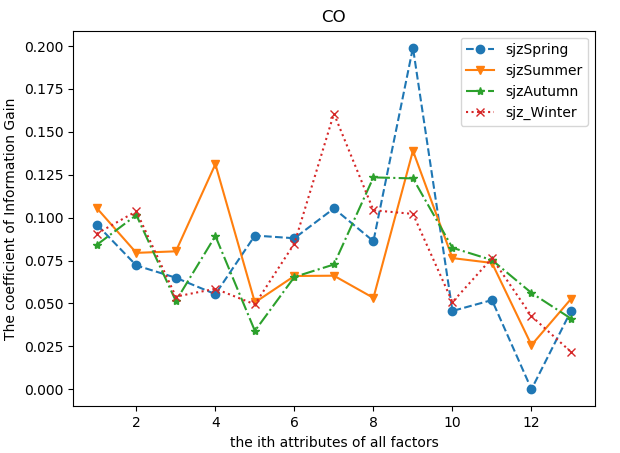
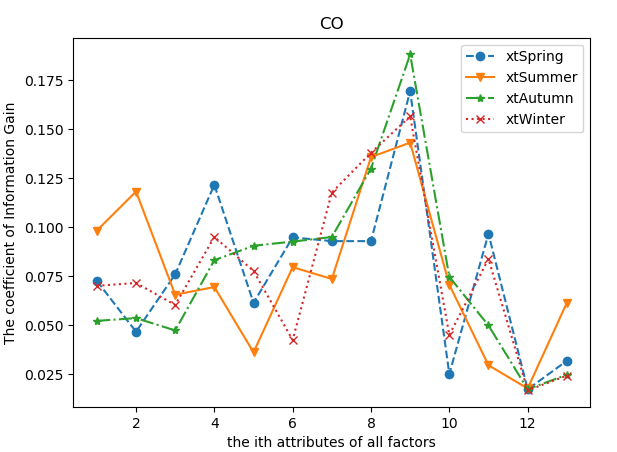
当一个特征不能变化时，系统的信息量是多少？这个信息量其实也有专门的名称，就叫做“条件熵”。但是如果一个特征*X*，它可能的取值有*n*种*（x1，x2，……，xn）*，当计算条件熵而需要把它固定的时候，每一种可能都要固定一下，计算*n*个值，然后取均值才是条件熵。而取均值也不是简单的加一加然后除以*n*，而是要用每个值出现的概率来算平均（简单理解，就是一个值出现的可能性比较大，固定在它上面时算出来的信息量占的比重就要多一些），即通过下式计算：

 （12）

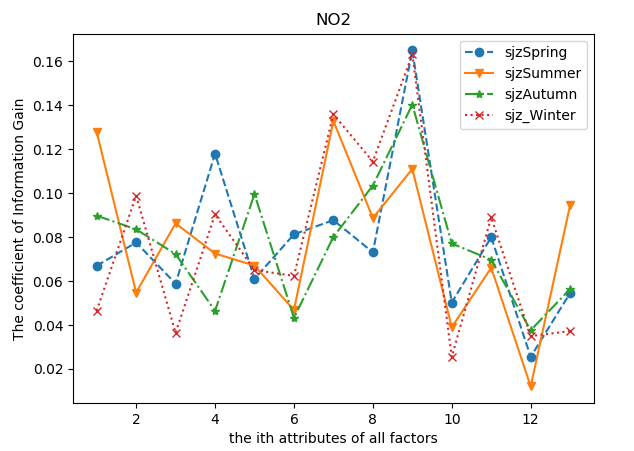
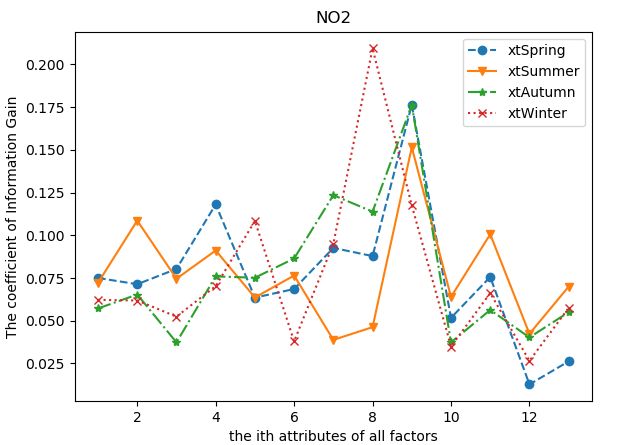
因此，特征X给系统带来的信息增益就可以写成系统原本的熵与固定特征X后的条件熵之差:

 （13）

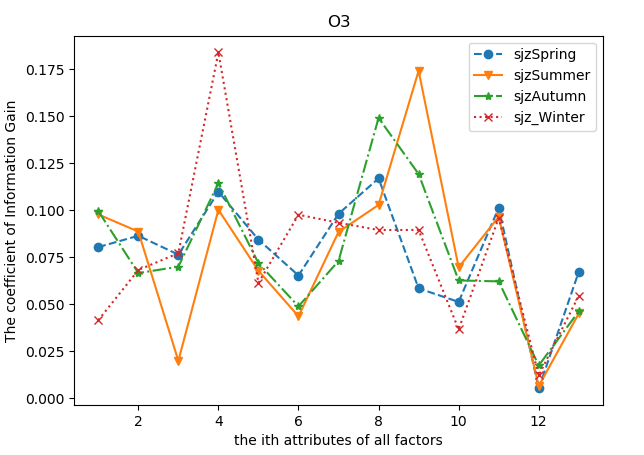
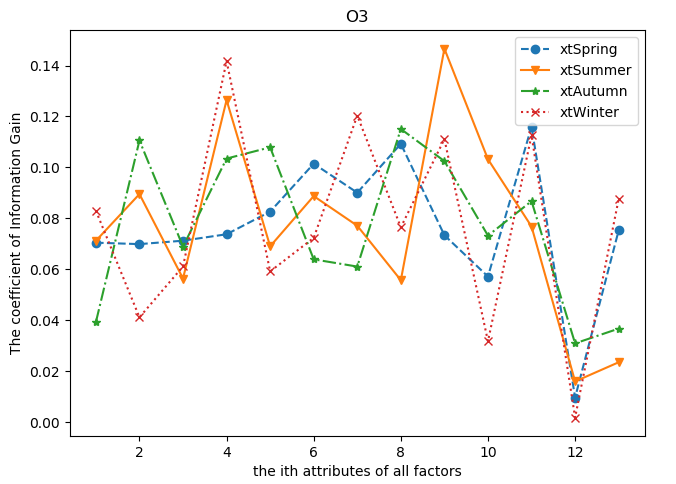
假定离散属性*a*有*V*个可能的取值{a1,a2,…,aV},若使用*a*来对样本集*D*进行划分。则会产生*V*个分支结点，其中第*v*个分支结点包含了*D*中所有在属性*a*上取值为*av*的样本，记为*Dv*。考虑到不同分支结点包含的样本数不同，分支节点赋予权重*|Dv|/|D|*，即样本数越多的分支结点影响越大，则用属性*a*对样本集*D*进行划分获得的信息增益计算公式同公式（3）信息增益越大，则意味着使用属性a来进行划分所获得的“纯度提升”越大。因此可以用信息增益来进行决策树的划分属性选择。由于不同城市不同季节污染物浓度与各气象因子的相关性是有所区别的。故按照不同城市，不同季节将数据集分为八组：石家庄春（sjzSpring）、石家庄夏（sjzSummer）、石家庄秋（sjzAutumn）、石家庄冬（sjzWinter）、邢台春（ztSpring）、邢台夏（xtSummer）、邢台秋（xtAutumn）和邢台冬（xtWinter），污染物有六种类型：CO、NO2、O3、PM2.5、PM10和SO2，而影响污染物浓度的属性因子总共有13个，按照表5-2中的属性因子分别训练得到两个城市四个季节六种污染物的信息增益值，其计算结果见附表1，具体的信息如图5-1所示，其中，左图为石家庄四个季节六种污染物的信息增益值，右图为邢台市的训练结果：

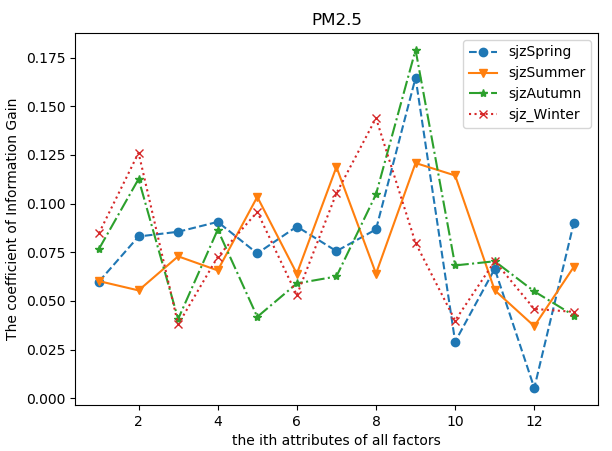
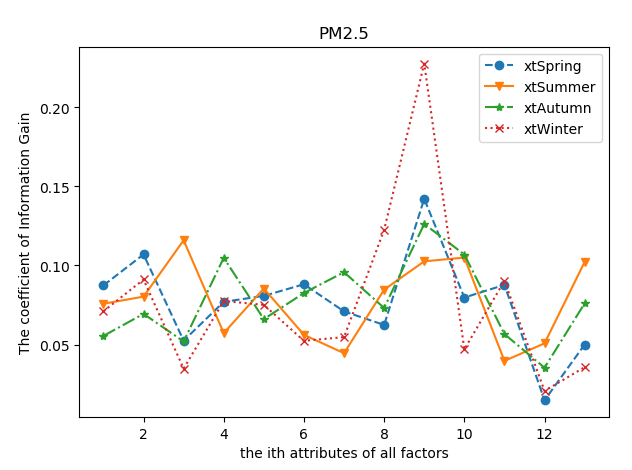
*a*（1）石家庄四季CO与特征关系 *b*（1）邢台四季CO与特征关系

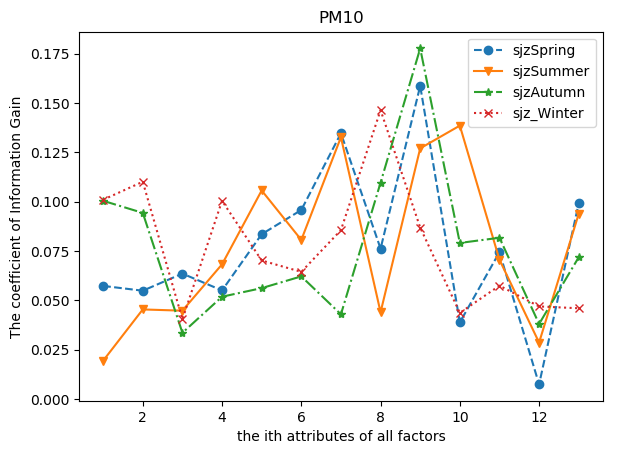
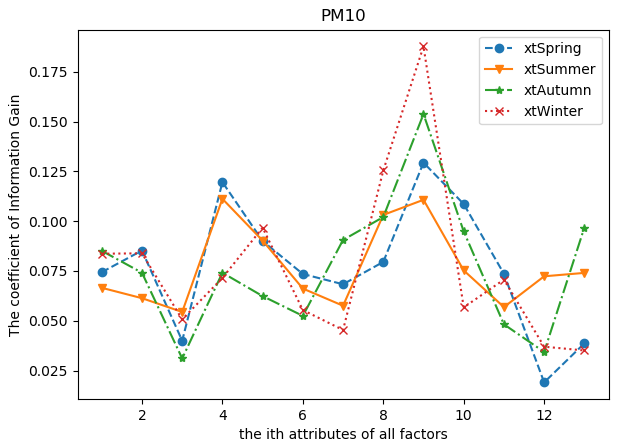
*a*（2）石家庄四季NO2与特征关系 *b*（2）邢台四季NO2与特征关系

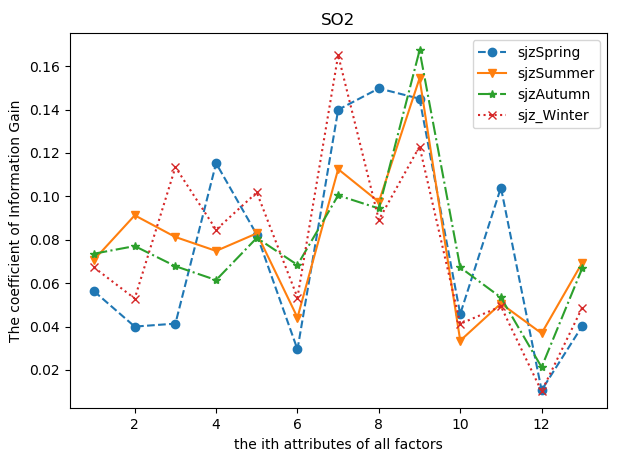
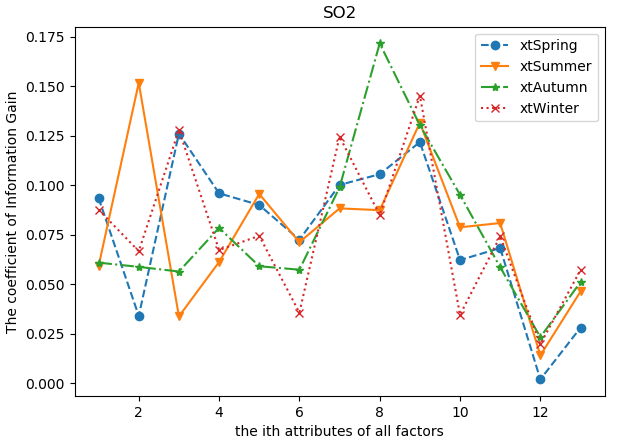
*a*（3）石家庄四季O3与特征关系 *b*（3）邢台四季O3与特征关系

*a*（4）石家庄四季PM2.5与特征关系 *b*（4）邢台四季PM2.5与特征关系

*a*（5）石家庄四季PM10与特征关系 *b*（5）邢台四季PM10与特征关系

*a*（6）石家庄四季SO2与特征关系 *b*（6）邢台四季SO2与特征关系

图5-1 基于信息增益的特征选择结果，*a*为石家庄特征选择结果图

*b*为邢台市特征选择结果

分析图中结果，按照不同城市不同季节各污染物与各因子之间的关系，石家庄市影响CO浓度最大的因子是静稳指数，其次是最高气温、最低气温与通风系数；影响NO2浓度最大的因子是静稳指数，其次是最低气温、通风系数与最高气温；影响O3浓度最大的因子是最高气温、静稳指数，其次是本站气压4次平均、气温4次平均； 影响PM2.5浓度最大的因子是静稳指数，其次是本站气压4次平均与混合层高度；影响PM10浓度最大的因子是静稳指数，其次为本站气压4次平均、日照时数合计与最低气温；影响SO2浓度最大的因子是最低气温与静稳指数，其次为本站气压4次平均、最高气温与相对湿度4次平均。邢台市影响CO浓度最大的因子是静稳指数，其次是本站气压4次平均、最高气温与混合层高度；影响NO2浓度最大的因子是本站气压4次平均与静稳指数，其次是最低气温与最高气温；影响O3浓度最大的因子是静稳指数与最高气温，其次是最低气温、气温4次平均与本站气压4次平均；影响PM2.5浓度最大的因子是静稳指数，其次是混合层高度、相对湿度4次平均与本站气压4次平均；影响PM10浓度最大的因子为静稳指数，其次是最高气温、本站气压4次平均和10分钟风速4次平均；影响SO2浓度最大的因子是本站气压4次平均，其次是混合层高度、静稳指数、相对湿度4次平均与最低气温。总体而言，各种污染物浓度与静稳指数的相关性都较强，而与20-20降水量的相关性都较强都较小。

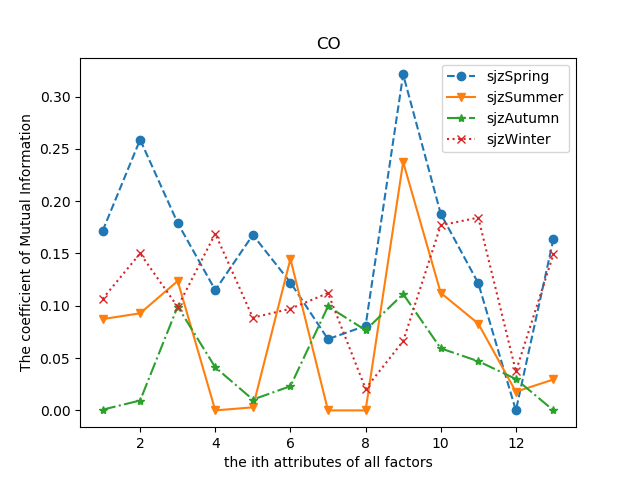
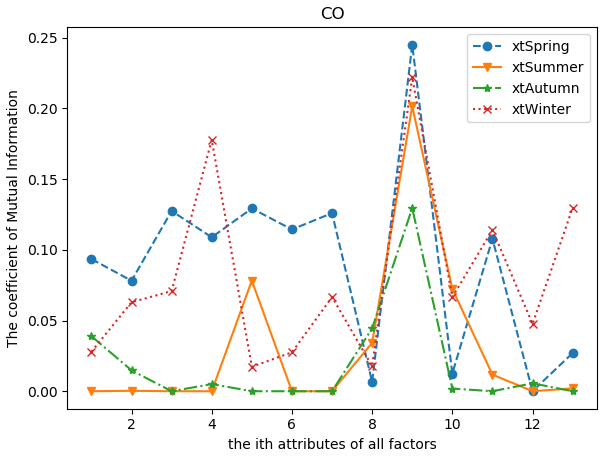
### 5.3.4 基于互信息的相关因子分析

有关系的信息才能消除不确定性，这个有关系有点模糊，最好能度量“相关性”才好。香农在信息论中提出了一个“互信息”的概念作为两个随机事件“相关性”的度量。假定有两个随机事件X和Y，它们的互信息的计算公式推导为：

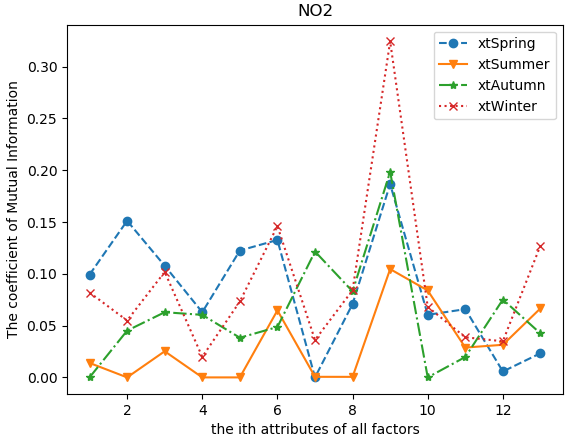
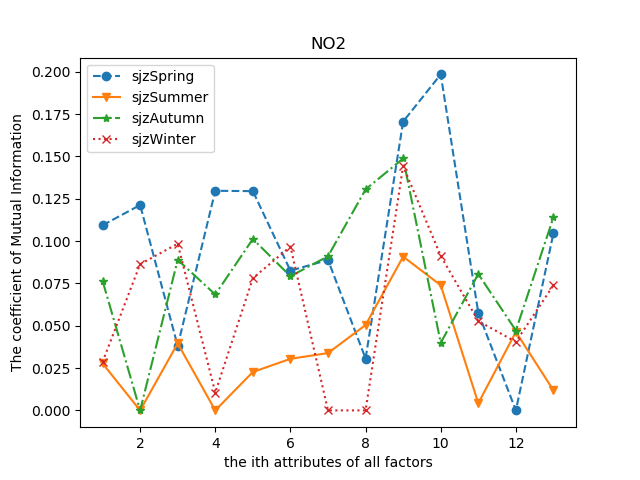
 （14）

其中，是*Y*的熵*，*定义为，*Y*的熵衡量的是Y的不确定度，*Y*分布的越离散的值越高，则表示在已知*X*的情况下*Y*的不确定度；因此*，*根据互信息公式可以看出，可以解释为由*X*引入而使Y的不确定性减小的的量，这个减小的量为，所以如果*X、Y*关系越密切，就越大；最大的取值是，此时的值为0，意义为*X*和*Y*完全相关，在X确定的情况下Y是一个确定的值，没有出现其他不确定情况的概率，所以的值为0；如果，此时，意义为*X*的出现不影响*Y*，就为0，即代表*X*与*Y*不相关。

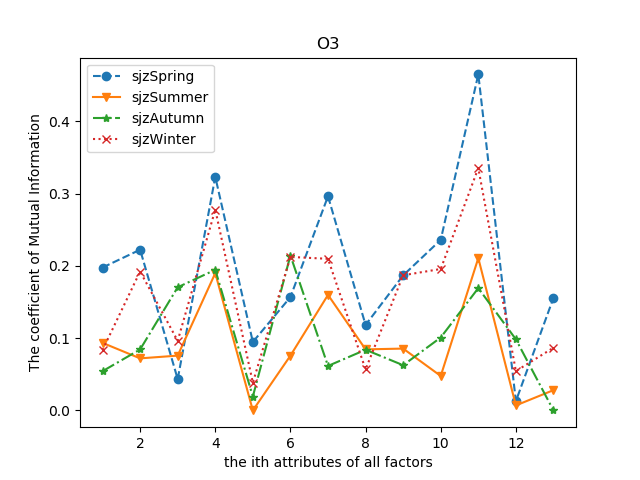
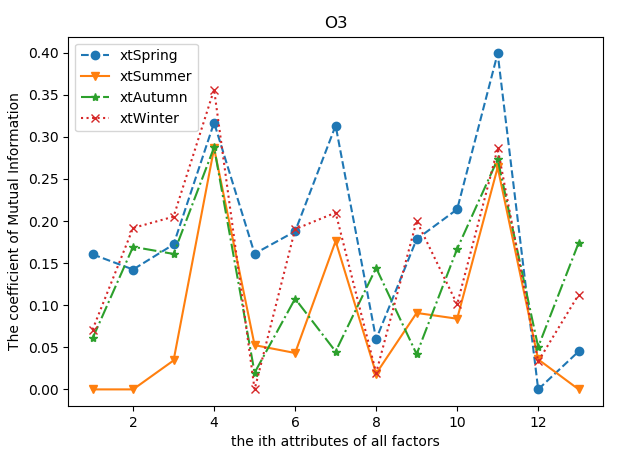
由于不同城市不同季节污染物浓度与各气象因子的相关性是有所区别的。故按照不同城市，不同季节将数据集分为八组：石家庄春（sjzSpring）、石家庄夏（sjzSummer）、石家庄秋（sjzAutumn）、石家庄冬（sjzWinter）、邢台春（ztSpring）、邢台夏（xtSummer）、邢台秋（xtAutumn）和邢台冬（xtWinter），又污染物有六种类型：CO、NO2、O3、PM2.5、PM10和SO2，而影响污染物浓度的属性因子总共有13个，按照表5-2中的属性因子分别训练得到两个城市四个季节六种污染物的互信息系数值，计算得到的互信息系数表见附表2，根据该表如图5-1所示的两个城市四个季节各个气象因子的互信息系数值，其中，左图为石家庄四个季节六种污染物的互信息系数，右图为邢台市的训练结果：

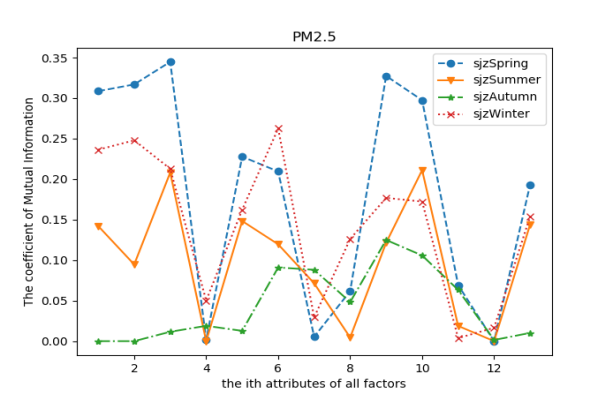
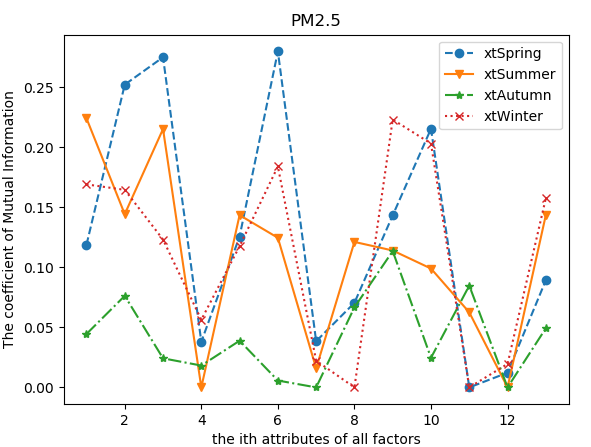
*a*（1）石家庄四季CO与特征关系 *b*（1）邢台四季CO与特征关系



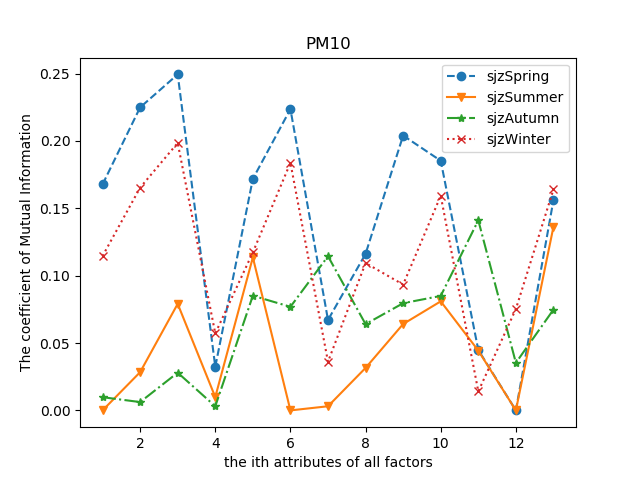
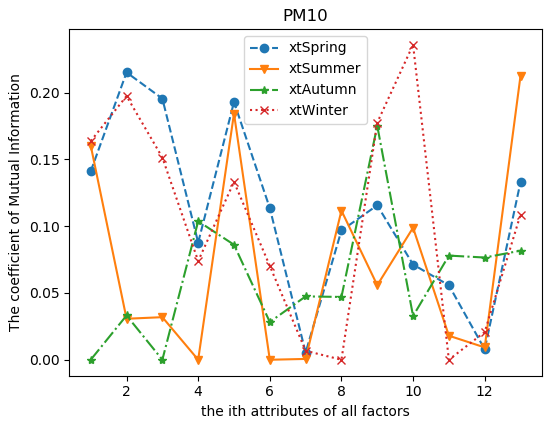
*a*（2）石家庄四季NO2与特征关系 *b*（2）邢台四季NO2与特征关系

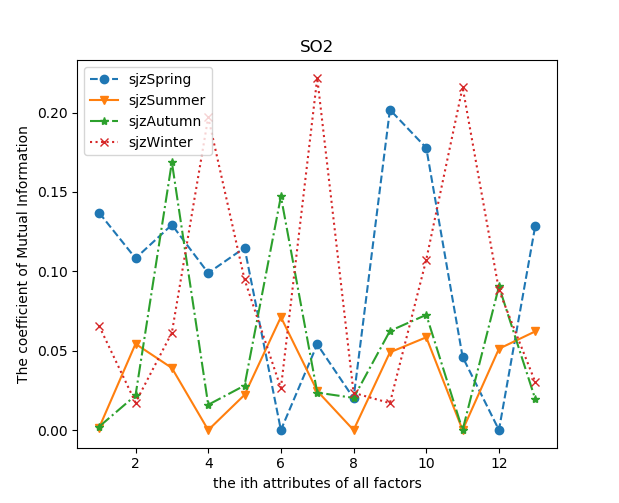
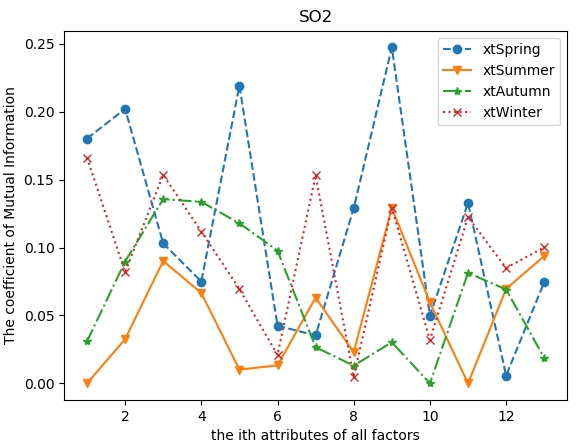
*a*（3）石家庄四季O3与特征关系 *b*（3）邢台四季O3与特征关系

*a*（4）石家庄四季PM2.5与特征关系 *b*（4）邢台四季PM2.5与特征关系

*a*（5）石家庄四季PM10与特征关系 *b*（5）邢台四季PM10与特征关系

*a*（6）石家庄四季SO2与特征关系 *b*（6）邢台四季SO2与特征关系

图5-2 基于互信息的特征选择结果, *a*为石家庄特征选择结果图

*b*为邢台市特征选择结果

分析图中结果发现，按照不同城市不同季节各污染物与各因子之间的关系，石家庄市影响CO浓度最大的因子是静稳指数，其次是混合层高度、气温4次平均和10分钟风速；影响NO2浓度最大的因子是日照时数合计，其次是静稳指数和日最大风风速；影响O3浓度最大的因子是气温4次平均，其次是最高气温、最低气温和日照时数合计，即其与温度的关系最大；影响PM2.5浓度最大的因子是日照时数合计，其次是最小相对湿度、相对湿度4次平均和静稳指数；影响PM10浓度最大的因子与影响PM2.5的因子大致相同；影响SO2浓度最大的因子是最低气温和气温4次平均，其次是静稳指数、最高气温与相对湿度4次平均。邢台市影响CO浓度最大的因子是静稳指数，其次为最高气温、日最大风速、10分钟风速4次平均；影响NO2浓度最大的因子是静稳指数，其次是混合层高度、最小相对湿度、10分钟风速4次平均；影响O3浓度最大的因子是气温4次平均，其次是最高气温、最低气温和静稳指数，即其与温度的关系最大；影响PM2.5浓度最大的因子是相对湿度4次平均和最小相对湿度，其次是通风系数、静稳指数和日照时数合计；影响PM10浓度最大的因子是日照时数合计和10分钟风速4次平均，其次是混合层高度、相对湿度4次平均、日最大风速、静稳指数；影响SO2浓度最大的因子是静稳指数，其次是混合层高度、日最大风速、相对湿度4次平均和最低气温。总体而言，冬季和春季污染程度与各因子之间的相关性更大，而夏季和秋季污染程度与各因子之间的关系较小。

### 5.3.5 相关因子选择模型的检验与评价

根据附表1、附表2、表5-1、图5-1及图5-2，即综合皮尔森系数、信息增益和互信息得到两个城市四个季节与污染物相关性最好的气象因子如表5-4所示：

表5-4 两个城市四个季节与污染物相关性最好的气象因子

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 污染物  季节 | CO | NO2 | O3 | PM2.5 | PM10 | SO2 |
| 石家庄春 | *a*9 | *a*10 *a*9 | *a*11 *a*4 *a*7 | *a*3 *a*9 *a*2 | *a*2 *a*3 *a*6 | *a*9 *a*10 |
| 石家庄夏 | *a*9 | *a*9 *a*10 | *a*11 *a*4 *a*7 | *a*10*a*3 | *a*13 *a*5 | *a*6 *a*13 |
| 石家庄秋 | *a*9 | *a*9 *a*8 | *a*6 *a*4 *a*11 | *a*9 *a*10 | *a*11 *a*7 | *a*3 *a*6 |
| 石家庄冬 | *a*11 *a*7 | *a*9 *a*3 *a6* | *a*11 *a*4 | *a*6 *a*2 *a*1 | *a*3 *a*6 | *a*7 *a*11 |
| 邢台春 | *a*9 | *a*9 *a*2 *a6* | *a*11 *a*4 *a*7 | *a*6 *a*3 *a*2 | *a*2 *a*3 *a5* | *a*9 *a*5 *a*2 |
| 邢台夏 | *a*9 | *a*9 | *a*4 *a*11 | *a*1 *a*3 | *a*13 *a*5 *a*1 | *a*9 *a*13 |
| 邢台秋 | *a*9 | *a*9 *a*7 | *a*4 *a*11 | *a*9 | *a*9 *a*4 | *a*3 *a*4 *a5* |
| 邢台冬 | *a*9 | *a*9 *a*3 *a*13 | *a*4 *a*11 | *a*9 *a*10 | *a*10 *a*2 | *a*1 *a*3 *a*7 |

通过对测试数据的分析结果可以看出，本文的特征选择模型对测试输出的回归是可以接受的，说明本文的预测基本合理。

从表5-4中可以看出，石家庄春、夏、秋季与CO浓度最相关的因子是静稳指数，石家庄冬季与CO浓度最相关的因子是气温4次平均；而邢台市一年四季与CO浓度最相关的因子是静稳指数，与NO2浓度最相关的也是静稳指数；石家庄市春季与NO2浓度最相关的因子是日照实数合计和静稳指数，石家庄夏、秋、冬季与NO2浓度最相关的因子是静稳指数；对于O3的浓度，石家庄市春、夏、冬季最相关的因子是气温4次平均，秋季是最小相对湿度，邢台市春季最相关的是气温4次平均，邢台夏、秋、冬季最相关的是最高气温；对于PM2.5，石家庄春、夏、秋、冬季最相关的分别是相对湿度4次平均，日照时数合计，静稳指数和最小相对湿度，邢台春、夏、秋、冬季最相关的分别是最小相对湿度、通风系数、静稳指数、静稳指数；对PM10，石家庄春、夏、秋、冬，邢台春、夏、秋、冬分别是混合层高度、10分钟风速4次平均、气温4次平均、相对湿度4次平均、混合层高度、10分钟风速4次平均、静稳指数和日照时数合计；而SO2，石家庄春、邢台春、夏最相关的是静稳指数，石家庄夏、秋、冬最相关的是最小相对湿度、相对湿度4次平均、最低气温，邢台秋、冬最相关的是相对湿度4次平均和通风系数。

## 5.4 问题二相邻两日空气污染物浓度差预报模型

### 5.4.1 相关概念

**CART决策树** CART是一种高效的决策树，CART决策树可以用于分类也可以用于回归，当CART决策树用于回归树时，整个数为一棵二叉树，也就是说对于每一个非叶节点，都有一个划分属性和一个划分的值，根据这个值将当前的数据集合划分为两类。相比于ID3决策算法与C4.5决策树算法来说，三者的实现过程在结构上是一样的，但是ID3决策树算法和C4.5决策树算法需要人为对数据进行分段，而CART决策树的最佳特征度量采用Gini Gain,因此calcShannonEnt方法被替换成calcGini方法，自动对数据进行分段，提高了工作效率。而且CART采用二分法，对于有多个取值的离散特征，需要首先获取最小二分序列及其GiniGain，因此splitDataSet方法需按照取值tuple分开、chooseBestFetureToSplit要返回最佳分叉点及其二分序列。当数据拥有多个特征时，构建全局模型就很有难度而且笨拙了。实际上，现实生活中的大量问题包括气象因子与污染物浓度的关系的问题都是非线性的，不可以再使用全局线性模型来回归任何数据。在这种情况下，CART回归决策树的方法就相当有用了，但是在CART回归决策树中叶节点的数据必须是连续的而不能是离散的，于是要对CART进行一些修改再进行数据回归。

**随机森林** 随机森林[8]就是用随机的方式建立一个森林，森林里面有很多的决策树组成，随机森林的每一棵决策树之间是没有关联的。在得到森林之后，当有一个新的输入样本进入的时候，就让森林中的每一棵决策树分别进行一下判断，看看这个样本应该属于哪一类（对于分类算法），然后看看哪一类被选择最多，就预测这个样本为那一类。在建立每一棵决策树的过程中，有两点需要注意：采样与完全分裂。首先是两个随机采样的过程，random forest对输入的数据要进行行、列的采样。对于行采样，采用有放回的方式，也就是在采样得到的样本集合中，可能有重复的样本。假设输入样本为N个，那么采样的样本也为N个。这样使得在训练的时候，每一棵树的输入样本都不是全部的样本，使得相对不容易出现over-fitting。然后进行列采样，从M 个feature中，选择m个(m << M)。之后就是对采样之后的数据使用完全分裂的方式建立出决策树，这样决策树的某一个叶子节点要么是无法继续分裂的，要么里面的所有样本的都是指向的同一 个分类。一般很多的决策树算法都一个重要的步骤 - 剪枝，但是这里不这样干，由于之前的两个随机采样的过程保证了随机性，所以就算不剪枝，也不会出现over-fitting。随机森林的优点：1）在数据集上表现良好；2）在当前的很多数据集上，相对其他算法有着很大的优势；3）它能够处理很高维度（feature很多）的数据，并且不用做特征选择；4）在训练完后，它能够给出哪些feature比较重要；5）在创建随机森林的时候，对generlization error使用的是无偏估计；6）训练速度快；7）在训练过程中，能够检测到feature间的互相影响；8）容易做成并行化方法；9）实现比较简单。

### 5.4.2 基于CART决策树+随机森林构建预报模型

随机森林主要应用于回归和分类。本文主要探讨基于随机森林的回归问题,对于回归树来说预测值为叶节点目标变量的加权均值。在决策树的根部，所有的样本都在这里，此时树还没有生长，这棵树的残差平方和就是回归的残差平方和。然后选择一个变量也就是一个特征（气象因子），这个变量使得通过这个进行分类后的两部分的分别的残差平方和的和最小。然后在分叉的两个节点处，再利用这样的准则，选择之后的分类属性。一直这样下去，直到生成一颗完整的树。以决策树为基本模型的bagging在每次bootstrap放回抽样之后，产生一棵决策树，抽多少样本就生成多少棵树，在生成这些树的时候没有进行更多的干预。而随机森林也是进行bootstrap抽样，但它与bagging的区别是：在生成每棵树的时候，每个节点变量都仅仅在随机选出的少数变量中产生。因此，不但样本是随机的，连每个节点变量的产生都是随机的。随机森林（random forest）是一种利用多个分类树对数据进行判别与分类的方法，它在对数据进行回归的同时，还可以给出各个变量（特征）的重要性评分，评估各个变量在回归中所起的作用。随机森林中的每一棵回归树为二叉树，其生成遵循自顶向下的递归分裂原则，即从根节点开始依次对训练集进行划分；在二叉树中，根节点包含全部训练数据， 按照节点纯度最小原则，分裂为左节点和右节点，它们分别包含训练数据的一个子集，按照同样的规则节点继续分裂，直到满足分支停止规则而停止生长。若节点n上的数据全部来自于同一类别，则此节点的纯度I(n)=0，纯度度量方法是Gini准则，即假设 是节点n上属于 类样本个数占训练数据总数的比例。

随机性主要体现在两个方面：1）训练每棵树时，从全部训练样本中选取一个子集进行训练（即bootstrap取样）。用剩余的数据进行评测，评估其误差；2）在每个节点，随机选取所有特征的一个子集，用来计算最佳分割方式。

回归树生成具体步骤是：

第一步：搜索分裂变量和分裂点。随机森林是有多颗CART树组成的，使用最小二乘法生成回归树，在训练数据集所在的输入空间中，递归地将每个区域划分为两个子区域并决定每个子区域并决定每个子区域上的输出值，假设将空间划分为M个区域，每个区域用对响应建模。在二叉划分中， 遍历变量，对固定的切分变量扫描切分点，选择使上式达到最小值的对。用选定的对划分区域并决定相应的输出值（假设搜索分裂变量*j*和分裂点*s*，定义一对半平面）：



 （15）

搜索分裂变量*j*和分裂点*s*的目标函数为：

 （16）

内部极小化可以用下式求解：



 （17）

第二步：树结构的控制。涉及两个方面，一个是何时停止分裂，另一个是对树进行剪枝。

何时停止分裂有两种方法：一种是仅当分裂是平方和的降低超过某个阈值时，才分裂；另一种是仅当达到最小节点大小时停止分裂。对树进行剪枝：思路是定义树的一些子树，从它们中找到在“对数据拟合程度 + 树模型的复杂度”准则下最优的一个，如下式：

 （18）

其中，，（ ，），参数来控制树的大小和对数据拟合程度之间的折中，对它的估计用5或10折交叉验证实现。

第三步：继续对两个子区域调用以上两个步骤，直至满足停止条件，然后将输入空间划分为M个区域，生成决策树：

 （19）

随机森林生成具体步骤是：

1）输入为训练数据集以及弱学习器迭代次数T。输出为最终的强分类器；

2）对于t=1,2...,T:对训练集进行第t次随机采样，共采集m次，得到包含m个样本的采样集;用采样集训练第m个决策树模型，在训练决策树模型的节点的时候， 在节点上所有的样本特征中选择一部分样本特征， 在这些随机选择的部分样本特征中选择一个最优的特征来做决策树的左右子树划分。

3）如果是分类算法预测，则T个弱学习器投出最多票数的类别或者类别之一为最终类别。如果是回归算法，T个弱学习器得到的回归结果进行算术平均得到的值为最终的模型输出。

在随机森林中某个特征*X*的重要性的计算方法如下：

1）对于随机森林中的每一颗决策树，使用相应的*OOB*(袋外)数据计算它的袋外数据误差，记为。

2）随机地对袋外数据所有样本的特征X加入噪声干扰（就可以随机的改变样本在特征X处的值），再次计算它的袋外数据误差，记为。

3）假设随机森林中有N颗树，那么对于特征X的重要性计算公式为：

 （20）

之所以可以用这个表达式来作为相应特征的重要性的度量值是因为：若给某个特征随机加入噪声之后，袋外的准确率大幅度下降，则说明这个特征对于样本的分类结果影响很大，也就是说它的重要程度比较高。

在本文中，首先随机森林RF使用了CART决策树作为弱学习器，其次，在使用决策树的基础上，RF对决策树的建立做了改进，对于普通的决策树，我们会在节点上所有的n个样本特征中选择一个最优的特征来做决策树的左右子树划分，但是RF通过随机选择节点上的一部分样本特征，这个数字小于n，假设为，然后在这些随机选择的个样本特征中，选择一个最优的特征来做决策树的左右子树划分。这样进一步增强了模型的泛化能力。

### 5.4.3 预报模型结果及分析

对于问题二，待求模型期望输出为相邻两日浓度差值，经过初步考虑有以下两种方式：

**方式1** 先建立当日污染物浓度预报模型，通过预报模型得到当日污染物浓度数据，再计算与前一日污染物的浓度差值，从而得到污染物相邻两日浓度差值的预报模型。

**方式2** 直接建立相邻两日浓度差值预报模型，将每日污染物浓度减去上一日浓度得到浓度差值数据，用浓度差值数据进行模型训练以得到污染物相邻两日浓度差值的预报模型。

对通过以上两种方式建立的BaseLine模型进行效果对比，本文选择方式1建立的模型作为最终的污染物相邻两日浓度差值预报模型并对其进行优化和调参。对于数据集，本文在对数据进行预处理之后，进一步使用DS+数据集进行模型的训练。同时，在DS+数据集的基础上提取出石家庄和邢台数据，分别建立当日浓度预报模型。模型的算法决策树及随机森林的实现使用python，并通过Sci-kit Learn包提供的GridSearch方法进行模型最优参数的搜寻。待优化参数包含生成决策树数量n\_estimators（范围为50-550）；最大特征计算方式max\_features(包括‘auto’,’sqrt’，‘log2’三种方式)；最大树深max\_depth（范围为5-20）。同时，将石家庄及邢台的DS+数据集进行训练集和测试集的切分（比例为0.8：0.2）。GridSearch使用10折交叉验证方式进行参数搜索。

模型结果分析：通过GridSearch搜寻当日污染物浓度模型的最好的参数为：max\_features=’auto’、 n\_estimators=50、max\_depth=17。图5-3，图5-4分别是石家庄和邢台当日污染物浓度预报模型在测试集上与真实值的对比图。

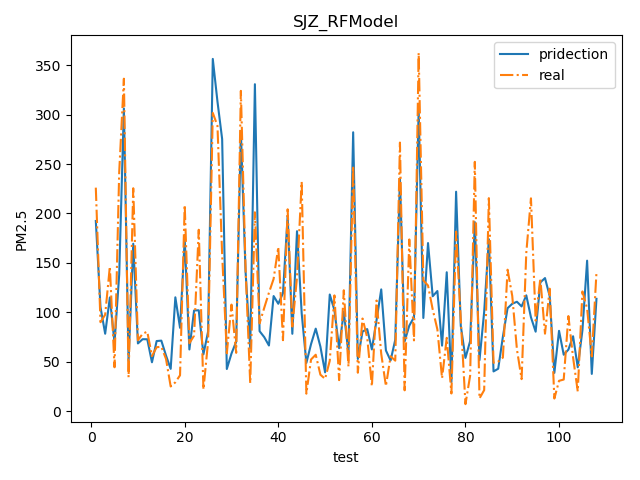


图5-3 石家庄当日污染物浓度模型在测试集上与真实值对比图

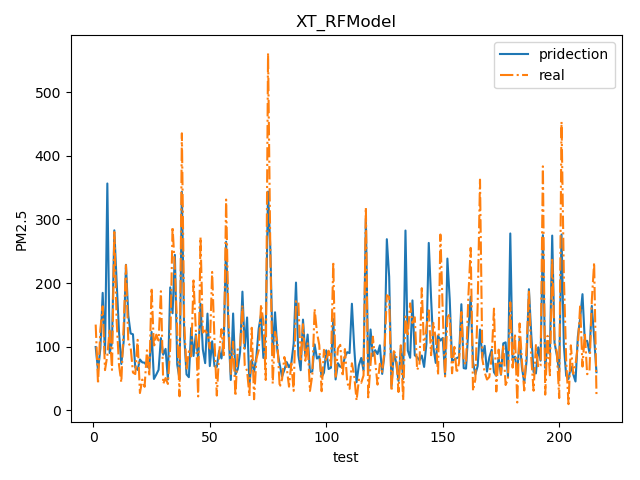


图5-4 邢台市当日污染物浓度模型在测试集上与真实值对比图

通过模型指标可以看出，石家庄当日污染物浓度预报结果与真实值比较接近，真实值与预测值的平均误差用RMSE表示。

表5-5 石家庄当日污染物浓度模型评估

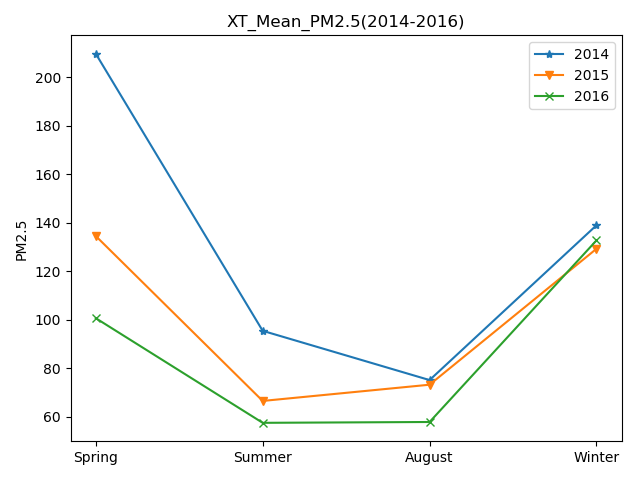
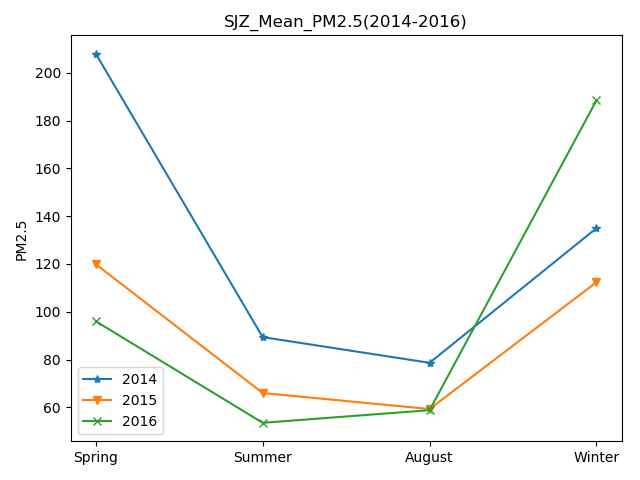
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 指标 | R2\_train | R2\_test | MAE | MSE | RMSE |
| 石家庄 | 0.84 | 0.69 | 35 | 1946 | 44 |

通过对比使用石家庄和邢台的数据共同训练与将两城市数据分开分别训练的实验结果分析发现，前者的预测效果不如后者。此外，分析发现邢台的当日污染物浓度预报模型效果与之相似,真实值与预测值的平均误差在56左右。通过对比图可以发现，当PM2.5值较大的情况下，模型的预测效果较差。对比模型的预测值和真实值轨迹，模型表现比较稳定。

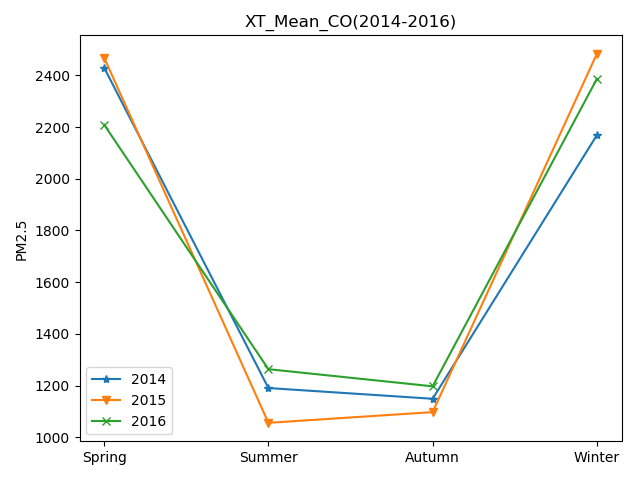
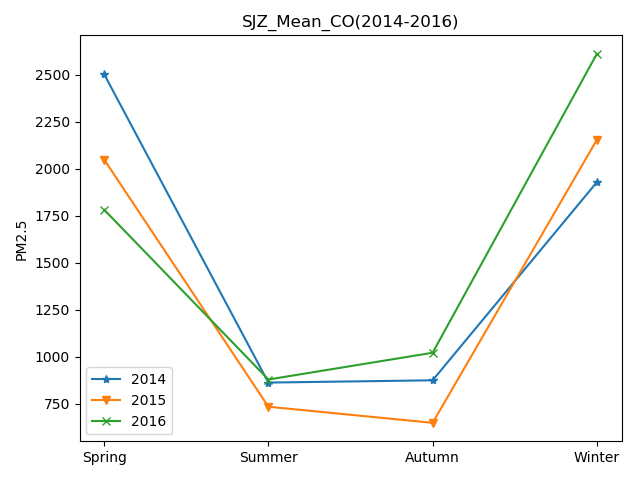
## 5.5 问题三不同季节污染物模型构建

### 5.5.1 按季节预报模型描述与构建

问题三旨在解决不同季节的污染物预报模型，首先，由于四个季度的数据量基本一致（在520-670之间，相差不大）；而且实验发现各个污染物在四个季度上平均浓度的变化趋势表现一致(呈现春冬高，夏秋低的趋势)，于是我们首先对数据仅以季度为指标进行实验。本文以两城市的PM2.5和CO两种污染物为例进行实验：



（1）石家庄PM2.5四季平均浓度变化 （2）邢台PM2.5四季平均浓度变化

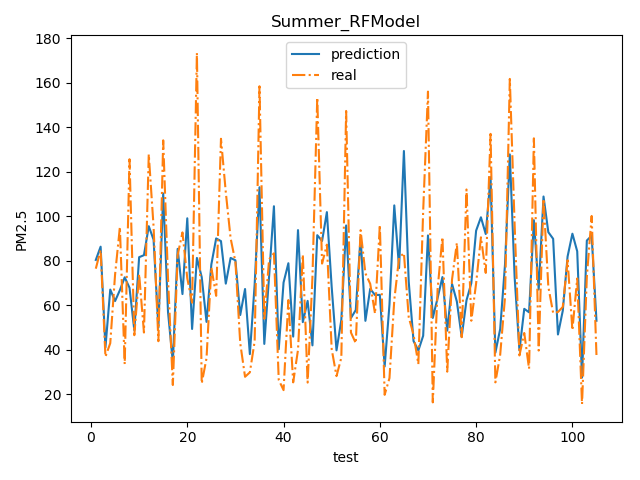
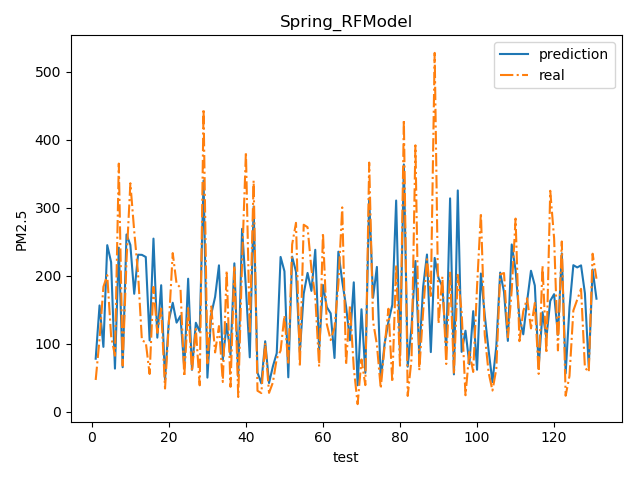


（3）石家庄CO四季平均浓度变化 （4）邢台CO四季平均浓度变化

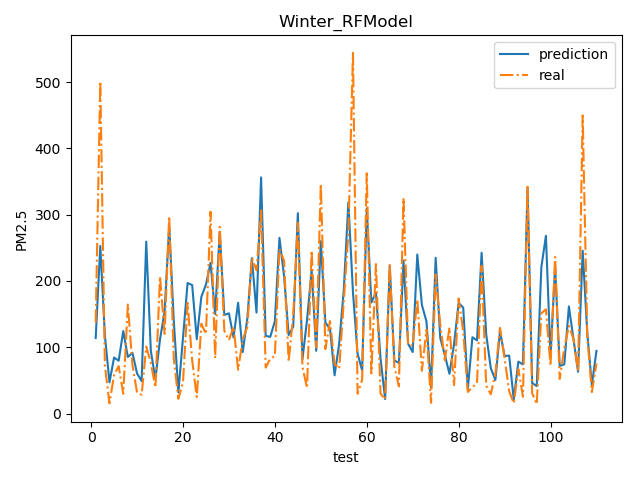
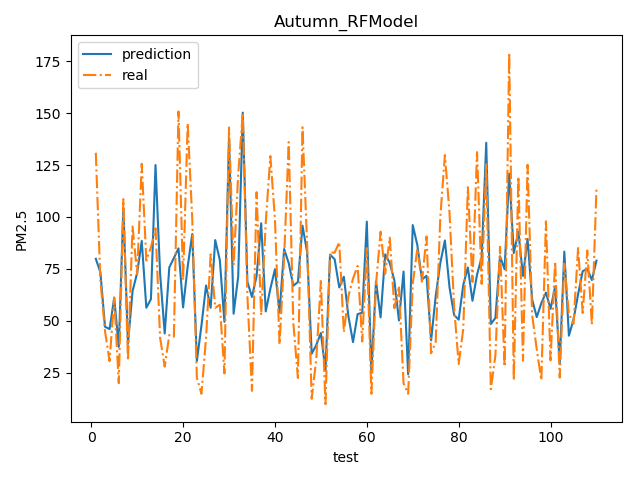
图5-4 两城市的PM2.5和CO四个季度上平均浓度的变化趋势

通过上图可以明确的看到春季与冬季的污染物浓度大幅度高于夏季和秋季，这也符合Pearson系数的分析，温度与污染物呈现负相关性质。（以石家庄的PM2.5为例，春冬两季的PM2.5浓度均值比夏秋两季高80，方差高60）。

我们同样使用CART和随机森林结合方法分别构建四个季度的污染物预报模型。由于题目三强调分季节建立模型，并且两个城市在四个季节上的表现行为一致，于是本文使用石家庄和邢台两城市的数据共同建立四个季节的模型。使用两城市在同一季节的数据，首先对比了两个城市分别建立各自的BaseLine模型以及两个城市共同建立一个BaseLine模型的预测效果发现，后者的预测效果更加理想，所以我们选择每个季度都使用两个城市的DS+混合数据集进行训练。同样使用GridSearch自动搜索最优参数。以污染物PM2.5为例，四个季节的预报模型最终效果如图5-5：



（1）春节PM2.5浓度预报与真实值 （2）夏节PM2.5浓度预报与真实值



（3）秋节PM2.5浓度预报与真实值 （4）冬节PM2.5浓度预报与真实值

图5-5 四个季节的PM2.5浓度预报与真实值

### 5.5.2 预报模型结果及分析

四个季节的PM2.5预报模型评估指标如下表所示。夏季和秋季的模型虽然R2比较小，但误差(RMSE)值是小于春冬季节模型的。这和上述分析一致，春冬季度数据均值高，方差大，模型的训练需要更强的学习能力。通过模型的真实值和预测值可以发现，模型在PM2.5值过大的情况下，预测效果很差。此外，模型的预测值的轨迹相对于真实值更加稳定。

表5-6 石家庄四季PM2.5预报模型评估

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 季节 | R2\_train | R2\_test | MAE | MSE | RMSE |
| Spring | 0.94 | 0.15 | 54 | 5027 | 70 |
| Summer | 0.92 | 0.42 | 21 | 767 | 27 |
| Autumn | 0.91 | 0.43 | 23 | 827 | 28 |
| Winter | 0.94 | 0.6 | 43 | 4496 | 67 |

# 六、模型的评价与优化

## 6.1 模型优点

本文构建的模型有以下几个创新性的优点：

1）在数据预处理阶段：根据已有研究的结论（重污染天气过程能够持续的气象条件：温度持续上升，相对湿度维持在50%以上，风速基本在2 m·s-1以下）创新性地新增3个维度的属性来表示相邻温度持续上升、两日湿度是否维持、相邻风速的差值，并通过实验对比发现，新添加的维度对模型有一定的提升效果。

2）在解决问题一的过程中，提出首先使用皮尔森系数来验证相关气象因子与污染物浓度的是否存在线性相关性，以及确定各气象因子与污染物浓度的正负相关性，同时，采用互信息和信息增益分别选择与污染物浓度相关性较好的气象因子，结合以上三种方法，根据不同城市不同季节分别得到与污染物浓度相关性较好的气象因子；

3）在解决问题二的过程中，采用基于CART决策树和随机森林的算法构建空气污染物相邻两日浓度差值的预报模型，使用GridSearch自动搜索模型最优参数，并构建不同属性维度的数据集（DS，DS+）形成对比实验以验证方法的性能，同时据此发现影响预测性能的关键因子以提升模型性能；

4）在解决问题三的过程中，经过比较详细数据统计分析，最终选择根据不同季节不区分城市构建污染物浓度模型，与问题二同样，采用基于CART决策树和随机森林的算法进行模型构建并自动搜索最优参数。

5）问题二、三模型均采用回归模型常用的评估指标R2，MAE，MSE和RMSE进行模型评估，评估效果全面且可靠。

## 6.2 模型缺点

在模型构建过程中通过分析数据，处理数据，并利用数据进行模型构建发现以下几个缺点以待进一步改进：

1）数据量小，特征提取有待进一步优化，模型训练结果有待进一步提升；训练集的大小会对模型的效果有影响，以问题二石家庄模型为例，训练集的比例为0.7，0.8，0.9情况下，具体情况如表6-1所示。训练的模型在测试集上表现越来越好。

表6-1 不同训练集比例的测试效果

|  |  |  |  |
| --- | --- | --- | --- |
| 训练集比例 | 0.7 | 0.8 | 0.9 |
| 测试集R2 | 0.41 | 0.47 | 0.55 |

2）用随机森林训练的模型，在测试集的R2与测试集的R2值相差有些大，有一定的过拟合问题，需要进一步的调整。

## 6.3 模型优化

目前模型存在一定程度的过拟合问题，需要进一步调参优化。鉴于DS+数据集的表现良好，但提升限度有限(仅将RMSE降低10%)，特征提取及特征的选择需要进一步的研究与使用，还可以在时间及算力充足的情况下，添加自动搜索参数的数量和范围。数据集少的问题，可以选择获取完整的2017年数据和2018年数据加入训练。

其次，目前使用的是进行回归预报模型的建立，以后可以考虑将问题转换成分类问题来解决。使用分类问题有两个优势：

1. 机器学习和深度学习中有丰富的分类算法及相关经验。
2. 传统的模型是人工抽取数据特征，而深度学习中的神经网络模型能自动提取特征并进行模型学习。

进一步的，问题二强调的相邻两日浓度差值，可以理解为前一条数据对后一条数据有影响。这类问题使用序列模型来处理效果会更好，典型的代表如深度学习中的RNN(Recurrent Neural Network)。以后有望使用它来进行模型预测。

# 七、参考文献

[1] 王欣睿, 叶剑军, 倪志鑫, 钟煜宏. 珠江口海域大气中重金属季节变化特征及其与气象因子的关系, 海洋通报, 1(6): 632-640, 2016.

[2] 张健, 樊曙先, 孙玉, 张悦, 魏锦成. 厦门春季PM10中PAHs成分谱特征及其与气象要素相关性分析, 环境科学, 15(4): 1173-1181, 2015.

[3] Yashuang Mu, Xiaodong Liu, Lidong Wang. A Pearson’s correlation coefficient based decision tree and its parallel implementation, Information Sciences, 435(1): 40-58, 2018.

[4] Lingyun Gao, Mingquan Ye, Xiaojie Lu, DaobinHuang. Hybrid Method Based on Information Gain and Support Vector Machine for Gene Selection in Cancer Classification, Genomics, Proteomics & Bioinformatics, 15(6): 389-395, 2017.

[5] WanfuGao, LiangHu, PingZhang. Class-specific mutual information variation for feature selection, Pattern Recognition, 79(1): 328-339, 2018.

[6] 郑美琴, 卢振礼, 日照市区PM10污染物特征及其与气象要素的关系, 南京气象学院学报, 29(3): 413—417, 2006.

[7] 李国翠, 王建国, 连志鸾. 石家庄市大气污染与沙尘天气的关系分析, 气象与环境学报, 23(2): l-5, 2007.

[8] Joaquín Abellán, Carlos J.Mantas, Javier G.Castellano. A Random Forest approach using imprecise probabilities, Knowledge-Based Systems, 134(1): 72-84, 2017.

[9] 王占山, 张大伟, 李云婷, 董欣, 孙瑞雯, 孙乃迪. 北京市夏季不同O3和PM2.5污染状况研究, 环境科学, 16(3): 807-815, 2016.

# 八、附录

## 附表1——信息增益值

附表1 两个城市不同季节污染物浓度与气象因子信息增益值

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | CO | NO2 | O3 | PM2.5 | PM10 | SO2 |
| 石家庄春 | 通风系数 | 0.092961 | 0.067905 | 0.088278 | 0.062264 | 0.056496 | 0.062529 |
| 混合层高度 | 0.078727 | 0.075858 | 0.082226 | 0.080067 | 0.064656 | 0.03804 |
| 相对湿度4次平均 | 0.066771 | 0.059438 | 0.07327 | 0.081543 | 0.066886 | 0.04747 |
| 最高气温 | 0.053489 | 0.120345 | 0.10916 | 0.089869 | 0.063606 | 0.113385 |
| 日最大风速 | 0.08895 | 0.059583 | 0.081753 | 0.074631 | 0.082449 | 0.077236 |
| 最小相对湿度 | 0.084681 | 0.081102 | 0.069504 | 0.091578 | 0.079887 | 0.032826 |
| 最低气温 | 0.101252 | 0.090369 | 0.098722 | 0.075806 | 0.136224 | 0.143646 |
| 本站气压4次平均 | 0.087242 | 0.073369 | 0.116846 | 0.086375 | 0.076135 | 0.142106 |
| 静稳指数 | 0.204386 | 0.161513 | 0.057319 | 0.165479 | 0.15636 | 0.138465 |
| 日照时数合计 | 0.045147 | 0.053542 | 0.047062 | 0.031447 | 0.038839 | 0.045727 |
| 气温4次平均 | 0.050342 | 0.07537 | 0.101976 | 0.06468 | 0.071032 | 0.110582 |
| 20-20降水量 | 0.000307 | 0.025533 | 0.005114 | 0.005105 | 0.00694 | 0.01063 |
| 10分钟风速4次平均 | 0.045745 | 0.056073 | 0.068772 | 0.091156 | 0.10049 | 0.037357 |
| 石家庄夏 | 通风系数 | 0.111286 | 0.137893 | 0.088871 | 0.059648 | 0.024151 | 0.071103 |
| 混合层高度 | 0.077327 | 0.054083 | 0.08709 | 0.050406 | 0.045916 | 0.091052 |
| 相对湿度4次平均 | 0.079775 | 0.088294 | 0.018728 | 0.071526 | 0.048042 | 0.081925 |
| 最高气温 | 0.130736 | 0.070239 | 0.095814 | 0.067679 | 0.066058 | 0.074591 |
| 日最大风速 | 0.046514 | 0.067227 | 0.068766 | 0.103837 | 0.105695 | 0.081647 |
| 最小相对湿度 | 0.063659 | 0.039338 | 0.042921 | 0.070207 | 0.081549 | 0.042975 |
| 最低气温 | 0.062618 | 0.13161 | 0.090323 | 0.117421 | 0.134048 | 0.113055 |
| 本站气压4次平均 | 0.054311 | 0.089539 | 0.104471 | 0.066807 | 0.041512 | 0.096809 |
| 静稳指数 | 0.139548 | 0.108812 | 0.174482 | 0.121088 | 0.127166 | 0.152964 |
| 日照时数合计 | 0.076534 | 0.039134 | 0.066977 | 0.114695 | 0.139172 | 0.035417 |
| 气温4次平均 | 0.075626 | 0.068267 | 0.102508 | 0.052605 | 0.072521 | 0.053361 |
| 20-20降水量 | 0.026065 | 0.01257 | 0.009277 | 0.035797 | 0.025744 | 0.037971 |
| 10分钟风速4次平均 | 0.055999 | 0.092996 | 0.049773 | 0.068283 | 0.088427 | 0.067131 |
| 石家庄秋 | 通风系数 | 0.088123 | 0.090825 | 0.098039 | 0.09119 | 0.100632 | 0.070627 |
| 混合层高度 | 0.090507 | 0.082556 | 0.070768 | 0.102016 | 0.093422 | 0.074679 |
| 相对湿度4次平均 | 0.041613 | 0.071895 | 0.06781 | 0.042668 | 0.038442 | 0.066904 |
| 最高气温 | 0.089488 | 0.044318 | 0.111417 | 0.08416 | 0.050847 | 0.063203 |
| 日最大风速 | 0.034568 | 0.095488 | 0.073291 | 0.038411 | 0.057984 | 0.085684 |
| 最小相对湿度 | 0.075105 | 0.043807 | 0.048462 | 0.058296 | 0.058991 | 0.062189 |
| 最低气温 | 0.068204 | 0.084075 | 0.080591 | 0.059172 | 0.047767 | 0.111528 |
| 本站气压4次平均 | 0.125691 | 0.103658 | 0.150436 | 0.103052 | 0.110209 | 0.092031 |
| 静稳指数 | 0.129112 | 0.143669 | 0.114901 | 0.179212 | 0.176645 | 0.160892 |
| 日照时数合计 | 0.083912 | 0.07657 | 0.061762 | 0.072485 | 0.079525 | 0.071517 |
| 气温4次平均 | 0.077685 | 0.06659 | 0.060457 | 0.07345 | 0.078357 | 0.049473 |
| 20-20降水量 | 0.057516 | 0.037519 | 0.016541 | 0.053953 | 0.037374 | 0.01928 |
| 10分钟风速4次平均 | 0.038477 | 0.059031 | 0.045526 | 0.041935 | 0.069803 | 0.071993 |
| 石家庄冬 | 通风系数 | 0.091338 | 0.041875 | 0.036343 | 0.080173 | 0.105556 | 0.064396 |
| 混合层高度 | 0.10259 | 0.110072 | 0.074694 | 0.1298 | 0.114058 | 0.053353 |
| 相对湿度4次平均 | 0.056182 | 0.032701 | 0.07488 | 0.033824 | 0.042481 | 0.117424 |
| 最高气温 | 0.057877 | 0.082733 | 0.180787 | 0.072419 | 0.092152 | 0.087695 |
| 日最大风速 | 0.049568 | 0.062506 | 0.061273 | 0.099774 | 0.069595 | 0.103989 |
| 最小相对湿度 | 0.087987 | 0.061563 | 0.09639 | 0.055465 | 0.061388 | 0.049206 |
| 最低气温 | 0.159133 | 0.137113 | 0.095316 | 0.109693 | 0.080735 | 0.162533 |
| 本站气压4次平均 | 0.10306 | 0.114425 | 0.090733 | 0.144429 | 0.14447 | 0.088534 |
| 静稳指数 | 0.101265 | 0.161384 | 0.090551 | 0.081239 | 0.088772 | 0.121108 |
| 日照时数合计 | 0.050228 | 0.025624 | 0.034627 | 0.036844 | 0.044075 | 0.045441 |
| 气温4次平均 | 0.073448 | 0.096964 | 0.100048 | 0.070292 | 0.062291 | 0.044675 |
| 20-20降水量 | 0.043013 | 0.036366 | 0.011962 | 0.045108 | 0.045088 | 0.01001 |
| 10分钟风速4次平均 | 0.02431 | 0.036673 | 0.052396 | 0.040941 | 0.049339 | 0.051636 |
| 邢台春 | 通风系数 | 0.076261 | 0.075615 | 0.065489 | 0.094073 | 0.071591 | 0.090923 |
| 混合层高度 | 0.04281 | 0.07316 | 0.071286 | 0.099515 | 0.084381 | 0.02948 |
| 相对湿度4次平均 | 0.073976 | 0.08124 | 0.07078 | 0.059253 | 0.040233 | 0.128125 |
| 最高气温 | 0.120973 | 0.11956 | 0.076406 | 0.080425 | 0.124364 | 0.093004 |
| 日最大风速 | 0.060173 | 0.064371 | 0.082543 | 0.079001 | 0.094811 | 0.086541 |
| 最小相对湿度 | 0.096258 | 0.06766 | 0.097607 | 0.088646 | 0.074433 | 0.073718 |
| 最低气温 | 0.095147 | 0.08862 | 0.088998 | 0.072578 | 0.065307 | 0.10089 |
| 本站气压4次平均 | 0.091949 | 0.088881 | 0.109696 | 0.061314 | 0.078369 | 0.108436 |
| 静稳指数 | 0.16586 | 0.174999 | 0.07746 | 0.142245 | 0.1318 | 0.122479 |
| 日照时数合计 | 0.025295 | 0.052718 | 0.061755 | 0.079238 | 0.108109 | 0.064619 |
| 气温4次平均 | 0.096732 | 0.073045 | 0.116845 | 0.085218 | 0.070482 | 0.067485 |
| 20-20降水量 | 0.018935 | 0.011753 | 0.008384 | 0.01365 | 0.019164 | 0.001933 |
| 10分钟风速4次平均 | 0.035631 | 0.028379 | 0.072751 | 0.044843 | 0.036957 | 0.032365 |
| 邢台夏 | 通风系数 | 0.093786 | 0.072503 | 0.068774 | 0.074432 | 0.071492 | 0.060723 |
| 混合层高度 | 0.119133 | 0.109122 | 0.093534 | 0.080138 | 0.065603 | 0.142913 |
| 相对湿度4次平均 | 0.064011 | 0.07173 | 0.054556 | 0.113412 | 0.053877 | 0.04332 |
| 最高气温 | 0.070002 | 0.088827 | 0.126471 | 0.061563 | 0.101719 | 0.063581 |
| 日最大风速 | 0.043135 | 0.064592 | 0.065293 | 0.083596 | 0.089117 | 0.096629 |
| 最小相对湿度 | 0.082532 | 0.074743 | 0.089137 | 0.059012 | 0.064946 | 0.068577 |
| 最低气温 | 0.073913 | 0.040856 | 0.072949 | 0.043582 | 0.061415 | 0.086522 |
| 本站气压4次平均 | 0.134176 | 0.051015 | 0.056473 | 0.083508 | 0.103061 | 0.089088 |
| 静稳指数 | 0.143299 | 0.153977 | 0.147107 | 0.101526 | 0.11167 | 0.129846 |
| 日照时数合计 | 0.06929 | 0.063109 | 0.09972 | 0.104128 | 0.07165 | 0.079202 |
| 气温4次平均 | 0.033033 | 0.102331 | 0.074841 | 0.035918 | 0.05901 | 0.077228 |
| 20-20降水量 | 0.018027 | 0.042291 | 0.018605 | 0.054417 | 0.075525 | 0.015634 |
| 10分钟风速4次平均 | 0.055662 | 0.064904 | 0.032541 | 0.104768 | 0.070916 | 0.046736 |
| 邢台秋 | 通风系数 | 0.053257 | 0.059313 | 0.037504 | 0.065384 | 0.078548 | 0.059856 |
| 混合层高度 | 0.056799 | 0.058514 | 0.111156 | 0.062348 | 0.079557 | 0.051747 |
| 相对湿度4次平均 | 0.04722 | 0.039899 | 0.070895 | 0.050192 | 0.032296 | 0.058489 |
| 最高气温 | 0.081597 | 0.075823 | 0.100736 | 0.109566 | 0.073919 | 0.069653 |
| 日最大风速 | 0.089888 | 0.073375 | 0.107995 | 0.061547 | 0.059562 | 0.059214 |
| 最小相对湿度 | 0.089374 | 0.08467 | 0.067683 | 0.081818 | 0.052056 | 0.062432 |
| 最低气温 | 0.096549 | 0.126396 | 0.062874 | 0.09241 | 0.091011 | 0.09936 |
| 本站气压4次平均 | 0.130671 | 0.113433 | 0.119739 | 0.074366 | 0.102014 | 0.166682 |
| 静稳指数 | 0.188472 | 0.17962 | 0.102898 | 0.123746 | 0.151123 | 0.129341 |
| 日照时数合计 | 0.073893 | 0.039787 | 0.073009 | 0.104044 | 0.097302 | 0.100372 |
| 气温4次平均 | 0.046843 | 0.05262 | 0.076663 | 0.057046 | 0.04735 | 0.060628 |
| 20-20降水量 | 0.017959 | 0.037383 | 0.030383 | 0.037195 | 0.034021 | 0.02907 |
| 10分钟风速4次平均 | 0.027479 | 0.059167 | 0.038465 | 0.080339 | 0.101242 | 0.053159 |
| 邢台冬 | 通风系数 | 0.072476 | 0.062484 | 0.080462 | 0.073035 | 0.089571 | 0.089417 |
| 混合层高度 | 0.063803 | 0.053915 | 0.040904 | 0.09835 | 0.080043 | 0.065611 |
| 相对湿度4次平均 | 0.05499 | 0.051102 | 0.056146 | 0.026642 | 0.047375 | 0.129133 |
| 最高气温 | 0.092952 | 0.065511 | 0.140701 | 0.079961 | 0.069577 | 0.070826 |
| 日最大风速 | 0.074759 | 0.11312 | 0.064074 | 0.071617 | 0.096515 | 0.07725 |
| 最小相对湿度 | 0.049317 | 0.048326 | 0.075143 | 0.052731 | 0.055065 | 0.033562 |
| 最低气温 | 0.117812 | 0.091101 | 0.125629 | 0.053423 | 0.047617 | 0.124929 |
| 本站气压4次平均 | 0.139828 | 0.208651 | 0.078135 | 0.123605 | 0.133239 | 0.083833 |
| 静稳指数 | 0.157477 | 0.119969 | 0.11379 | 0.23211 | 0.184675 | 0.139475 |
| 日照时数合计 | 0.047875 | 0.03329 | 0.028822 | 0.048034 | 0.050111 | 0.035198 |
| 气温4次平均 | 0.08836 | 0.069608 | 0.108626 | 0.0884 | 0.070007 | 0.072172 |
| 20-20降水量 | 0.016841 | 0.028801 | 0.001754 | 0.019711 | 0.040297 | 0.020406 |
| 10分钟风速4次平均 | 0.02351 | 0.054123 | 0.085814 | 0.032381 | 0.035908 | 0.058187 |

## 附表2——互信息系数值

附表2 两个城市不同季节污染物浓度与气象因子互信息系数

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | CO | NO2 | O3 | PM2.5 | PM10 | SO2 |
| 石家庄春 | 通风系数 | 0.171363 | 0.109329 | 0.197735 | 0.308611 | 0.167704 | 0.136855 |
| 混合层高度 | 0.258803 | 0.121278 | 0.222398 | 0.316902 | 0.224846 | 0.108209 |
| 相对湿度4次平均 | 0.179077 | 0.038103 | 0.043296 | 0.344774 | 0.249345 | 0.12932 |
| 最高气温 | 0.114955 | 0.129696 | 0.323248 | 0.001619 | 0.03238 | 0.098939 |
| 日最大风速 | 0.167749 | 0.129481 | 0.0947 | 0.227777 | 0.171811 | 0.114857 |
| 最小相对湿度 | 0.122202 | 0.082511 | 0.156832 | 0.209463 | 0.223576 | 0 |
| 最低气温 | 0.068114 | 0.088852 | 0.296355 | 0.005367 | 0.067111 | 0.053988 |
| 本站气压4次平均 | 0.080806 | 0.030623 | 0.11827 | 0.06172 | 0.11618 | 0.020074 |
| 静稳指数 | 0.321144 | 0.170563 | 0.186857 | 0.327368 | 0.203955 | 0.201801 |
| 日照时数合计 | 0.187923 | 0.198426 | 0.236037 | 0.297414 | 0.185274 | 0.177767 |
| 气温4次平均 | 0.121515 | 0.057604 | 0.465035 | 0.069174 | 0.04484 | 0.046067 |
| 20-20降水量 | 0 | 0 | 0.012785 | 0 | 0 | 5.47E-05 |
| 10分钟风速4次平均 | 0.163865 | 0.104682 | 0.155695 | 0.193229 | 0.156104 | 0.128786 |
| 石家庄夏 | 通风系数 | 0.087261 | 0.027757 | 0.093445 | 0.142079 | 0 | 0.001546 |
| 混合层高度 | 0.092808 | 0 | 0.07202 | 0.094636 | 0.028369 | 0.054319 |
| 相对湿度4次平均 | 0.123718 | 0.039568 | 0.075806 | 0.208083 | 0.079068 | 0.039105 |
| 最高气温 | 0 | 0 | 0.189973 | 0 | 0.009889 | 0 |
| 日最大风速 | 0.00293 | 0.022683 | 0 | 0.148275 | 0.113529 | 0.022183 |
| 最小相对湿度 | 0.144816 | 0.03045 | 0.075784 | 0.119547 | 0 | 0.071163 |
| 最低气温 | 0 | 0.033872 | 0.159869 | 0.071395 | 0.003071 | 0.024601 |
| 本站气压4次平均 | 0 | 0.050405 | 0.084455 | 0.004468 | 0.031526 | 0 |
| 静稳指数 | 0.237458 | 0.090809 | 0.08555 | 0.121434 | 0.064112 | 0.048971 |
| 日照时数合计 | 0.112322 | 0.073917 | 0.047656 | 0.211107 | 0.080929 | 0.058434 |
| 气温4次平均 | 0.082896 | 0.004258 | 0.21078 | 0.018812 | 0.044388 | 0 |
| 20-20降水量 | 0.01782 | 0.046373 | 0.006882 | 0 | 0 | 0.051217 |
| 10分钟风速4次平均 | 0.029557 | 0.011862 | 0.028021 | 0.143915 | 0.136277 | 0.062175 |
| 石家庄秋 | 通风系数 | 0.000715 | 0.076554 | 0.054272 | 0 | 0.009803 | 0.002183 |
| 混合层高度 | 0.009452 | 0 | 0.084367 | 0 | 0.006177 | 0.021819 |
| 相对湿度4次平均 | 0.099301 | 0.088908 | 0.170418 | 0.011583 | 0.027996 | 0.168764 |
| 最高气温 | 0.041156 | 0.068456 | 0.194627 | 0.018885 | 0.00306 | 0.015929 |
| 日最大风速 | 0.01034 | 0.101233 | 0.019227 | 0.012708 | 0.085112 | 0.028054 |
| 最小相对湿度 | 0.02314 | 0.07924 | 0.214371 | 0.09095 | 0.076882 | 0.147187 |
| 最低气温 | 0.099386 | 0.090997 | 0.061526 | 0.088054 | 0.11441 | 0.02361 |
| 本站气压4次平均 | 0.076683 | 0.130517 | 0.083941 | 0.048235 | 0.064008 | 0.02027 |
| 静稳指数 | 0.111221 | 0.148769 | 0.062537 | 0.124907 | 0.079555 | 0.06244 |
| 日照时数合计 | 0.059288 | 0.039898 | 0.100896 | 0.105775 | 0.084956 | 0.072586 |
| 气温4次平均 | 0.046992 | 0.080527 | 0.169255 | 0.063273 | 0.141058 | 0 |
| 20-20降水量 | 0.029993 | 0.047173 | 0.098366 | 0.001446 | 0.035314 | 0.090466 |
| 10分钟风速4次平均 | 0 | 0.114067 | 0 | 0.010165 | 0.074779 | 0.019431 |
| 石家庄冬 | 通风系数 | 0.106545 | 0.028586 | 0.083526 | 0.236314 | 0.114538 | 0.065523 |
| 混合层高度 | 0.150073 | 0.086353 | 0.192239 | 0.247924 | 0.165296 | 0.017347 |
| 相对湿度4次平均 | 0.09896 | 0.098427 | 0.096314 | 0.212757 | 0.198488 | 0.061334 |
| 最高气温 | 0.168916 | 0.010542 | 0.277893 | 0.049994 | 0.057096 | 0.196993 |
| 日最大风速 | 0.088645 | 0.078066 | 0.038348 | 0.162256 | 0.117276 | 0.094877 |
| 最小相对湿度 | 0.097137 | 0.096746 | 0.212424 | 0.262505 | 0.183727 | 0.026352 |
| 最低气温 | 0.112125 | 0 | 0.20978 | 0.029353 | 0.036085 | 0.221803 |
| 本站气压4次平均 | 0.020703 | 0 | 0.056745 | 0.126159 | 0.109378 | 0.023439 |
| 静稳指数 | 0.066231 | 0.144332 | 0.187023 | 0.176854 | 0.093352 | 0.016986 |
| 日照时数合计 | 0.176965 | 0.091247 | 0.195712 | 0.17225 | 0.159276 | 0.107345 |
| 气温4次平均 | 0.184235 | 0.052776 | 0.335392 | 0.003731 | 0.014087 | 0.216054 |
| 20-20降水量 | 0.037359 | 0.040473 | 0.054382 | 0.016572 | 0.075265 | 0.088088 |
| 10分钟风速4次平均 | 0.149272 | 0.073943 | 0.086744 | 0.153958 | 0.163979 | 0.030185 |
| 邢台春 | 通风系数 | 0.0935 | 0.099183 | 0.160413 | 0.118831 | 0.14106 | 0.180246 |
| 混合层高度 | 0.078138 | 0.151194 | 0.142396 | 0.252383 | 0.215031 | 0.202113 |
| 相对湿度4次平均 | 0.127486 | 0.107769 | 0.172164 | 0.275212 | 0.195429 | 0.103023 |
| 最高气温 | 0.108892 | 0.062936 | 0.316808 | 0.038001 | 0.0877 | 0.074997 |
| 日最大风速 | 0.129278 | 0.122473 | 0.161181 | 0.125758 | 0.192911 | 0.219121 |
| 最小相对湿度 | 0.114416 | 0.132664 | 0.187665 | 0.280088 | 0.113462 | 0.042453 |
| 最低气温 | 0.12597 | 0 | 0.31326 | 0.0385 | 0.00538 | 0.035284 |
| 本站气压4次平均 | 0.006575 | 0.07088 | 0.059594 | 0.070484 | 0.096728 | 0.128906 |
| 静稳指数 | 0.245083 | 0.186636 | 0.178415 | 0.143415 | 0.115569 | 0.24736 |
| 日照时数合计 | 0.01251 | 0.060437 | 0.213894 | 0.21555 | 0.071266 | 0.049846 |
| 气温4次平均 | 0.107814 | 0.065867 | 0.399061 | 0 | 0.055726 | 0.132854 |
| 20-20降水量 | 0 | 0.005708 | 0 | 0.012242 | 0.007809 | 0.00576 |
| 10分钟风速4次平均 | 0.027 | 0.02333 | 0.04518 | 0.089305 | 0.132942 | 0.074593 |
| 邢台夏 | 通风系数 | 0 | 0.013956 | 0 | 0.224366 | 0.160255 | 0 |
| 混合层高度 | 0.00031 | 0 | 0 | 0.144414 | 0.030682 | 0.032597 |
| 相对湿度4次平均 | 0 | 0.025217 | 0.034395 | 0.215376 | 0.031863 | 0.090113 |
| 最高气温 | 0 | 0 | 0.286389 | 0 | 0 | 0.066594 |
| 日最大风速 | 0.078279 | 0 | 0.052567 | 0.143599 | 0.184279 | 0.01013 |
| 最小相对湿度 | 0 | 0.064706 | 0.043232 | 0.124725 | 0 | 0.013237 |
| 最低气温 | 0 | 0.000563 | 0.176048 | 0.015881 | 0.000615 | 0.063137 |
| 本站气压4次平均 | 0.034227 | 0.000516 | 0.018757 | 0.121269 | 0.111643 | 0.022958 |
| 静稳指数 | 0.201529 | 0.104564 | 0.090821 | 0.114112 | 0.055806 | 0.129175 |
| 日照时数合计 | 0.072284 | 0.083901 | 0.083944 | 0.099071 | 0.098811 | 0.059646 |
| 气温4次平均 | 0.011752 | 0.028807 | 0.263659 | 0.062558 | 0.01798 | 0 |
| 20-20降水量 | 0 | 0.031478 | 0.035634 | 0 | 0.009235 | 0.069553 |
| 10分钟风速4次平均 | 0.002025 | 0.066651 | 0 | 0.143661 | 0.212472 | 0.093799 |
| 邢台秋 | 通风系数 | 0.038956 | 0 | 0.060578 | 0.04452 | 0 | 0.030921 |
| 混合层高度 | 0.014761 | 0.045131 | 0.169349 | 0.076315 | 0.033129 | 0.089735 |
| 相对湿度4次平均 | 0 | 0.062915 | 0.160884 | 0.024243 | 0 | 0.135667 |
| 最高气温 | 0.005056 | 0.060384 | 0.287504 | 0.018202 | 0.103669 | 0.133622 |
| 日最大风速 | 0 | 0.038298 | 0.019819 | 0.0389 | 0.086072 | 0.11787 |
| 最小相对湿度 | 0 | 0.048654 | 0.107784 | 0.005722 | 0.028248 | 0.097779 |
| 最低气温 | 0 | 0.121523 | 0.045003 | 0 | 0.047415 | 0.02664 |
| 本站气压4次平均 | 0.044476 | 0.083043 | 0.144279 | 0.066725 | 0.046991 | 0.013067 |
| 静稳指数 | 0.129524 | 0.198476 | 0.041818 | 0.113638 | 0.175065 | 0.030304 |
| 日照时数合计 | 0.001892 | 0 | 0.166933 | 0.024667 | 0.03272 | 0 |
| 气温4次平均 | 0 | 0.019709 | 0.273377 | 0.084864 | 0.077946 | 0.081408 |
| 20-20降水量 | 0.005582 | 0.07518 | 0.050423 | 0 | 0.076502 | 0.069109 |
| 10分钟风速4次平均 | 0 | 0.04242 | 0.174153 | 0.049296 | 0.081624 | 0.018661 |
| 邢台冬 | 通风系数 | 0.027971 | 0.081938 | 0.070949 | 0.169237 | 0.163658 | 0.166282 |
| 混合层高度 | 0.062921 | 0.054886 | 0.191733 | 0.164683 | 0.197355 | 0.081968 |
| 相对湿度4次平均 | 0.07085 | 0.101837 | 0.205154 | 0.123272 | 0.151313 | 0.153535 |
| 最高气温 | 0.177796 | 0.019832 | 0.356104 | 0.056375 | 0.073875 | 0.111814 |
| 日最大风速 | 0.017295 | 0.073277 | 0 | 0.117733 | 0.133122 | 0.069542 |
| 最小相对湿度 | 0.027549 | 0.146105 | 0.190268 | 0.184174 | 0.070193 | 0.020902 |
| 最低气温 | 0.066641 | 0.036131 | 0.209907 | 0.022269 | 0.006716 | 0.153066 |
| 本站气压4次平均 | 0.017748 | 0.084162 | 0.019522 | 0 | 0 | 0.004705 |
| 静稳指数 | 0.222072 | 0.324382 | 0.200136 | 0.223302 | 0.177456 | 0.129275 |
| 日照时数合计 | 0.06652 | 0.06807 | 0.101072 | 0.203267 | 0.235735 | 0.031916 |
| 气温4次平均 | 0.114138 | 0.038729 | 0.28612 | 0 | 0 | 0.122455 |
| 20-20降水量 | 0.047606 | 0.03486 | 0.033243 | 0.019546 | 0.020739 | 0.084899 |
| 10分钟风速4次平均 | 0.129675 | 0.127327 | 0.112689 | 0.157702 | 0.10862 | 0.100244 |