基于 ERA-interim 数据和机器学习的 2016 年北京地区黑碳反演方法

一. 背景

黑碳气溶胶是大气气溶胶的重要组成部分，黑碳主要是由富含碳的物质不完全燃烧产生的，比如化石原料和生物质原料燃烧等，现实生活中如汽车，火车轮船飞机的尾气排放中都含有黑碳。黑碳的粒度仅 0.01~0.05 微米，属于 pm2.5 中数量最多危害最大一种污染物。从其性质上看黑碳是一种强烈的吸光性物质，颗强烈吸收太阳短波辐射，同时释放红外辐射加热周围大气，因此可以产生区域增温效应，其增温效果要高于 co2。黑碳的反照率也低，如果沉降将在冰雪表面，会降低表面反照率，加速冰雪的消融，目前北极和高亚洲冰川都发现了黑碳的存在，对全球变暖有着推动的作用。黑碳也属于 2B 类致癌物质，对人体的心血管，神经，呼吸系统都有损害。

近十多年来，国家对 pm2.5 的重视加大，PM2.5 遥感方法的研究在多个环境相关研究领域得到了广泛的发展和应用，利用卫星数据反演 pm2.5 的方法已经做得相对比较成熟。而关于 pm2.5 组成分中的危害巨大的黑碳获得的关注度没有那么高。而我国目前仍是黑碳排放量最多的国家，但根据日本国立研究开发法人海洋研究开发机构的观测显示中国排放的黑碳在过去 10 年的时间里大幅减少了 40%。目前我国能够检测黑碳的站点数只有 43 无法覆盖到中国全境，如果能够利用卫星数据对黑碳进行反演，那么就可以低成本地了解到中国的黑碳分布情况。对于后续的治理和政策制定都有积极推动作用。

二. 数据 & 方法

1. 数据处理

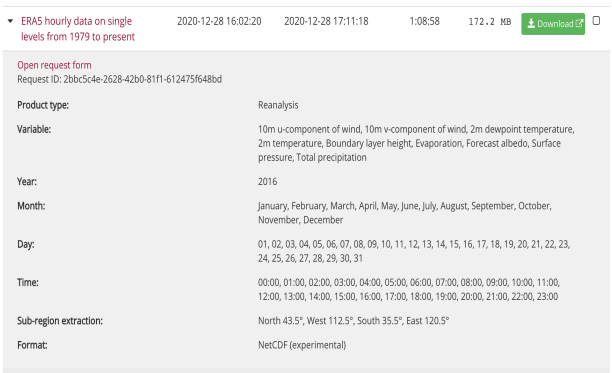

图 1

number	station	x	y	number	station	x	y
51058	阿克达拉	87.967	47.1	54662	大连	121.64	38.908
51747	塔中	83.667	39	54725	惠民	117.533	37.5
52203	哈密	93.517	42.817	55591	拉萨	91.133	29.667
52267	额济纳旗	101.067	41.95	56294	成都	104.0167	30.66667
52418	敦煌	94.683	40.15	57083	郑州	113.667	34.7
53276	朱日和	112.9	42.4	57131	泾河	108.967	34.433
53646	榆林	109.783	38.267	57278	襄樊	112.077	32.003
53787	榆社	112.983	37.067	57461	宜昌	111.22	30.44
54102	锡林浩特	116.07	43.57	57494	武汉	114.051	30.598
54135	通辽	122.267	43.6	57596	金沙	114.205	29.635
54161	长春	125.217	43.9	57957	桂林	110.3	25.317
54339	鞍山	122.999	41.09	58215	寿县	116.783	32.433
54342	沈阳	123.51	41.733	58370	浦东	121.533	31.233
54346	本溪	123.775	41.307	58448	临安	119.7	30.217
54351	抚顺	124.069	41.926	58457	杭州	120.167	30.233
54421	密云上甸子	117.117	40.65	58477	定海	122.1	30.033
54499	昌平	116.217	40.217	58506	庐山	115.983	29.567
54511	北京	116.467	39.8	58659	温州	120.65	28.033
54594	大兴	116.21	39.43	58665	洪家	121.417	28.617
				59287	广州	113.482	23.21
				59289	东莞	113.739	22.966
				59431	南宁	108.55	22.783
				59481	番禺	113.319	22.938
				59493	深圳	114.003	22.542

图 2

如图 1，欧洲再分析数据变量有：10m u-component of wind; 10m v-component of wind; 2m temperature; Boundary layer height; 2m dewpoint temperature; Evaporation; Forecast albedo; Surface pressure; Total precipitation，时间分辨率为 1 小时，空间分辨率为 0.25 度，时间范围为 2016 年全年。2016 年黑碳数据，由气象局提供，其站点分布如图 2。绘制了一张 2016 年 1 月 1 日 8:00 时京津冀 2m 高的温度图如图 3，可以大致判断温度分布的趋势合理，温度范围合理。

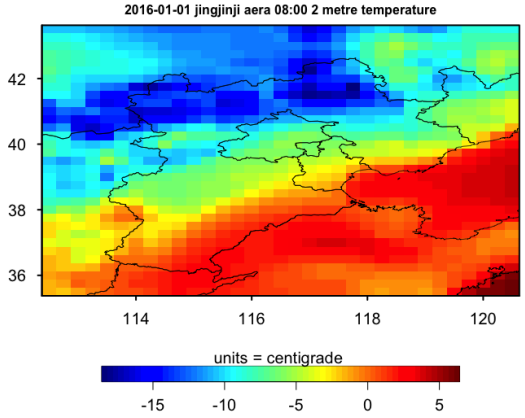

图 3

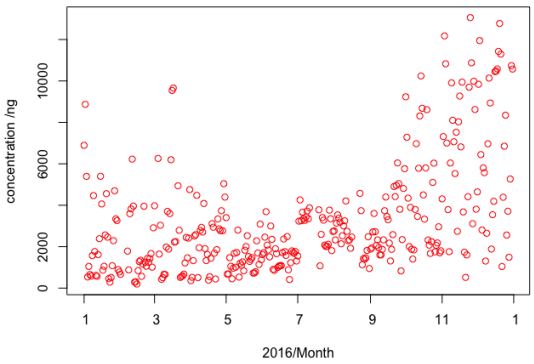

图 4

图 4 为昌平地区 2016 年全年的黑碳浓度分布图，黑碳浓度单位为 ng/m3，可以看出高浓度黑碳主要集中在春季和冬季。将 u-component of wind 和 v-component of wind 根据公式（1）和公式（2）转化成了风速（WS）和风向（WD），单位分别为 m/s 和度。画出各变量之间的关系如图 5。

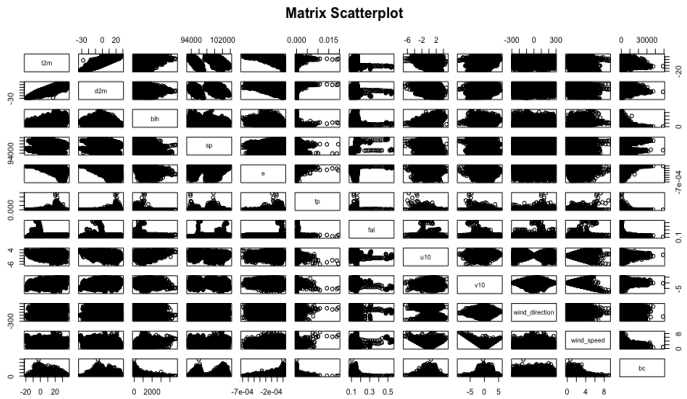

图 5

通过图 5 可以观察到 t2m 和 d2m 变量之间存在一个较为明显的线性关系，之后利用 d2m 和 t2m，根据公式（3）：Tetens 公式计算出实际水汽压和饱和水汽压，然后相比较计算出当前的相对湿度，其中 t 为摄氏度，然后将 d2m 变量去掉。

$$u - wind = WS * \sin(WD/180 * \pi) \quad \text{（公式 1）}$$

$$v - wind = WS * \cos(WD/180 * \pi) \quad \text{（公式 2）}$$

$$\mathrm{E} = 6.112\exp\left(17.67 * \mathrm{t}/(\mathrm{t} + 243.5)\right) \quad \text{（公式 3）}$$

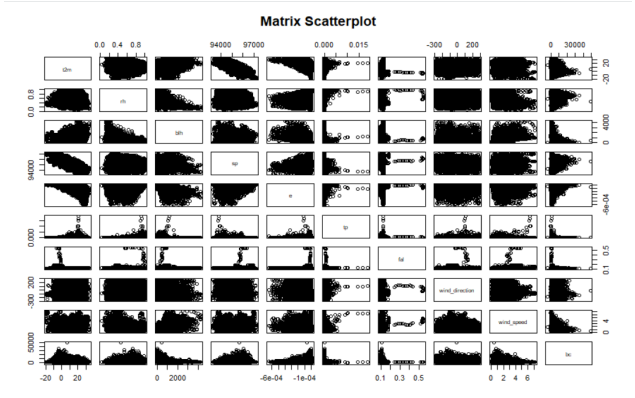

图 6

图 6 为更新变量后各变量之间的关系。

2. 方法

（1） 线性回归

```
> model <-lm(bc ~ t2m + rh + e + fal + wind_direction + wind_speed + blh + sp + tp, data = all_use)
> coef(model)
   (Intercept)            t2m             rh              e            fal wind_direction
  3.457830e+04  -8.385147e+01   2.740597e+03  -4.991694e+06  -1.995386e+04   1.436017e+00
    wind_speed            blh             sp             tp
 -3.724136e+02  -2.731020e-01  -3.092918e-01  -5.294546e+05
```

图 7

如图 7 所示使用线性回归对这些变量进行拟合，得到个变量的系数。图 8 可以看到各变量的 P 值都小于 0.05，所以该模型的变量是显著的。但是 R-square 的值较低，对于目的是预测值来说并不好。接下来将使用机器学习中随机森林的方法对这些数据进行拟合。

```
> summary(model)

Call:
lm(formula = bc ~ t2m + rh + e + fal + wind_direction + wind_speed +
    blh + sp + tp, data = all_use)

Residuals:
   Min     1Q Median     3Q    Max
 -6851  -2109   -668   1086  45162

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.458e+04  7.279e+03   4.750 2.06e-06 ***
t2m            -8.385e+01  5.853e+00 -14.325  < 2e-16 ***
rh              2.741e+03  2.033e+02  13.478  < 2e-16 ***
e              -4.992e+06  4.944e+05 -10.097  < 2e-16 ***
fal            -1.995e+04  1.650e+03 -12.091  < 2e-16 ***
wind_direction  1.436e+00  2.734e-01   5.252 1.54e-07 ***
wind_speed     -3.724e+02  4.117e+01  -9.046  < 2e-16 ***
blh            -2.731e-01  8.591e-02  -3.179  0.00148 **
sp             -3.093e-01  7.577e-02  -4.082 4.50e-05 ***
tp             -5.295e+05  7.278e+04  -7.274 3.78e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3285 on 8766 degrees of freedom
Multiple R-squared:  0.1218,    Adjusted R-squared:  0.1209
F-statistic: 135.1 on 9 and 8766 DF,  p-value: < 2.2e-16
```

图 8

（2） 随机森林

使用随机森林这个机器学习方法来对数据进行拟合，在拟合之前，使用 MinMaxScaler()将所有变量都进行归一化操作，如图 9。

```
#all data
all_data = pd.read_csv('data_rh_all.csv')
X1 = all_data.iloc[:,1:10].values
print('all_data = ',X1)
Y = all_data.iloc[:,10].values
print(Y)
scaler=MinMaxScaler()
X = scaler.fit_transform(X1)
print(X)
print(X1[1,:])
```

图 9

```
#Fitting Random_Forest Classification Model
print('模型拟合中')
from sklearn.ensemble import RandomForestRegressor
classifier = RandomForestRegressor(n_estimators =200,criterion='mse',
                                   random_state=0)
classifier.fit(X_train,Y_train)
print('模型拟合完毕')

#Predicting the test set results
Y_train_pred = classifier.predict(X_train)
Y_test_pred = classifier.predict(X_test)

print("MSE Train: %.3f, Test: %.3f" % (mean_squared_error(Y_train,Y_train_pred),
                                       mean_squared_error(Y_test,Y_test_pred)))
print("R2_Score Train: %.3f, Test: %.3f" % (r2_score(Y_train,Y_train_pred),
                                            r2_score(Y_test,Y_test_pred)))

#判断特征重要性
names = ['t2m','rh','sp','e','tp','fal','wind_direction','wind_speed']
print("Features sorted by their score:")
print(sorted(zip(map(lambda x: round(x, 4), classifier.feature_importances_), names), reverse=True))
```

图 10

然后调用随机森林，对北京地区昌平，大兴，密云三个地区的 2016 年所有数据进行拟合，输出 MSE 和 R^2 以及特征重要行排序，并画出预测值和真实值的图像，代码如图 10。

三. 结果

（1）模型表现

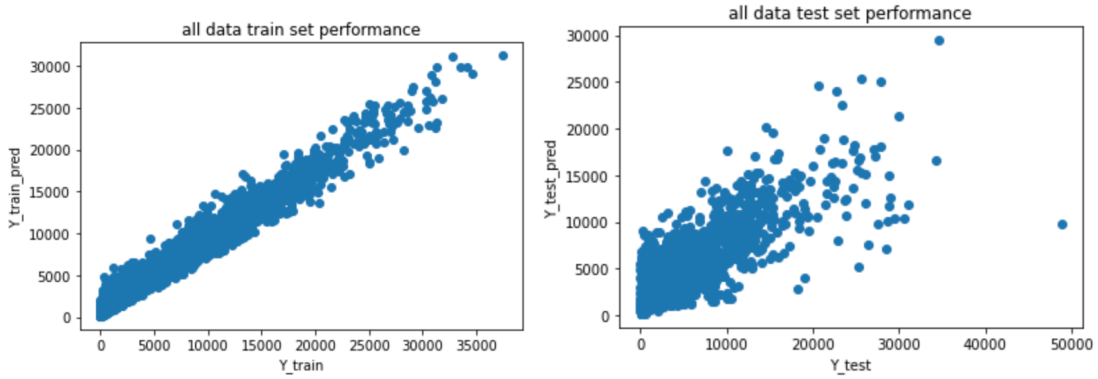

图 11

```
MSE Train: 716745.429, Test: 6306043.290
R2_Score Train: 0.960, Test: 0.685
Features sorted by their score:
[(0.243, 'sp'), (0.2223, 'fal'), (0.1136, 't2m'), (0.1021, 'rh'), (0.0846, 'wind_speed'), (0.0818, 'blh'), (0.0742,
'e'), (0.0587, 'wind_direction'), (0.0197, 'tp')]
```

图 12

如图 11，从左到右分别为训练集和测试集在该模型下的真实值和预测值的比较。图 12 为该模型下的训练集和测试集的 R^2 和 MSE 一级个变量的重要性排序。可以看到训练集的表现很好但是测试集的表现不太好，差距太大。

（2）模型应用

虽然模型目前表现不是很好，但是把整个流程做完，将该模型应用于昌平(116.217,40.217)，大兴 (116.21,39.43)，密云上甸子(117.117,40.65)三个站点附近范围为(纬度:39-41.25; 经度:115.75-117.5) 2016 年 1 月 1 日 8:00 的黑碳浓度预测浓度，并绘图如图 13 所示。.

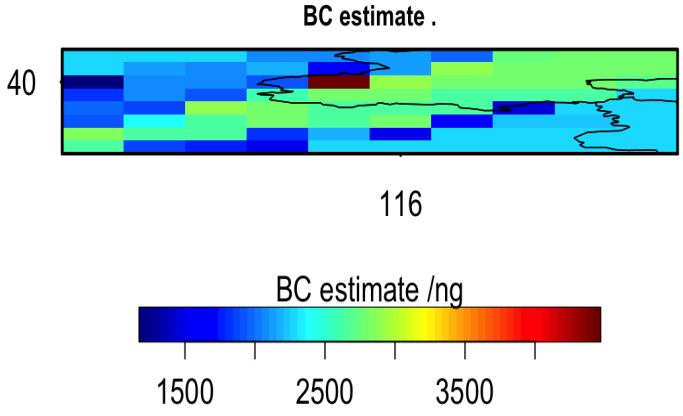

图 13

四. 思考

（1） 该模型变量的选择是依靠卫星反演 pm2.5 主流模型的气象变量选择，多加了一个反照率变量。MODI 每天过境北京地区的时间是固定的时间 10:00-14:00，所以如果使用其 AOD 数据在时间经度上只能适用于日变化。该项目的时间分辨率为小时。

（2） 模型变量的选择上这些气象变量不足以很好地繁衍出黑碳的浓度，需要根据黑碳的物理化学特性增加删除一些变量。

（3） 本次项目缺乏时间和空间上的联系，以后需要加以考虑进来。