颗粒物组分质控说明文档

1. 研发背景

随着我国工业化和城市化进程加快，化石燃料消耗量不断加大，同时机动车保有量急剧增长，导致我国大气环境污染问题日益凸显。大量对雾霾现象研究的不断深入表明影响雾霾现象主要原因是大气颗粒物污染，特别是细颗粒物 PM25 污染。研究发现大气 PM25 中主要由水溶性离子、碳质组分和无机元素等组成，其中：

水溶性离子主要包括 SO42-、NO3-、NH4+、Na+、Cl-、K+、Mg2+、Ca2+等，其中 SO42-、NO3-和 NH4+（二次无机离子，SNA）是水溶性离子中主要化学组分，水溶性离子总质量浓度在 PM25 质量浓度中占比范围为 30～70%，且在极端天气条件下该比值可能会有所上升。SO42-主要来源于 SO2 在大气中氧化反应，其主要包括光化学作用、气相均相中的氧化反应和液相中的氧化反应等，气相均相作用是指空气中存在的 SO2 于羟基等自由基之间发生的氧化反应，液相反应是指大气中存在的 SO2 在液态表明被氧化的反应；NO3-主要通过 NOx 转化而来，主要转化途径分别为白天羟基等自由基将 NO2 氧化，与 NH3 发生作用后形成硝酸盐，夜间 NO2 与 O3 发生作用氧化后在颗粒物表面水合形成硝酸盐；NH4+主要通过空气中氨气与酸性气体发生反应后形成；Na+来源与海洋排放和土壤扬尘等；Cl-主要来源包括化石燃料燃烧产生和海盐离子的挥发等；K+主要是由于生物质燃烧排放产生；Mg2+和 Ca2+通常来源于土壤扬尘。

炭质组分主要包括有机碳 OC 和元素碳 EC，此外还含有少量碳酸盐，通常因含量较低而被忽略。OC 主要由一次有机碳 POC 和二次有机碳 SOC 组成，其中 POC 主要来源于化石燃料燃烧、生物质等燃烧直接排放，而 SOC 主要来源于大气中存在可挥发性的有机物经过化学反应，而形成的污染物；EC 则主要来源于化石燃料等污染物的直接排放。

无机元素包含组分较多，现已经测定元素共 70 多种，其中包括地壳元素（Si、Al、Ca、Mg 等）和微量元素（As、Pb、Hg 等）。无机元素主要来源于人为源和自然源排放，并且由于多数无机元素性质较为稳定，因此主要为一次颗粒物。

由于地壳中 Si、Al、Ca、Mg 等元素于土壤扬尘和生产水泥等源类活动水平相关，因此常被作为地壳物质的失踪元素。

为更好的获取上述各类组分数据，各地颗粒物组分观测超级站可实时采集组成数据，受仪器稳定性影响，目前组分网组分数据质量较差，存在大量缺失和异常值。通过了解组分观测数据审核历史经验，在经验指标的基础上，设计灵活适用性更高的自动审核方案，并结合机器学习方法，拓展目前质量控制方法体系以提高所采集颗粒物组分数据的准确性和有效性。

2. 研发目标

- 结合总站审核指标，提出基于统计方法的颗粒物质控方案；
- 基于历史数据分析，提出基于机器学习算法的颗粒物质控方案；
- 设计可执行的方案，开发相关功能模块，测试各种方案的效果；
- 针对各阶段的问题，提出针对性的优化方案，开展功能优化；

3. 方法综述

3.1 经验指标

通过调研相关文献，并结合目前审核规范，针对不同组分内容，设计一套多维度的审核指标用于控制组分数据质量。

3.1.1 水溶性离子

针对水溶性离子在线观测，主要审核组分包括：SO42-、NO3-、NH4+、Na+、Cl-、K+、Mg2+、Ca2+、F-等，同类气态暂不纳入数据有效性考核范围。与水溶性离子质控相关指标：

- **阴阳离子平衡 AE/CE**：阴阳离子电荷浓度平衡指数。

$$AE = \frac{SO_4^{2-}}{48} + \frac{NO_3^-}{62} + \frac{Cl^-}{35.5} + \frac{F^-}{19}$$

$$CE = \frac{Na^+}{23} + \frac{NH_4^+}{18} + \frac{K^+}{39} + \frac{Mg^{2+}}{12} + \frac{Ca^+}{20}$$

- **二次离子 SNA 平衡**：针对 SNA 主要三类组分，计算阴阳离子平衡指数，观测其是否在指定阈值 0.7 ~ 1.3 之间。

$$SNA\_rate = \left(\frac{SO_4^{2-}}{48} + \frac{NO_3^-}{62} \right) / \left(\frac{NH_4^+}{18} \right) \epsilon \ [0.7, 1.3]$$

- **单组分阈值**：单组分在一段时间范围内出现单点的异常高值和异常低值。

$$curr\_val \ vs. \ lower\&upper\_thd$$

- **二次离子 SNA 数据波动**：当 SO42-、NH4+、NO3- 三类组分数据在前后观测时次偏差均超过 40%则该数据可能异常。

$$\frac{abs(curr\_val - prev\_val)}{prev\_val} \ \epsilon \ [0, 0.4]$$

$$\frac{abs(next\_val - curr\_val)}{curr\_val} \ \epsilon \ [0, 0.4]$$

式中：

AE	阳离子电荷浓度
CE	阴离子电荷浓度
curr_val	当时次观测值
prev_val	前一时次观测值
next_val	后一时次观测值
lower_thd	阈值下限
upper_thd	阈值上限

3.1.2 OCEC 离子

碳组分在线观测数据，主要审核组分在同一观测时次与 PM25 的占比关系、占比值在前后时次的波动情况、组分之间比例关系，相关指标有：

- OC/PM25：OC 占 PM25 的比例超过 30%可能为异常。
- EC/PM25：EC 占 PM25 的比例超过 20%可能为异常。
- (OC+EC)/PM25：OC 与 EC 之和占 PM25 的比例超过 50%可能为异常。

$$\frac{OC}{PM25} \ \epsilon \ [0, 0.3] \ \& \ \frac{EC}{PM25} \ \epsilon \ [0, 0.2] \ \& \ \frac{OC + EC}{PM25} \ \epsilon \ [0, 0.5]$$

- OC/EC：OC 与 EC 的比值一般在 0.6～10 之间，超过范围则肯能异常。

$$\frac{OC}{EC} \epsilon [0.6, 10]$$

- **单组分** PM25 占比波动：OC 组分占 PM 的值在观测前后时次的偏差均超过 100%则当时次 OC 数据异常；EC 组分占 PM 的值在观测前后时次的偏差均超过 300%则当时次 EC 数据异常。

$$ratio = \frac{OC}{PM}$$

$$\frac{abs(curr\_ratio - prev\_ratio)}{prev\_ratio} \& \frac{abs(next\_ratio - curr\_ratio)}{prev\_ratio} \epsilon [0, 100\%]$$

$$ratio = \frac{EC}{PM}$$

$$\frac{abs(curr\_ratio - prev\_ratio)}{prev\_ratio} \& \frac{abs(next\_ratio - curr\_ratio)}{prev\_ratio} \epsilon [0, 300\%]$$

3.1.3 重金属离子

所有组分之和超过 PM25 数据则整组观测数据异常，进一步判断异常组分；单一组分的重金属离子数据整体较小，考虑单组分观测出现异常高值和低值的离群点为异常数据，针对特点组分存在特点阈值范围，其余大部分阈值均在 0～1 之间：

$$Fe \in [0, 2],\ \{Ca, Si\} \in [0.1, 1],\ Al \in [0, 5],\ other \in [0, 1]$$

3.1.4 小结

通过调研文献可知，不同区域的颗粒物组分受地形、气象条件、经济产业结构等因素影响，其污染分布和颗粒物组分特征都不相同，并表现出一定的季节特征，由此分析目前审核规范存在优劣方面：

- 优势：结合了历史观测数据和专家经验的分析，针对不同组分提出了不同的异常审核指标，其定义便于量化考察，同时有效性也经过考验。
- 劣势：确定量化指标的同时给定确定阈值，不易适用于分布在全国各地的观测站点实际数据情况，同时不同季节观测数据存在较大差别，审核指标也需要进行适当调整。

3.2 统计方法

3.2.1 简介

针对审核规范目前无法进行地区和时间的调整, 对组分网 2019 年 82 个观测站点的数据进行提取分析, 查看现有指标确定阈值的审核情况, 对各项指标在一段时间内容的观测数据进行计算并获取其统计分布情况, 选择置信区间内的统计量作为确定阈值的更新值, 从而达到不同站定不同时间段阈值的动态适配。

3.2.2 方案

- **数据准备**

原始数据从数据库提取, 先取得指定时间范围的观测数据, 该数据以观测时次为文件, 包含该时次所有观测站点所有组分数据, 并以此作为质控输入数据集。

通过提取站点在每个时次的观测数据, 形成 82 个站点在指定时间范围内的所有组分观测序列数据, 该数据集以站点信息为文件, 作为指标统计的输入数据。

- **统计指标**

结合经验指标所设计内容, 对每个站点的每类组分, 对应计算在一段时期内每个观测时次相应指标的数据, 其中: 水溶性离子包括 AE、CE、阴阳离子平衡、SNA 组分平衡、SNA 组分观测时次变化比率; OCEC 离子包括 PM25 占比、OC/EC 比例、PM25 占比观测时次变化比率; 重金属离子包括各组分异常的剔除。

针对以上各组分计算所得指标, 通过计算其统计分布, 参考经验阈值的确定阈值, 对各指标在 90%以上分位数进行对比分析, 选取 95 ~ 97%百分位数作为统计量参数以动态调整统计阈值, 其中部分统计过程如下:

	count	mean	std	min	2%	5%	10%	50%	80%	90%	95%	97%	99%	max
PM10	8704.0	68.939290	65.321860	0.007000	7.320600	12.566000	18.192000	47.615000	100.522000	149.090000	193.973000	229.332400	310.069300	916.600000
PM25	8705.0	45.263637	40.918336	0.039000	6.783200	9.456400	12.340000	32.940000	65.660000	91.640000	128.688000	147.350800	196.977600	425.020000
PM10/PM25	8693.0	2.379388	9.450552	0.000412	0.257582	0.437419	0.591460	1.451844	2.644588	3.940394	5.933142	7.931307	12.836278	450.576923
PM10-PM25	8693.0	23.660967	55.465864	-394.880000	-53.415600	-33.292000	-19.146000	14.540000	45.188000	72.298000	104.096000	139.047200	240.203200	839.790000
diff_PM10	8672.0	-0.006229	34.591530	-727.506000	-55.297200	-32.848000	-21.349000	-0.135000	12.800000	21.010000	30.794500	39.529600	83.855400	724.316000
diff_PM25	8673.0	0.002600	15.828890	-378.550000	-24.943440	-14.114000	-9.068000	0.212000	5.456000	9.120000	13.584200	17.733600	33.834000	305.510000
PM10_ratio	8672.0	2.370677	90.059293	0.000000	0.006951	0.017322	0.033678	0.202093	0.443169	0.647608	0.921192	1.232938	2.487479	6995.250000
PM25_ratio	8673.0	0.398026	8.358597	0.000000	0.004261	0.010794	0.021059	0.122297	0.278226	0.421960	0.595552	0.754387	1.263521	596.179487

颗粒物指标统计

	PM10	PM25	OC	EC	oc/pm25	ec/pm25	ocec/pm25	oc/ec
count	8583.000000	8583.000000	8583.000000	8583.000000	8583.000000	8583.000000	8583.000000	8583.000000
mean	68.871765	45.334546	10.227934	1.994792	0.488306	0.082334	0.570640	4.724572
std	65.520722	40.981893	519.936839	66.040908	32.702307	4.157273	36.856863	2.801153
min	0.007000	0.039000	0.040000	0.010000	0.000816	0.000666	0.003467	0.166667
5%	12.551000	9.491000	1.096400	0.230000	0.046671	0.009911	0.065042	1.203337
10%	18.150000	12.370000	1.355200	0.287000	0.058786	0.012650	0.079423	1.586516
50%	47.430000	32.950000	3.860000	0.915000	0.110442	0.026792	0.141794	4.427640
80%	100.428000	65.736000	6.451600	1.752600	0.160160	0.044839	0.198405	6.306906
85%	119.517000	75.967000	7.140700	2.046700	0.174645	0.050517	0.214325	6.817113
90%	149.152000	91.672000	8.211600	2.482800	0.197793	0.058787	0.239045	7.617373
95%	194.317000	128.901000	10.199600	3.466900	0.247447	0.073531	0.294184	8.943691
97%	229.643200	147.670800	12.205620	4.460640	0.287893	0.084069	0.349394	9.997490
99%	311.280400	197.476800	16.190140	6.893560	0.429472	0.109014	0.503344	12.711072
max	916.600000	425.020000	48171.970000	6118.159000	3029.683648	384.789874	3414.473522	90.600000

OCEC 离子指标统计

	PM25	Hg	Br	As	Se	Te	V	Ti	Ba	Sc	Pd	Co	Mo
count	8705.000000	7748.000000	8251.000000	6260.000000	5705.000000	1019.000000	3412.000000	7581.000000	7454.000000	5.426000e+03	1459.000000	5135.000000	4400.000000
mean	45.263637	0.001285	0.024287	0.022493	0.008322	0.926707	0.006361	0.044468	0.054525	4.178224e-03	0.107136	0.021901	0.073918
std	40.918336	0.002534	0.022886	0.604443	0.012486	10.414635	0.031742	0.074506	0.202249	6.564319e-03	1.440946	1.134479	0.070389
min	0.039000	0.000010	0.000007	0.000001	0.000001	0.000287	0.000001	0.000053	0.000030	7.000000e-07	0.000026	0.000001	0.000004
50%	32.940000	0.000720	0.018909	0.005765	0.006874	0.116591	0.002507	0.025472	0.026802	2.904000e-03	0.014289	0.005001	0.059649
70%	51.448000	0.001100	0.027007	0.008969	0.009999	0.217795	0.004352	0.036984	0.040798	4.925500e-03	0.027207	0.007378	0.092326
80%	65.660000	0.001501	0.033875	0.011441	0.012006	0.361487	0.005980	0.048916	0.053475	6.508000e-03	0.051974	0.008926	0.114816
90%	91.640000	0.002509	0.049052	0.014946	0.015348	0.811045	0.010981	0.083445	0.082557	8.999000e-03	0.116930	0.011236	0.149786
95%	128.688000	0.004083	0.064900	0.018050	0.018709	1.498306	0.019111	0.176618	0.153770	1.130075e-02	0.263904	0.013402	0.176741
99%	196.977600	0.009834	0.104456	0.031395	0.031045	5.716154	0.037059	0.332397	0.403430	1.758125e-02	0.700335	0.019050	0.298560
max	425.020000	0.093739	0.979907	30.725630	0.553440	226.705400	0.882820	1.663812	7.307677	3.016530e-01	51.314590	81.297870	1.714948

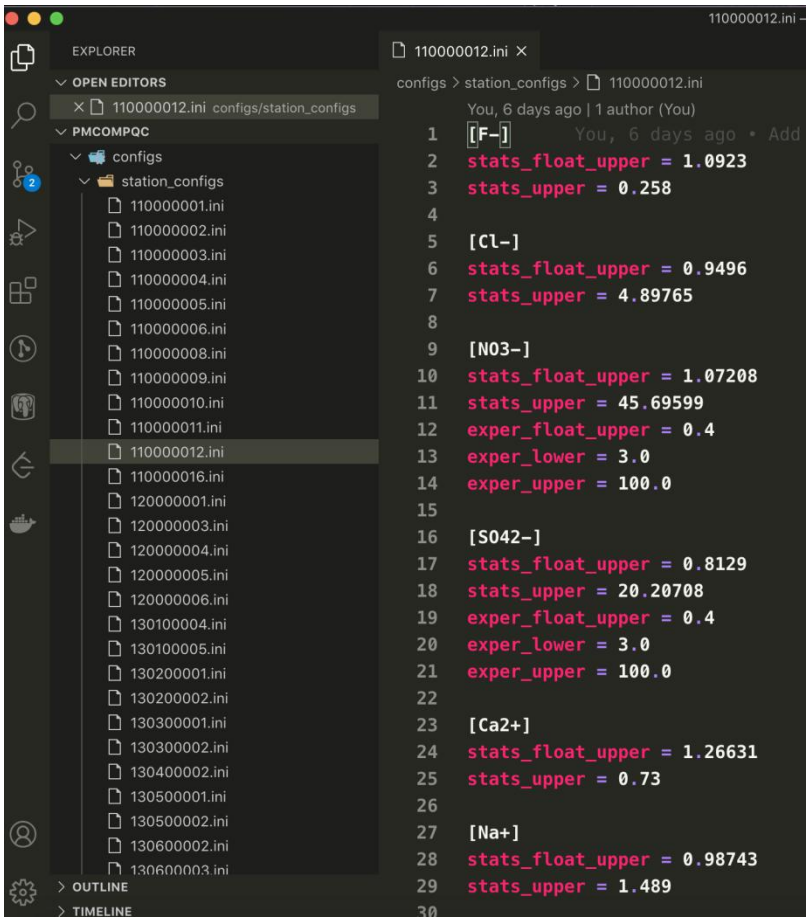
K	Pb	TI	Cu	Cr	Cs	Ag	Zn	Sb	Sn	Mn	Cd	Ni	Ga
8058.000000	7.284000e+03	4337.000000	5153.000000	6559.000000	5429.000000	899.000000	6.299000e+03	1322.000000	966.000000	8078.000000	8.240000e+02	4131.000000	5453.000000
0.497796	4.406314e-02	0.004933	0.116596	0.010367	0.006626	0.202938	1.218175e-01	0.498493	0.495430	0.065093	1.818783e-01	0.012197	0.024479
1.210809	7.599653e-01	0.008169	0.110047	0.010884	0.008536	2.062605	3.046551e-01	5.786266	6.538493	0.061643	2.258709e+00	0.014184	0.679642
0.000013	3.000000e-07	0.000001	0.000003	0.000002	0.000003	0.000081	3.000000e-07	0.000228	0.000183	0.000008	1.000000e-07	0.000009	0.000003
0.311528	2.406150e-02	0.003634	0.093181	0.008154	0.004594	0.037613	8.580900e-02	0.104079	0.068750	0.049734	2.773300e-02	0.009390	0.007453
0.492803	3.904290e-02	0.005457	0.143774	0.012357	0.007515	0.069040	1.392452e-01	0.180949	0.137217	0.073582	5.366380e-02	0.013942	0.010941
0.640739	5.127400e-02	0.006801	0.182180	0.015212	0.009773	0.103242	1.828674e-01	0.297371	0.233156	0.094570	8.237280e-02	0.017132	0.013100
0.947072	7.710630e-02	0.008929	0.243012	0.019841	0.013822	0.261450	2.548322e-01	0.606195	0.464676	0.128080	1.940288e-01	0.021912	0.016280
1.249081	1.051000e-01	0.010982	0.292797	0.024187	0.018495	0.407415	3.321736e-01	1.142203	0.785576	0.167180	2.685156e-01	0.027092	0.018893
2.642618	1.858702e-01	0.031986	0.466431	0.055807	0.036138	0.819943	5.424289e-01	3.489347	2.553145	0.278287	1.013259e+00	0.086378	0.028087
37.105270	6.479658e+01	0.222277	1.772066	0.200741	0.206000	47.340170	2.233501e+01	148.686200	144.247300	1.082875	4.609736e+01	0.154149	29.410720

重金属指标统计

3.2.3 统计指标调用、更新

对于经验指标，其特性结合了专家历史观测经验，对不同组分的数据值融合了多维度的考量，同时其确定的阈值无法适配全国范围不同区域及不同季节的组分观测数据特征，结合各站点历史数据对各地区各时段的不同指标进行计算。

为保障业务化运行时针对不同站点执行不同的指标阈值，在对所有站点各组分的统计时，将每个站点的组分指标同时记录经验阈值和统计阈值并保存在单个站点的配置文件中。质控过程可根据不同站点调用特定配置文件，以实现审核指标的地区和时段的适配；同时可设定时间段对于配置文件的进行更新。



站点阈值文件

3.3 机器学习方法

3.3.1 异常识别

结合颗粒物各组分数据异常特征，以水溶性离子单组分、SNA 组分、水溶性离子全组分、OC、EC、OC+EC 组分分别为特征进行无监督异常识别模型的训练，训练预设 9 个模型，依次为：

ABOD 基于角度的异常值检测，它考虑每个点与其邻近点之间的关系，但不考虑这些邻近点之间的关系，其加权余弦分值与所有相邻分值的方差可视为离群评分。ABOD 在多维数据上表现良好，有两种不同形式的 ABOD：快速 ABOD：使用 k 近邻来估计；原始 ABOD：考虑所有具有高时间复杂性的训练点。

LOF 基于密度检测方法。可量化每个数据点的异常程度，适用中等高维数

据，每一个样本的异常分数称为局部异常因子。局部性由 k 近邻给出，其距离用于估计局部密度。通常使用欧几里得距离将样本的局部密度与其邻居的局部密度进行比较，密度明显低于其邻居的样本。这些被认为是异常值。一个样本点周围的样本点所处位置的平均密度比上该样本点所在位置的密度。比值越接近 1，越可能是正常样本；比值越大于 1，则该点所在位置的密度越小于其周围样本所在位置的密度，这个点就越有可能是异常点。

k-Nearest Neighbors 检测（KNN）对于任何数据点，到其第 k 个最近邻居的距离可以被视为离群评分，常用三个 kNN 检测：最大：使用第 k 个邻居的距离作为离群评分；平均值：使用所有 k 个邻居的平均值作为离群评分；中位数：使用与 k 个邻居距离的中位数作为离群评分。

Feature Bagging, 功能装袋检测器在数据集的各种子样本上预设一些基本检测器。它使用平均或其他组合方法来提高预测精度默认情况下，Local Outlier Factor（LOF）用作基本估算器，其他模型可以用作基本估计器，例如 kNN 和 ABOD；特征装袋首先通过随机选择特征子集来构造 n 个子样本，通过平均或取所有基本检测器的最大值来生成预测分数。

CBLOF 算法是基于聚类组的本地异常因子计算异常值分数。CBLOF 将数据集和由聚类算法生成的聚类模型作为输入。它使用参数 alpha 和 beta 将群集分为小群集和大群集。然后基于该点所属的聚类的大小以及到最近的大聚类的距离来计算异常分数。

HBOS 基于直方图的离群值检测, 假设特征独立并通过构建直方图来计算边远程度。该方法为每一个样本进行异常评分，评分越高越可能是异常点，分别为每个特征作一个直方图，连乘所有特征中该实例密度估计。

PCA 降纬子空间，使用加权投影距离与特征向量超平面的总和作为异常得分。通过降维，将数据映射到低维特征空间，然后在特征空间不同维度上查看每个数据点跟其它数据的偏差；将数据映射到低维特征空间得到 k 个特征向量，再根据这 k 个特征向量从低维特征空间投射回原空间，将重构的数据与原有数据做比较，观察重构误差。

OCSVM 基于密度检测方法, 异常检测中高维数据, 且数据中含有离群点 (异常点），或者对上层数据的分布没有任何假设。当训练数据中包含离群点，模型训练时要匹配训练数据的中心样本，忽视训练样本中的其他异常点。指定要在算

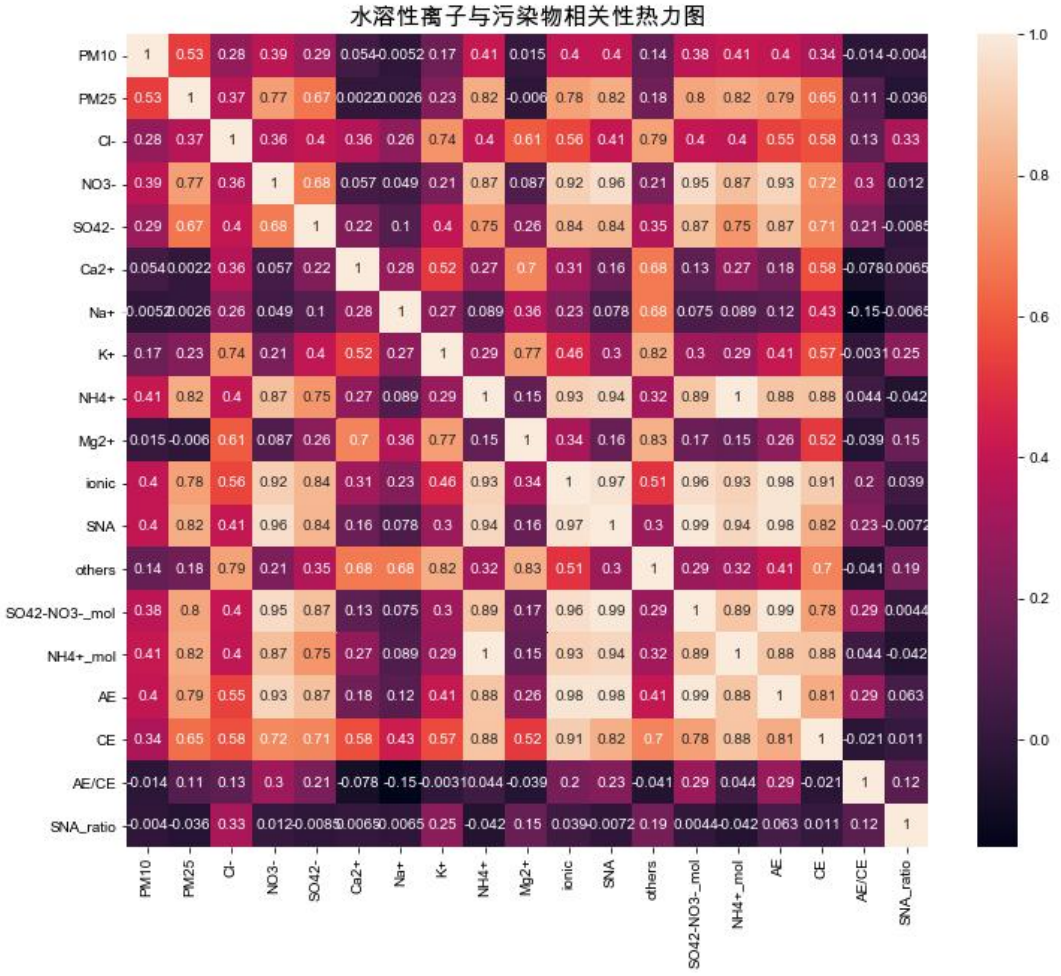
法中使用的内核类型为 Linear，Poly，Rbf，Sigmoid 等。

IForest 孤立森林使用一组树完成数据分区，根据这个数据点在结构中的孤立程度提供异常分数，然后使用异常分数来识别正常观察的异常值，孤立森林在多维数据上表现较好。

在模型实际预测过程，通过对 9 类模型的结果进行投票，判定超过 5 个即为异常情况，以此消除不同模型对于数据的特异性，提高整体模型群的泛化能力。

3.3.2 数据填补

通过统计颗粒物组分历史数据，计算各组分之间的相关性后可发现 PM25～SO42-，PM25～NO3-，PM25～NH4+，PM25～OC，PM25～EC 间的相关性较高，可达 0.7 以上，故通过训练相关回归模型以达到对于缺失、质控剔除数据的填补，水溶性离子组分相关性热力图如下：



基于以上相关性情况，预设 5 类回归模型，其中包括：

Linear Regression 线性回归，线性回归假设目标值与特征之间存在线性相关，即满足一个多元一次方程。通过构建损失函数，来求解损失函数最小时的参数 w 和 b，一般通过最小二乘法或梯度下降法开发求解。

Support Vector Regression 支持向量回归，算法主要是通过升维后，在高维空间中构造线性决策函数来实现线性回归。为适应训练样本集的非线性，传统的拟合方法通常是在线性方程后面加高阶项，增加的可调参数也增加了过拟合的风险。支持向量回归算法采用核函数解决这一矛盾。用核函数代替线性方程中的线性项，引进核函数达到了“升维”的目的，而增加的可调参数使得过拟合依然能控制。

Decision Tree Regression 决策树回归，回归树将特征空间划分成若干单元，每一个划分单元有一个特定的输出，因为每个结点都是“是”和“否”的判断，划分的边界是平行于坐标轴的，对于测试数据只要按照特征将其归到某个单元，便得到对应的输出值。

Random Forest Regression 随机森林回归，利用多棵树对样本进行训练并预测的一种集成模型，处理回归问题时，则以每棵决策树输出的均值为最终结果。

Bagging Regression 集成回归，让算法训练多轮，每轮的训练集由从初始的训练集中随机取出的 n 个训练样本组成，某个初始训练样本在某轮训练集中可以出现多次或根本不出现，训练之后可得到一个预测函数序列 H = {h1.....hn}，最终的预测函数 H 对回归问题采用简单平均方法对结果进行计算。

3.3.3 功能集成

颗粒物各组分各项审核指标的统计阈值计算集成在指定脚本，独立进行质控站点全部组分的各项指标统计更新，用于审核指标配置文件的定期更新。

异常式别与数据填补模型的训练、保存、加载、预测过程经过线下测试，集成在对应模型训练的脚本内，模型加载预测集成在质控流程内；训练与预测过程独立解耦，用于各模型的定期更新。

3.4 数据质控

3.4.1 质控过程

质控流程从不同质控方法角度出发，有缺失记录、统计方法质控、异常模型识别质控、数据填补质控几个阶段：

1）缺失记录，数据由质控时次文本文件读入，对于-999 的缺失值进行质控日志记录，同时修改为 nan 便于后面处理；

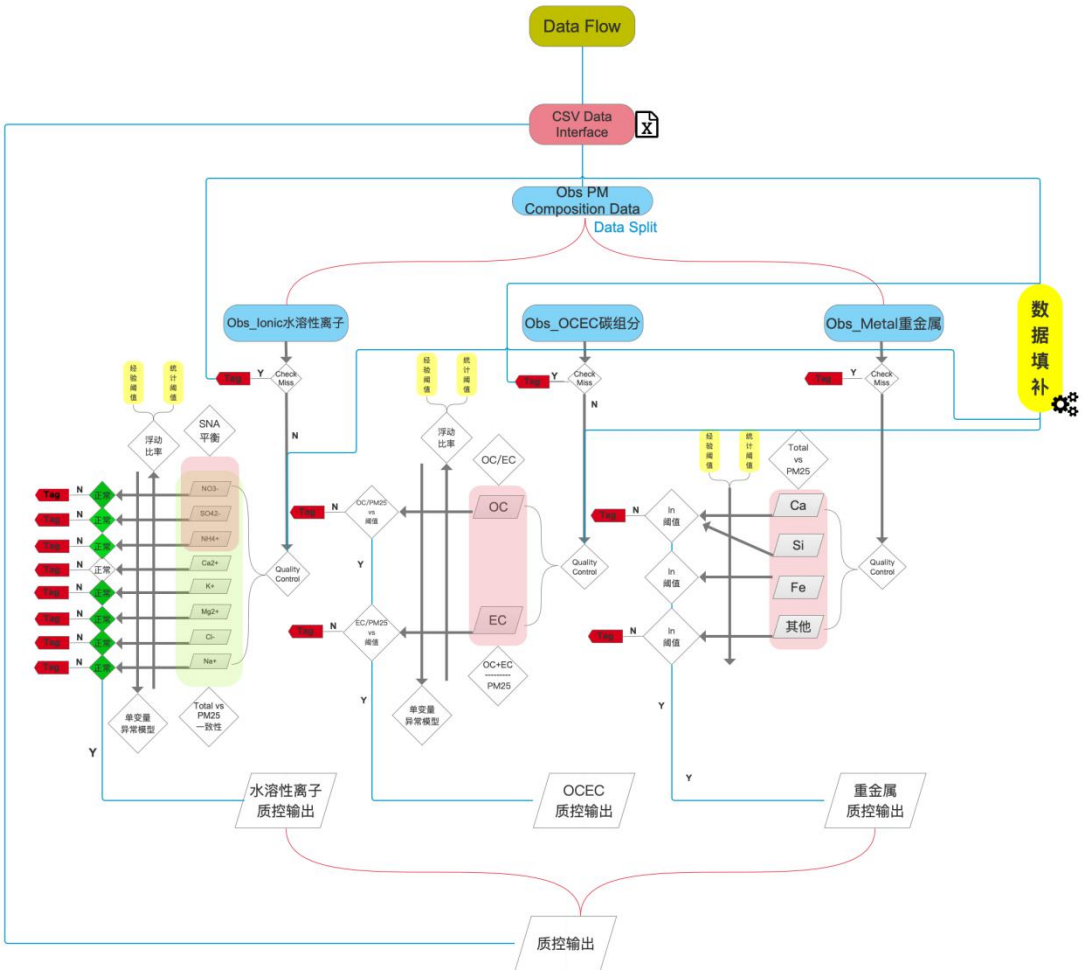
2）统计质控，结合总站审核规划指标，对质控时次及相关前后时次数据进行对应指标计算，利用质控指标进行第一阶段质控审核，对于审核的数据进行异常记录，并剔除送入下一阶段质控；

3）异常识别，加载异常识别模型群，通过不同模型的异常判定结果进行综合判定，对判定异常的数据进行记录与剔除，保存为质控剔除阶段数据文件，进行下一阶段质控；

4）数据填补，针对当时次由于原始缺失或质控剔除而形成缺失的 SNA 与 OCEC 数据进行填补，基于 PM25 数据，加载预训练的回归模型，进行对应缺失的预测填补，保存为质控填补阶段数据文件。

3.4.2 数据流程

任务总体框架的数据流程，分为输入接口数据获取，质控流程遵循以上过程，输出接口数据和日志文件写入等主要过程，具体形式入下图：



4. 项目测试

4.1 环境配置

代码及环境配置参考地址：http://47.92.132.84:3000/buzh/PmCompQC，配置好运行环境后进行案例的具体配置。

4.2 测试案例

4.2.1 公共配置

修改项目根目录下 base_cfg.py 内各模块统一公共的信息，主要是项目目录。

4.2.2 提取数据

通过格式化拼接 SQL，从数据库提取每小时观测站点的所有组分数据；通过提取每小时站点数据拼接所有站点的一段时间的组分观测数据。

base_cfg.py - 为上层目录 base_cfg.py 软连接；

cfg.py - 配置脚本；

大部分内容为静态配置

修改提取时间段的起止时间

paths 为相对目录，一般不需要修改 也可按照喜好修改目录名称

utils_workflow.py - 工具脚本 包括其他脚本需要的功能函数；

main_db.py - 数据库提取主脚本 ；

根据配置文件的数据库、站点、时间、路径信息，从数据库提取单小时观测数据，文件名称格式为：obs_com_${yyyy}${mm}${dd}${HH}.txt

```
[buzh@node1 PmCompQC]$ cd extract/
[buzh@node1 extract]$ ll
total 20
lrwxrwxrwx 1 buzh buzh   14 Jul  2 09:57 base_cfg.py -> ../base_cfg.py
-rw-rw-r-- 1 buzh buzh 2749 Jul  2 09:57 cfg.py
-rw-rw-r-- 1 buzh buzh 6904 Jul  2 09:57 main_db.py
-rw-rw-r-- 1 buzh buzh 2047 Jul  2 09:57 main_extract.py
-rw-rw-r-- 1 buzh buzh 3136 Jul  2 09:57 utils.py
[buzh@node1 extract]$ /public/home/buzh/miniconda3/bin/python main_db.py
processing 2019-02-12 00 ...
processing 2019-02-12 01 ...
processing 2019-02-12 02 ...
processing 2019-02-12 03 ...
...
```

main_extract.py - 数据抽取；

从单小时观测数据中提取单一站点起止时间内所有组分数据到独立文件，文件名称格式为 ${站点编号}.txt

```
[buzh@node1 extract]$ ll
total 28
lrwxrwxrwx 1 buzh buzh   14 Jul  2 09:57 base_cfg.py -> ../base_cfg.py
-rw-rw-r-- 1 buzh buzh 2749 Jul  2 09:57 cfg.py
-rw-rw-r-- 1 buzh buzh 6904 Jul  2 09:57 main_db.py
-rw-rw-r-- 1 buzh buzh 2079 Jul  2 10:05 main_extract.py
drwxrwxr-x 2 buzh buzh 4096 Jul  2 10:00 obs_hourly_data
drwxrwxr-x 2 buzh buzh 4096 Jul  2 09:59 __pycache__
-rw-rw-r-- 1 buzh buzh 3136 Jul  2 09:57 utils.py
[buzh@node1 extract]$ /public/home/buzh/miniconda3/bin/python main_extract.py
handling station 1: 110000006
handling station 2: 371500001
handling station 3: 150100001
handling station 4: 1320100001
handling station 5: 410200001
...
```

4.2.3 填补数据

目前数据库提取的组分观测数据中颗粒物 PM 数据大量缺失，设计匹配临近

污染观测站点相同时间的 PM10、PM25 数据作为补充，其数据依赖于数据库提取的原始组分观测数据和匹配的临近污染观测站 PM 观测数据。

通过匹配生成 input 目录下 2019 年质控输入数据，以及 data 目录用于模型训练的数据集。此功能目前是针对 2019 年全年数据量，咱不支持指定区间段的填补，目前仅支持 2019 年全年数据的测试。

base_cfg.py - 为上层目录 base_cfg.py 软连接；

cfg.py - 本模块配置文件；

fill_obs_pm.py - 提供小时分辨率文件 obs_comp_${yyyymmddHH}.txt 添加 PM 数据；

fill_sta_pm.py - 提供单站点文件 ${站点编号}.txt 添加 PM 数据；

nearby.json - 组分观测站点与污染观测站点匹配文件 {k: v} = {组分观测站点编号: 污染观测站点编号}；

nearby.txt - 匹配后的污染观测站点信息文件；

4.2.4 数据质控

包括数据质控、统计阈值计算写入配置文件、异常识别模型训练、数据填补模型训练等功能，其数据依赖于数据库提取、PM 颗粒物填补后的各类数据。

base_cfg.py - 为上层目录 base_cfg.py 软连接；

cfg.py - 本模块配置文件；

无监督异常识别模型配置

回归补数模型配置

计算统计指标统计上线分位数

质控起止时间

相关路径

calc_qc_index.py - 统计阈值脚本；

统计各站点历史数据各项指标并写入各站点的阈值文件，保存至 configs 目录；执行计算统计，则更新 configs 目录下各站点审核指标.ini 文件中各类组分各项指标。

```
[buzh@node1 src]$ pwd
/public/home/buzh/PmCompQC/src
[buzh@node1 src]$ /public/home/buzh/miniconda3/bin/python calc_qc_index.py
*****************************************************************************************************
1       Calc 371400002 ...
      PM10  PM25  F-  Cl-  NO3-  SO42-  Ca2+  ...  Ionic_ratio  SO42-NO3-_mol  NH4+_mol  SNA_balance  AE  CE
4068  78.0  36.0 NaN  NaN   NaN    NaN   NaN  ...          0.0            NaN       NaN          NaN NaN NaN

[1 rows x 19 columns]
-----------------------------------------------------------------------------------------------------
       PM10  PM25  OC  EC  OC/PM25  EC/PM25  OCEC/PM25  OC/EC
3430  107.0  18.0 NaN NaN      NaN      NaN        NaN    NaN
*****************************************************************************************************
2       Calc 131100002 ...
...
```

qc_model_train.py - 训练无监督学习模型脚本；

模型自动保存至 models/qc_models 目录；执行训练操作，读取配置中无监督学习配置内容，训练指定特征的无监督异常判别模型，并按照特征、模型名称保存。

```
[buzh@node1 src]$ pwd
/public/home/buzh/PmCompQC/src
[buzh@node1 src]$ /public/home/buzh/miniconda3/bin/python qc_model_train.py
```

fill_model_train.py - 训练回归补数模型脚本；

模型自动保存至 models/fill_models 目录；执行训练操作，读取配置中监督回归学习配置内容，训练指定特征的回归补数机器学习模型，并按照特征、模型名称保存。

```
[buzh@node1 src]$ pwd
/public/home/buzh/PmCompQC/src
[buzh@node1 src]$ /public/home/buzh/miniconda3/bin/python fill_model_train.py
```

main.py - 质控主脚本；

质控输入 input

质控输出 output

质控数据输出 output/qc_data

质控剔除 qc_${yyyymmddHH}.txt

质控填补 fill_${yyyymmddHH}.txt

质控日志输出 output/qc_logs

质控日志 qc_${yyyymmddHH}.json

执行质控操作，系统自动针对配置时间段内数据进行质控操作，此时 input 为链接目录，为避免出现同时读取可复制 input 到项目所在目录；同时需注意由于质控过程考虑了前后时次数据，其时间起止设置与之前时间内会存在差别。

```
[buzh@node1 src]$ pwd
/public/home/buzh/PmCompQC/src
[buzh@node1 src]$ /public/home/buzh/miniconda3/bin/python main.py>&qc.test.log&
```

utils - 工具；

mylog - 日志记录；

dao - data access object 数据读取；

controllers - 质控控制类；

Controller.py - 质控基类

IonicController.py - 水溶性离子质控类

OcecController.py - OCEC 质控类

MetalController.py - 重金属质控类

4.2.5 后处理提数

分别提取质控站点在质控时间段内的观测原始、质控剔除、质控填补的数据，其输入数据依赖于质控过程的输入输出数据，形成三类数据用于绘图分析一段时间的质控效果。

base_cfg.py - 为上层目录 base_cfg.py 软连接；

cfg.py - 本模块配置文件；

配置组分观测站点信息文件目录

配置提取 SRC 项，如 obs qc fill

main_extract.py - 数据提取脚本；

数据提取需分别配置，运行三次提取到质控站点质控期间 obs qc fill 三类数据集。

```
[buzh@node1 post_analysis]$ pwd
/public/home/buzh/PmCompQC/post_analysis
[buzh@node1 post_analysis]$ vim cfg.phy
[buzh@node1 post_analysis]$ /public/home/buzh/miniconda3/bin/python main_extract.py
handling station 1: 110000006
handling station 2: 371500001
handling station 3: 150100001
handling station 4: 1320100001
handling station 5: 410200001
handling station 6: 140100002
...
```

4.2.6 后处理画图

质控可视化输入数据依赖于 post_analysis 模块提取质控时间段内的三类数

据。

base_cfg.py - 为上层目录 base_cfg.py 软连接；

utils.py - 工具；

cfg.py - 本模块配置文件；

配置组分观测站点信息文件目录

配置 obs、qc、fill 三类数据目录

配置图片输出模块

sequence_plot.py - 绘制质控对比图；

```
[buzh@node1 visualization]$ pwd
/public/home/buzh/PmCompQC/visualization
[buzh@node1 visualization]$ /public/home/buzh/miniconda3/bin/python  sequence_plot.py
2019-02-12 02:00:00 2019-02-13 21:00:00
Plot 1: 110000006
Plot 2: 371500001
Plot 3: 150100001
Plot 4: 1320100001
...
...
```

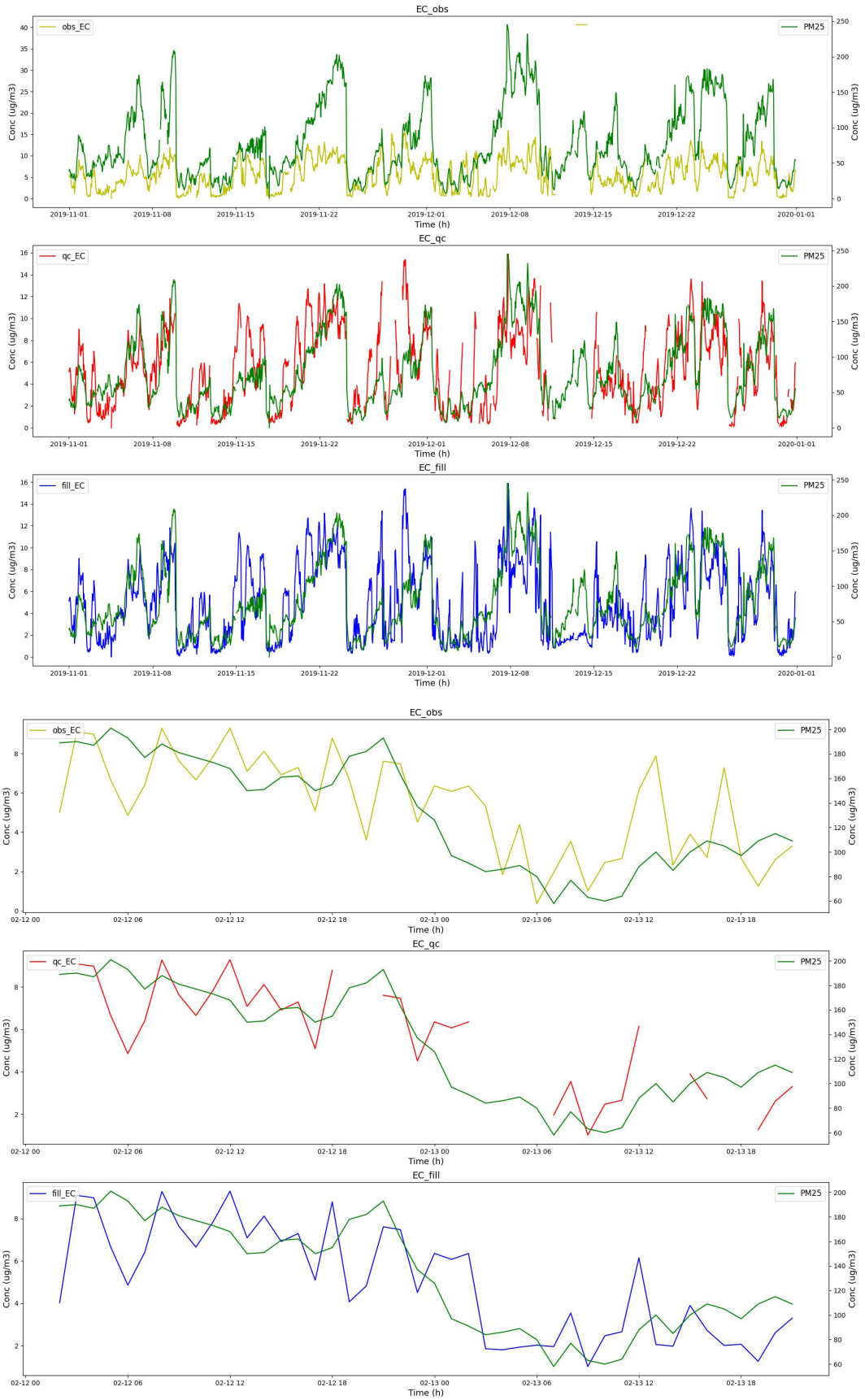
4.3 输出结果

质控输出数据均为文本文件，可直接查看；质控日志为 json 文件，可直接查看也可以用文本编辑器查看，质控输出目录结构如下：

```
[buzh@node1 output]$ pwd
/public/home/buzh/PmCompQC/output
[buzh@node1 output]$ tree -L 2
.
├── qc_data
    ├── fill_obs_2019021201.txt
    ├── fill_obs_2019021202.txt
    ├── fill_obs_2019021203.txt
    ...
    ├── qc_obs_2019021201.txt
    ├── qc_obs_2019021202.txt
    ├── qc_obs_2019021203.txt
    ...

└── qc_logs
    ├── qc_2019021201.json
    ├── qc_2019021202.json
    ├── qc_2019021203.json
    ...
```

绘制质控效果对比图，输出主要五种组分数据内容：


EC_obs
obs_EC
PM25
Conc (ug/m3)
Time (h)
EC_qc
qc_EC
EC_fill
fill_EC
2019-11-01
2019-11-08
2019-11-15
2019-11-22
2019-12-01
2019-12-08
2019-12-15
2019-12-22
2020-01-01
02-12 00
02-12 06
02-12 12
02-12 18
02-13 00
02-13 06
02-13 12
02-13 18

质控过程效果对比图