ICS 35.240.01
CCS L 70

团　　体　　标　　准

T/ACEF □□□－20□□

大数据优化区域空气质量模拟排放输入数据技术规范

Technical specification for using big data optimizate emission input data in regional air quality modeling

（征求意见稿）

20□□-□□-□□发布　　20□□-□□-□□实施

中 华 环 保 联 合 会 发布

目　次

前　言

本文件按照GB/T1.1-2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由生态环境部环境规划院提出。

本文件由中华环保联合会归口。

本文件起草单位：生态环境部环境规划院、北京市生态环境监测中心。

本文件主要起草人：赵大地、王建童、卢亚灵、李勃、王莉华。

大数据优化区域空气质量数值模拟排放输入数据技术规范

1　范围

本文件规定了大数据优化区域空气质量数值模拟排放输入数据方法的基本要求、技术要求、技术方法。

本文件适用于指导使用大数据方法优化开展区域空气质量数值模拟所需的排放输入数据的实施与应用。

2　规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

本文件没有需要界定的术语和定义。

3　术语和定义

下列术语和定义适用于本文件。

3.1

大数据　big data

具有体量巨大、来源多样、生成极快且多变等特征并且难以用传统数据体系结构有效处理的包含大量数据集的数据。

[来源：GB/T 35295-2017，　2.1.46]

3.2

区域空气质量模拟　regional air quality modeling

基于人类对大气物理和化学过程的科学认知，以整个大气圈层为整体，运用气象学原理和数学方法，从水平和垂直方向上对几十公里到几千公里范围区域内大气中的多种污染物之

间的复杂物理、化学反应过程开展仿真模拟，再现污染物在大气中输送、反应、清除等过程。

3.3

排放输入数据　emission input data

开展区域空气质量模拟时需输入的污染物排放量数据，需精确到小时、空间网格及每个化学物种的具体排放量。

3.4

时间序列数据　time series data

在多个时间点观察或测量的并按照时间排列的一组数据。

3.5

空间分配　spatial allocation

污染源的地理位置或者采用与污染源有相同空间变化特征的空间地理信息数据，利用基于地理信息系统（GIS）的计算机技术，将污染物排放量分配到网格化地图中，以表征污染源排放空间特征。

3.6

时间分配 time allocation

根据污染源在某段时间的污染物排放总量，基于污染源活动水平的时间变化特征数据将逐步分解到小时尺度排放量，以反映污染物排放的时间变化特征。

3.7

固定污染源　stationary environmental pollution source

由同一排污单位组织管理，位于相同或相邻位置的产生、处理或排放污染物的生产设施、污染治理设施和排放口等的集合。

3.8

移动污染源　moving environmental pollution source

空间位置随时间变化而变化的污染源，主要包括汽车、飞机、船舶、非道路移动机械等交通工具。

4　基本要求

开展排放数据优化工作所选取的大数据，应能够满足以下基本要求：

a)　应能够区分不同污染源；

b)　应能够反映污染源活动水平或支持识别与污染物排放量的关系；

c)　应能够识别污染源排放的时间变化特征；

d)　应能够反映污染源排放的空间位置信息。

5　技术要求

5.1　数据采集

5.1.1　数据采集方式

数据采集活动的目标是获得能够反映污染源污染物排放水平的数据，数据采集方式包括但不限于：

a)　网络数据采集。通过网络爬虫或公开 API 等方式获取数据。

b)　通过传感器获取。传感器包括温度传感器、电视、汽车、摄像头等公共和个人的智能设备。

c)　系统数据。组织内部的系统在运行过程中采集和产生的业务数据，以及各种系统、程序和服务运行产生的大量运维和日志数据等。

d)　从其他来源获取。通过线上或线下等方式从其他来源获取数据。

数据采集活动主要操作包括但不限于：发现数据源、传输数据、生成数据、缓存数据、创建元数据、数据转换、数据完整性验证等。

5.1.2　数据安全要求

开展数据采集活动时，应：

a) 选取适当的数据来源，明确数据采集范围和用途；

b) 遵循合规原则，确保数据采集的合法性、正当性和必要性；

c) 遵循数据最小化原则，只采集满足业务所需的最少数据；

d) 遵循质量保障原则，制定数据质量保障的策略、规程和要求；

e) 遵循确保安全原则，对采集的数据进行分类分级标识，并对不同类和级别的数据实施相应的安全管理策略和保障措施。对数据采集环境、设施和技术采取必要的安全管控措施。

5.2 数据存储要求

数据存储应满足以下要求：

a) 应支持通用存储组件；

b) 应支持通用数据类型；

c) 应支持行存储或列存储；

d) 应支持数据压缩；

e) 应支持数据索引；

f) 应支持冗余存储策略；

g) 应支持安全备份策略。

5.3 数据计算要求

数据计算能力满足以下要求：

a) 应支持海量数据批量计算；

b) 应支持流数据处理；

c) 应支持分析查询引擎；

d) 应支持常用的 SQL 分析功能（如统计、连接、任意指标维度的下钻）；

e) 应支持机器学习算法。

5.4 数据处理要求

完成大数据采取后，应开展数据清洗、数据整合、数据处理等工作：

a) 应开展数据清洗，剔除无效数据并对数据中的异常值进行分析处理；

b) 应开展数据整合，统一不同来源、格式、类型数据格式、单位、编码等，确保数据

的一致性；

c) 应开展数据归一化、标准化、离散化等处理，以适应不同的数据处理需求；

d) 须针对敏感数据进行加密处理，以保障数据安全；

e) 应定期对数据进行更新和维护，保证数据的时效性。

6 技术方法

6.1 方法原理

识别已获取的大数据与污染源污染物排放数据间的关系，量化污染源近实时的污染物排放量数据。

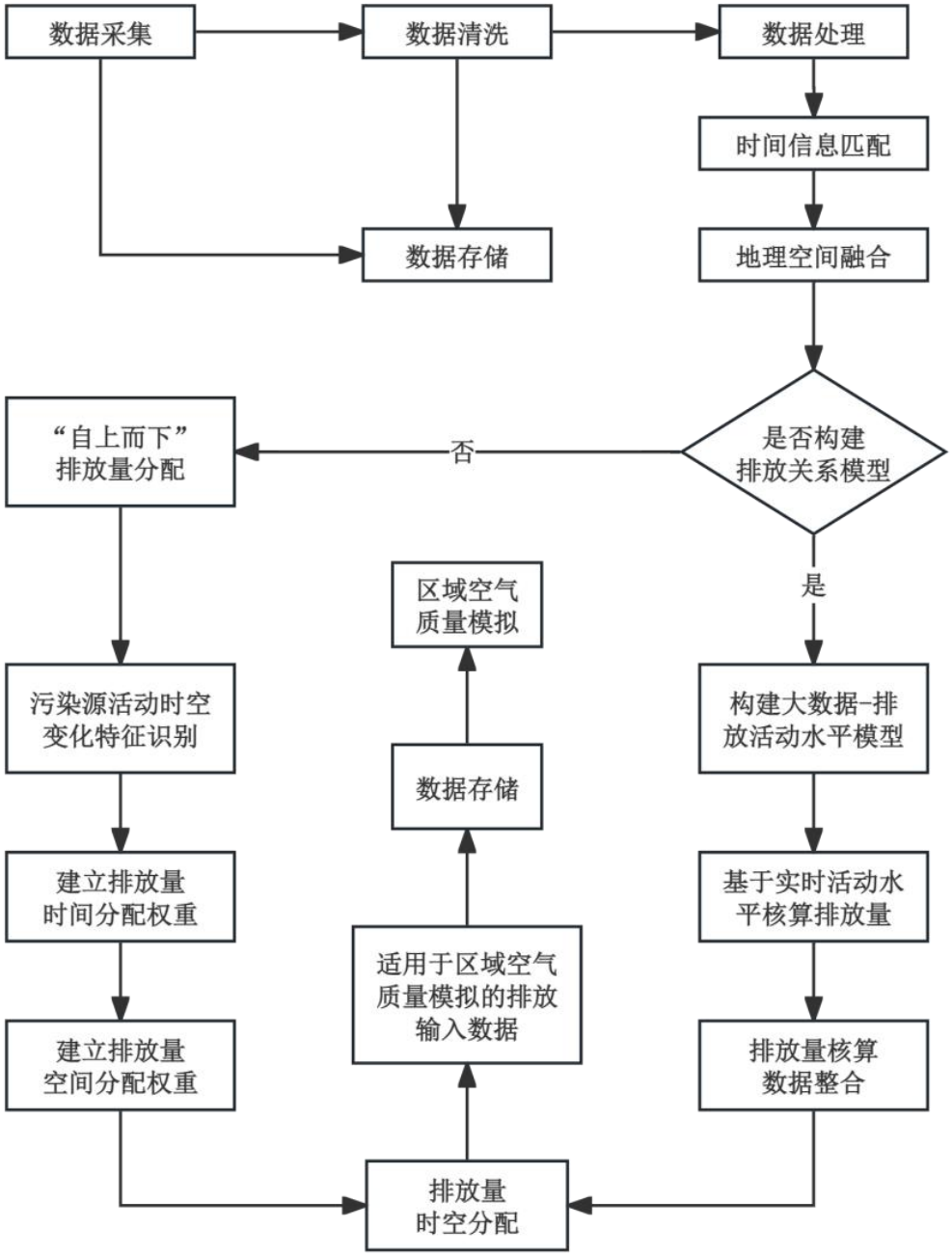

图 1 大数据优化排放输入数据流程

6.2 排放关系建模法

基于获取到的大数据分析发掘污染源排放的时空变化规律，构建大数据与污染源活动水平和污染物排放量的关系模型，量化小时尺度下的排放量信息，得到实时动态变化的高时空分辨率排放清单。具体步骤如下：

a) 按照污染源类型、排放单位初步筛分大数据，分类处理数据；

b) 根据污染源文本字段进行时间信息匹配，形成具有时间序列的污染源大数据；

c) 根据污染源文本字段进行地理空间融合，赋予时间序列数据空间位置信息；

d) 构建大数据-污染源排放活动水平模型，识别所采集数据与污染源排放活动关系，模型构建形式应可表达为：

$$P_i = f(x_i) \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots（1）$$

式中：

P_i——为可表征污染源排放特征的活动水平数据；

i——为污染物类型；

x_i——为获取到的和污染源相关的大数据。

e) 构建污染源活动水平－排放量关系模型，匹配污染源活动水平与排放因子数据，核算污染源在不同时间段、不同地点（经纬度、路段、区域等）的污染物排放量，排放量核算计算公式可表达为：

$$E_i = P_i \times EF_i \times (1 - \eta) \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots（2）$$

式中：

EF_i——为污染源污染物排放系数；

i——为污染物类型；

P_i——为污染源排放活动水平数据；

η——为污染源治污设施的污染物去除效率，若无治污设施，则 η 设为 0。

f) 集合区域内所有污染源，形成具有时间序列和空间信息的污染源污染物排放数据集；

g) 按照数据集时间和空间信息，分配到数值模拟所需的小时尺度和空间网格尺度，形成高时空分辨率的大气污染物排放清单数据；

h) 将排放清单数据转换为适用于空气质量数值模拟的数据输入格式。

6.3 优化时空分配系数法

基于获取到的大数据，通过优化时空分配参数实现对已有的长时间尺度、大空间范围排放清单的精细化准确分配。具体步骤如下：

a) 按照污染源类型、排放单位初步筛分大数据，分类处理数据；

b) 根据污染源文本字段进行时间信息匹配，形成具有时间序列的污染源大数据；

c) 根据污染源文本字段进行地理空间融合，赋予时间序列数据空间位置信息；

d) 按照大数据变化特征，量化污染源活动水平数据的时间变化趋势和空间分布特征；

e) 基于污染源活动水平时间变化趋势，建立排放量时间（年、月、日、小时）分配权重；

f) 基于污染源空间信息（如经纬度、路网、所在地块位置等），与数值模拟拟采用的空间网格进行空间融合；针对固定污染源，其空间分配权重可设置为所在网格为 1，其他网格为 0；对于移动污染源，基于移动路线（如道路、铁路、河道等）所在网格，统计大数据中污染源在不同点位出现频次，以频率计算相关网格分配比例（路线途径网格权重总和应为 1，分路线途径的其他网格权重应均为 0）；

g) 对已有的污染源排放清单数据进行时间与空间分配，形成高时空分辨率排放清单数据；

时间分配的具体方法见下方公式：

$$E_m = E_y \times M \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots（3）$$

$$E_d = E_m \times D \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots（4）$$

$$E_h = E_d \times H \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots（5）$$

式中：

m、d、h——分别代表月、天和小时；

E——污染物排放量；

M、D、H——分别为当月、天、小时的时间分配权重；

排放空间分配具体方法见下方公式：

$$Q_g = DW_g \times E \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots（6）$$

式中：

Q_g 为 g 网格污染物排放量；

g 为网格编号；

DW_g 为 g 网格的空间分配比例；

E 为污染物总排放量。

h) 将排放清单数据转换为适用于空气质量数值模拟的数据输入格式。

参 考 文 献

[1] GB/T 35295 信息技术 大数据 术语

[2] GB/T 38675 信息技术 大数据计算系统通用要求

[3] GB/T 41818 信息技术 大数据 面向分析的数据存储与检索技术要求

[4] GB/T 42201 智能制造 工业大数据时间序列数据采集与存储管理

[5] GB/T 42528 时空大数据技术规范

[6] GB/T 44216 信息技术 大数据 批流融合计算技术要求

[7] HJ/T 416 环境信息术语

[8] HJ 608 排污单位编码规则