

DOI: 10.11949/0438-1157.

面向炼化生产运行的高质量多源时序数据集构建方法研究

刘晋瑞¹, 刘雪鹏¹, 刘速¹, 吕庭龙¹, 戴赞², 王振雄³

(¹ 中石油(北京)数智研究院有限公司, 北京 102206; ² 中国石油天然气股份有限公司长庆石化分公司, 陕西 咸阳 712042; ³ 中国石油天然气股份有限公司兰州石化分公司, 甘肃 兰州 730060)

摘要: 炼化生产过程中的 DCS、LIMS 与 MES 数据是支撑时序大模型训练与智能化应用的关键前提。针对此类数据普遍存在的结构不一致、质量参差、时间错配和语义关系割裂等问题, 本文提出一套面向炼化生产运行的高质量多源时序数据集构建方法。该方法涵盖数据采集、存储、标准化、质量评估与处理分析五个核心环节: 首先, 基于分层分类目录结构实现多源数据的高效组织; 其次, 定义 DCS、LIMS 和 MES 数据统一时间戳与变量标识的标准化格式; 进而, 构建包含完整性、规范性、合理性与稳定性四个维度的可量化质量评估体系; 最后, 采用时间对齐与语义映射策略, 有效解决不同来源数据在时间粒度与语义表达上的差异问题。本研究通过所提方法, 构建了一套高质量炼化生产运行时序数据集, 实现数据集的结构统一、质量可控和语义协同, 为炼化行业数据资产化与智能化升级提供了可复用的数据构建范式。

关键词: 石油化工; 生产运行; 过程系统; 高质量时序数据集

中图分类号: TQ 056.1; TP 391

文献标志码: A

文章编号: 0438-1157 (XXXX) XX-0001-11

Research on construction methods for high-quality multi-source time series datasets for refining and chemical production operations

LIU Jinrui¹, LIU Xuepeng¹, LIU Su¹, LYU Tinglong¹, DAI Yun², WANG Zhenxiong³

(¹ PetroChina (Beijing) Digital Intelligent Research Institute Co., Ltd., Beijing 102206, China; ² PetroChina Changqing Petrochemical Company, Xianyang 712042, Shaanxi, China; ³ PetroChina Lanzhou Petrochemical Company, Lanzhou 730060, Gansu, China)

Abstract: The DCS(Distributed Control System), LIMS(Laboratory Information Management System), and MES (Manufacturing Execution System) data in the refining and chemical production process are crucial prerequisites for supporting the training of large time-series models and intelligent applications. Addressing the prevalent issues in such data, such as inconsistent structures, varying quality, time mismatches, and fragmented semantic relationships, this paper proposes a method for constructing a high-quality multi-source time-series dataset tailored for refining and chemical production operations. This method encompasses five core aspects: data collection, storage, standardization, quality assessment, and processing analysis. Firstly, it achieves efficient organization of multi-source data based on a hierarchical classification directory structure. Secondly, it defines standardized formats for unified timestamps and variable identifiers for DCS, LIMS, and MES data. Furthermore, it constructs a quantifiable

收稿日期: 2026-03-27 修回日期: 2026-04-23

通信作者: 刘雪鹏(1983—),男,博士,高级工程师,liuxuepeng01@petrochina.com.cn

第一作者: 刘晋瑞(1999—),女,硕士,工程师,liu_jinrui@petrochina.com.cn

基金项目: 中国石油天然气股份有限公司科技项目(2025ZG86)

引用本文: 刘晋瑞, 刘雪鹏, 刘速, 吕庭龙, 戴赞, 王振雄. 面向炼化生产运行的高质量多源时序数据集构建方法研究[J]. 化工学报, XXXX, XX (XX): 1-11

Citation: LIU Jinrui, LIU Xuepeng, LIU Su, LYU Tinglong, DAI Yun, WANG Zhenxiong. Research on construction methods for high-quality multi-source time series datasets for refining and chemical production operations[J]. CIESC Journal, XXXX, XX(X): 1-11

quality assessment system encompassing four dimensions: completeness, normalization, rationality, and stability. Lastly, it employs time alignment and semantic mapping strategies to effectively address the differences in time granularity and semantic expression among data from different sources. Through the proposed method, this study constructs a high-quality time-series dataset for refining and chemical production operations, achieving unified structure, controllable quality, and semantic coordination of the dataset. This provides a reusable data construction paradigm for the data assetization and intelligent upgrading of the refining and chemical industry.

Keywords: petrochemical industry; production operation; process system; high-quality time-series dataset

引 言

在国家大力推进“数字中国”与“智能制造”的战略背景下,数据作为新型生产要素的核心价值日益凸显。《“数据要素×”三年行动计划(2024—2026年)》明确提出,要推动工业领域数据资源深度开发利用,提升数据赋能产业转型升级的能力[1]。高质量、标准化的数据集是人工智能大模型、智能体平台等前沿技术在工业场景落地的前提与基石[2-4]。针对炼化领域中的复杂工艺流程,构建多来源、高质量的时序数据集,已成为实现数据驱动智能化业务场景构建的关键支撑[5-7]。

炼化生产运行数据主要来源于分布式控制系统(Distributed Control System, DCS)、实验室信息管理系统(Laboratory Information Management System, LIMS)和制造执行系统(Manufacturing Execution System, MES)[8-10]。其中,DCS数据反映各生产装置实时动态,可采集秒级的连续过程变量(如温度、压力、液位、流量);LIMS数据则以小时或班次为周期,提供原料、中间产品及成品的关键物性分析数据(如组分含量、油品馏程),其采样频率低且时延较高;MES数据包含装置实时变量、能耗统计、操作指令等信息。三者和时间尺度、数据结构与表示含义上高度互补,若能有效融合,将为工艺操作优化、质量预测、能效优化与异常诊断等智能化任务提供全面的数据支撑。炼化生产运行数据集的构建存在三大困难,一是数据来源与格式的多样性,国内各炼化企业各装置所使用的DCS系统不同,不同来源的DCS系统导出解析后的数据格式不一致,LIMS数据和MES数据也因各炼化企业使用的系统原因存在不同的标准规范,导致数据形态不一。二是此类炼化时序数据基本为秒级或分钟级采样数据,数据质量参差不齐,缺乏质量评价标准,导致数据处理难度大。三是不同来源数据之间存在时间错配和语义差异问题,导致后续建模存在数据使用

难度。

近年来,相关研究在工业数据集建设方面取得显著进展,工业数据治理与数据集建设领域已形成通用治理框架[11-14]。周军等人面向测井数据,建设统一测井数据库,采用分布式存储架构实现各类数据、模型、资料等的存储管理和一体化应用[15]。李梁等人通过研究提出新型数据治理方法论,以治理组织和数智中台为支撑,设计多模态数据标准体系,构建AI场景下的数据治理体系和实施策略[16]。刘珊珊等人研究阐述了大语言模型在提升数据质量和数据治理的价值和潜力[17]。Shu等人面向大型石化工厂的工业物联网数据集构建与管理挑战,重点探讨了异构无线传感网络融合、众包感知机制及中间件架构在数据采集、整合与实时处理中的关键作用[18]。上述成果为工业数据治理提供了通用范式与技术参考,但面向炼化生产场景的时序数据集构建仍存在空白和短板:一是通用治理框架侧重管理规范与组织流程,对炼化时序数据的工艺特性等适配不足;二是现有工业数据集建设多聚焦存储整合或单源治理,缺少覆盖“采集-存储-标准化-评估-处理与分析”全链路、可量化质控、可工程落地的炼化时序数据集专属构建框架,难以支撑高质量、标准化、可复用的炼化领域数据基座建设。

因此,本文聚焦炼化生产运行高质量时序数据集构建工作的标准化、全流程范式,提出高质量炼化时序数据集构建方法,支撑建设多源数据格式统一、质量标准高、时间和语义统一匹配的大规模炼化生产运行时序数据集。研究重点包括:设计分层分类的数据组织架构,构建各类数据标准化格式,建立覆盖完整性、准确性、一致性与时效性的数据质量评估体系,并提出基于时间对齐与语义映射的多源融合策略。在此基础上,进一步探讨所构建数据集的智能化应用路径,旨在为炼化行业数据资产化与智能化升级提供可复用的技术范式与实践参考。

1 多源炼化生产运行数据集构建方法

本文提出一套高质量多源炼化生产运行时序数据集构建方法,流程如图1所示,包括五个核心环节:数据采集、数据存储、数据标准化、数据质量评估、数据处理与分析。首先,通过对接DCS、LIMS与MES等工业信息系统,实现多源异构数据的采集与初步整理和解析;其次,基于分层分类的目录结构

完成高效存储与组织;进而,按照统一标准化的数据格式结构,进行各类型数据标准化处理;在此基础上,构建涵盖完整性、规范性、合理性与稳定性的质量评估量化体系,支撑数据高质量判定;随后,采用时间对齐与语义映射相结合的策略开展多源融合的数据处理与分析。该方法体系兼顾技术可行性与工业落地性,为高质量炼化时序数据集构建提出一套系统方法架构,进而为炼化智能化应用提供坚实的数据基座。

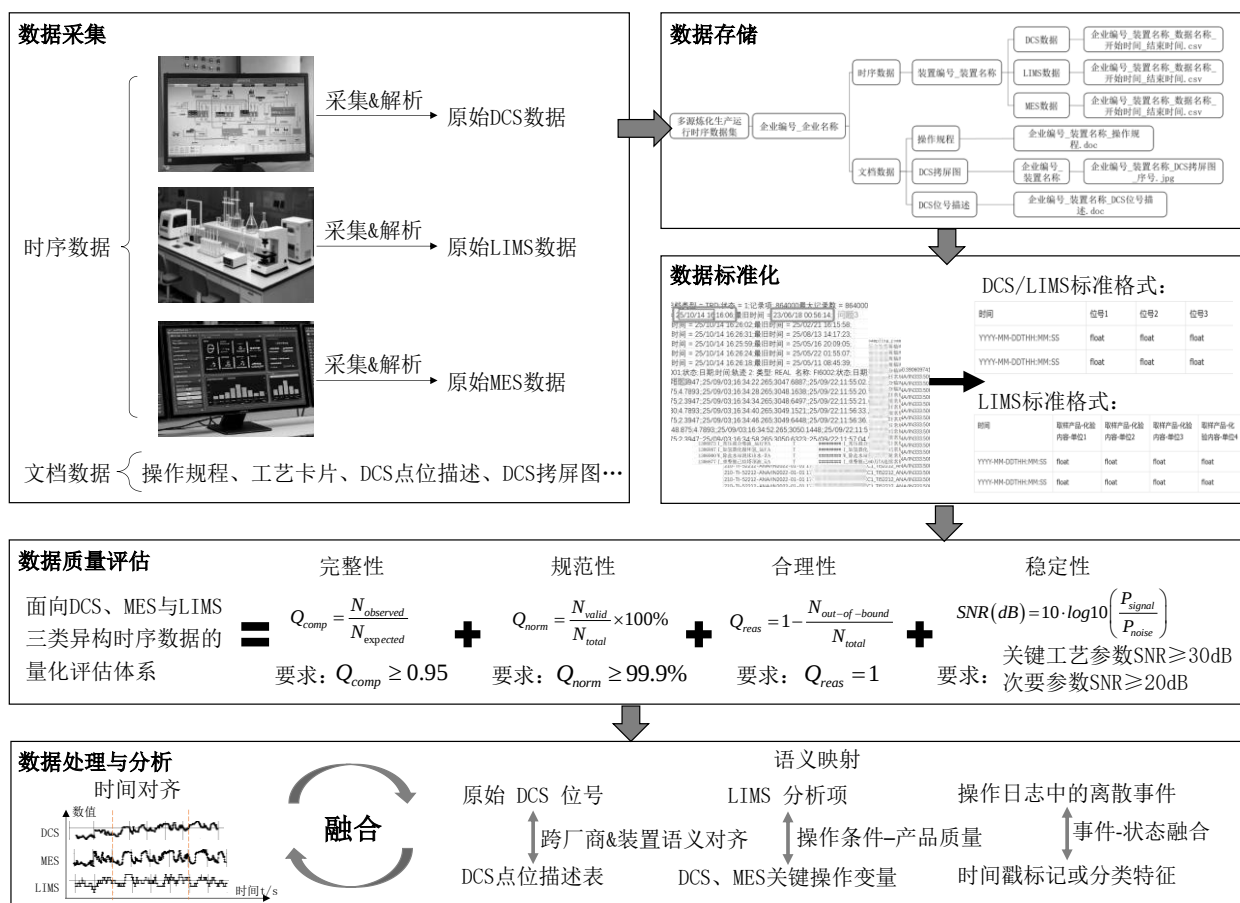


图1 高质量多源炼化生产运行时序数据集构建方法流程图

Fig.1 Flowchart of construction method for high-quality multi-source refining and chemical production operational time series dataset

1.1 数据采集

本文所用数据均来源于各炼化企业实际运行中的核心信息系统。其中,DCS数据是获取装置运行状态最核心的数据来源。根据生产装置特性,DCS通常覆盖上千个测量点位,包括反应器温度、压力、进料流量、液位及关键阀门开度等过程变量,采样频率大多为1秒。这些高维、高频、连续的时间序列数据完整刻画了装置的动态行为,日均数据量多者可达上亿条,具有强时序性、高相关性与低延迟特性。该类数据根据各系统存储历史数据周期的

实际情况,通过各装置所属工程师站定期导出并人工拷贝获取。

LIMS数据则提供与产品质量密切相关的物性分析数据,包括对原料、中间物料及产品的物性与组分分析结果,典型指标如组分含量、初馏点、PH值等。受限于人工取样与实验室分析流程,LIMS数据通常以班次(8小时)或更长周期采集,单点每日仅产生1-3条记录,并伴随30分钟至数小时的滞后。尽管LIMS数据时间分辨率远低于DCS,但其分析精度高、工艺意义明确,是模型监督训练与产品质量

判定的关键依据。该类数据将直接从实验室信息管理系统中定期批量导出,并按装置类型进行数据拆分。

MES数据集成装置物料平衡、能耗统计、操作指令及事件日志等多源信息,不仅包含DCS中的关键运行参数(如温度、压力、液位、流量等),还记录能耗(能源介质消耗量)、收率等结构化或半结构化数据。其时间粒度通常为分钟级,兼具连续变量与离散事件特征,在衔接底层过程动态与上层生产管理方面发挥桥梁作用,为能效评估、操作优化与跨系统协同提供重要支撑。该类数据均以文件形式(如CSV、SQL、TXT等)由工程师定期导出,并经初步整理后用于后续处理流程。

此外,收集各类文档数据,包括操作规程、工艺卡片、DCS点位描述、DCS拷屏图等文档类数据,用于数据分析过程中的语义对齐、数据理解和数据标注等工作。

1.2 数据存储

为解决多源异构数据的使用管理问题,实现炼化生产运行数据的系统化标准化,本文提出分层分类的数据组织架构。该架构以炼化企业为顶层单元,依次按数据类型、工艺装置及数据系统来源进行四级逻辑划分:首先以“企业编号_企业名称”标识企业级实体;其次划分为“时序数据”与“文档数据”两类核心数据类型;然后,按“装置编号_装置名称”组织装置级数据单元;最终细分为DCS、MES和LIMS三类子系统数据源,具体数据集结构如图2所示。其中,时序数据目录集中归档为结构化时间序列,涵盖DCS秒级过程变量、MES分钟级变量以及LIMS小时级物性分析结果;文档数据目录则统一管理为非结构化或半结构化辅助资料,包括操作规程、DCS位号描述表、历史操作日志及DCS拷屏图等,用于支撑变量语义解析与上下文关联。这种结构不仅便于数据检索和访问,也为后续数据的标准化、预处理和模型训练等方面提供有利条件。

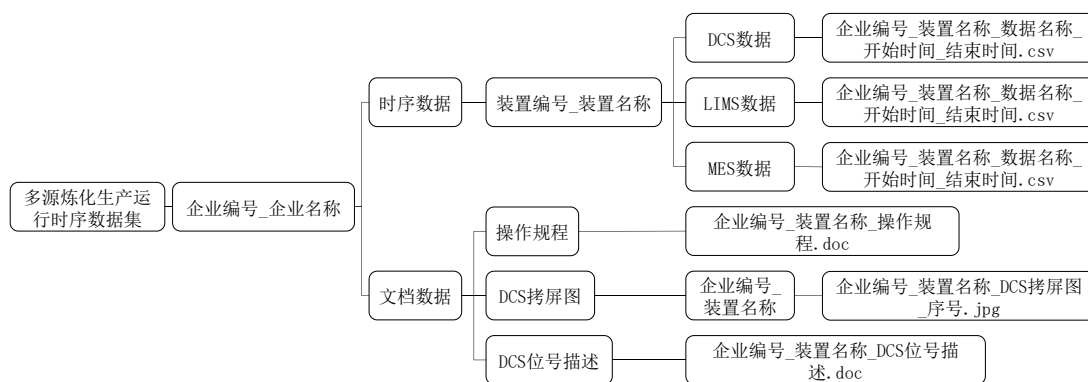


图2 多源炼化生产运行时序数据集结构图

Fig.2 Structure diagram of multi-source refining and chemical production operation time series dataset

1.3 数据标准化

炼化生产运行数据的标准化面临显著挑战:DCS、LIMS与MES数据受系统品牌、装置工艺类型、企业管理规范及人工操作习惯等多重因素影响,导致原始数据在字段命名、单位体系、时间表达及记

录结构等方面存在显著不一致性,如表1-3所示。未进行数据标准化的原始数据会导致应用困难,严重制约后续的数据融合与模型训练,本文通过通用规则的制定和自动化工具直接实现格式统一。

表1 DCS数据现状
Table 1 DCS data status

DCS品牌	DCS数据描述
ECS-700系统	秒级采样频率;不同装置存在不同数量的位号,会将位号数据记录分散在多个文件中(数量不定)
ACB800F7.2系统	不同装置的采样频率设置不同;采样时间间隔不均匀;每个文件所包含位号不超过6个,一套装置约有两三百个文件,且存在位号重复出现于多个文件的情况。
CS-3000系统	不同装置的采样频率设置不同;以单个趋势组8个位号保存,一套装置约有一二百个的文件。

表2 LIMS数据现状
Table 2 LIMS data status

LIMS 数据 企业 来源	LIMS数据描述
企业A process_unit;装置名称;product_name;取样产品;name:化验内容;result_order_number;数值;units_display:单位	
企业B A列:采样时间;B列:装置名称;M列:样品名称;D列:分析项目;E列:数值(分析结果);H列:单位;T列:采样点位置;O列:发布时间 (样品化验出结果后,上传录入系统的时间)	

表3 MES数据现状
Table 3 MES data status

MES数据来源	MES数据描述
某完整性平台	分钟级采样频率;全企业所有装置的位号数据按以下字段存储:tag_no-‘仪表位号’,formatted_time-‘显示时间’,data_time-‘数据采样时间’,phd_tag-‘实时数据库位号’

为解决上述问题,本文提出一种面向多源异构系统的格式层级结构规范化方法,旨在解决因DCS、LIMS与MES系统品牌差异及工程实施多样性所导致的数据表结构不一致问题。具体而言,针对已解析的DCS、LIMS与MES时序数据,制定统一的数据表结构规范:所有数据文件均以CSV格式存储,首行为变量标识行(包含原始位号或分析项名称),首列为ISO 8601标准时间戳(精确至秒),其余单元格按时间对齐填充对应数值。DCS、LIMS和MES的数据标准化格式如表4-6所示。该结构不改变原始位号命名,但强制统一时空组织方式,确保不同来源数据在后续处理中具备一致的读取接口与对齐基础。同时,配套建立轻量级元数据注册机制,记录

每张数据表对应的装置编号、数据源类型、采样频率、时间范围及原始系统信息,用于追溯与上下文关联。该方法在保留原始数据语义完整性的同时,有效消除了系统间的数据表达歧义,为多源时序数据的自动化清洗、对齐、质量评估与模型输入预处理提供了结构化、可解析、可复用的数据基础。

表4 DCS数据标准化格式(企业X-装置X)
Table 4 DCS data standardization format (Enterprise X – Device X)

时间	位号1	位号2	位号3	位号...
YYYY-MM-DDTHH:MM:SS	float	float	float	...
YYYY-MM-DDTHH:MM:SS	float	float	float	...

表5 LIMS数据标准化格式(企业X-装置X)
Table 5 LIMS data standardization format (Enterprise X – Device X)

时间	取样产品-化验内容-单 位1	取样产品-化验内容-单 位2	取样产品-化验内容-单 位3	取样产品-化验内容-单 位...
YYYY-MM-DDTHH:MM:SS	float	float	float	...
YYYY-MM-DDTHH:MM:SS	float	float	float	...

表6 MES数据标准化格式(企业X-装置X)
Table 6 MES data standardization format (Enterprise X – Device X)

时间	位号1	位号2	位号3	位号...
YYYY-MM-DDTHH:MM:SS	float	float	float	...
YYYY-MM-DDTHH:MM:SS	float	float	float	...

1.4 数据质量评估体系

建立一套可量化的数据质量评估体系是确保炼化时序数据集高质量构建的关键[4, 19]。本文构建一套面向DCS、MES与LIMS三类时序数据的量化

评估体系。该体系包含四个可量化计算的核心维度:完整性、规范性、合理性与稳定性。以下四个量化指标通过专家经验和效果测试,规定相应的指标阈值,以构建高质量炼化时序数据集。

1.4.1 完整性 计算时间序列在采样周期下的完整覆盖程度。对于变量 v_i 在时间窗口 $[t_s, t_e]$ 内,应有记录数 $N_{expected}$ 为:

$$N_{expected} = \left\lfloor \frac{t_e - t_s}{\Delta t_i} \right\rfloor + 1 \quad (1)$$

其中 Δt_i 为数据采样间隔,完整性 Q_{comp} 计算定义见式(2):

$$Q_{comp} = \frac{N_{observed}}{N_{expected}} \quad (2)$$

$N_{expected}$ 表示完整时间范围内应有记录的数量;

$N_{observed}$ 表示完整时间范围内实际记录的数量。

DCS、MES 为秒级/分钟级过程数据,是时序数据集构建和后续场景应用的核心输入,因此依据炼化装置连续运行要求设定阈值。依据专家经验,按数据类型设定阈值:DCS 和 MES 数据要求 $Q_{comp} \geq 0.95$;LIMS 数据按班次计,允许单月最多缺失 1 个班次。

1.4.2 规范性 评估数据是否严格遵循标准化结构。校验项包括:(1)首列是否为标准时间戳格式(YYYY-MM-DDTHH:MM:SS);(2)数值列是否均为可解析浮点数(非空、非文本、非异常编码如“#N/A”等);(3)变量名(首行)是否非空且唯一。规范性 Q_{norm} 计算定义为合规单元格占比,见式(3):

$$Q_{norm} = \frac{N_{valid}}{N_{total}} \times 100\% \quad (3)$$

N_{valid} 表示合规单元格的数目;

N_{total} 表示全部单元格的数目。

标准化格式是多源数据批量处理与应用的基础,为保证数据接口统一与程序可读性,强制要求规范性 $Q_{norm} \geq 99.9\%$,否则视为不合格,需重新格式标准化。

1.4.3 合理性 即变量值是否处于物理或工艺允许范围内。例如:压力 ≥ 0 MPa、温度 $\in [-50, 600]$ °C、流量 ≥ 0 t/h 等。合理性 Q_{reas} 得分定义见式(4):

$$Q_{reas} = 1 - \frac{N_{out-of-bound}}{N_{total}} \quad (4)$$

$N_{out-of-bound}$ 表示超出允许范围的变量数量;

N_{total} 表示全部变量的数量。

工艺参数范围来自装置工艺卡片、设备设计说明书及行业安全运行规范,超出范围可能受传感器干扰、数据传输等因素影响,因此要求所有数据源均满足 $Q_{reas} = 1$,其中,超出允许范围的数据点需通过插值方法进行校正。

1.4.4 稳定性 用于评估连续型过程变量的时序动态是否受随机噪声显著干扰。对 DCS 与 MES 中的连续变量(如温度、压力、流量),以信噪比 SNR 计算量化[20]见式(5):

$$SNR(\text{dB}) = 10 \cdot \log_{10} \left(\frac{P_{signal}}{P_{noise}} \right) \quad (5)$$

信噪比依据已有工作在化工领域场景中对信

号噪声的分析结果确定[21],要求数据信噪比 $\geq 20\text{dB}$ 。LIMS 数据为离散分析值,不适用此项评估。

基于以上几个维度的量化评估,对于发现的数据质量问题,需使用插值填充、异常值剔除及逻辑校正等手段,进行数据修复。缺失值处理采用邻域瞬时插值,仅补全瞬时缺失数据,不进行平滑滤波与趋势修正,最大限度保留原始时序信号的真实动态特征。仅当每个变量在其适用的所有维度上均达标时,其对应数据段才被纳入高质量炼化时序数据集。

1.5 数据处理与分析

在完成多源数据的采集、存储、标准化与质量评估后,本文设计基于时间对齐与语义映射的多源融合方法,旨在解决不同来源数据之间的时间和语义差异问题,为后续建模提供时间和语义对齐的输入基础。

首先,时间对齐以构建统一的时间基准为目标,确保同一装置内 DCS、MES 与 LIMS 数据在时间基准上严格对应。具体而言,DCS 数据以其原始高频率采样点作为主时间轴;将 MES 数据时间戳映射至距离 DCS 数据所属时段最近的时刻;LIMS 数据为班次级分析结果(如每 8 小时一条),严格采用记录的采样时间作为 LIMS 数据时间戳,并映射至对应班次内对应的 DCS/MES 时间点,形成时间基准的多源观测样本。其中,LIMS 数据从现场采样,到记录采样时间和结果发布时间,均存在时间滞后,但因现场实际因素影响,无法准确获取真实现场采样时间,因此均保持记录采样时间为统一标准。尽管记录采样时间与真实采样时刻存在滞后无法完全消除,但该策略可最大限度保证在现有工业数据条件下时序建模的合规性与公平性。该策略避免了因原始时间戳错位导致的同一装置中跨系统变量误配,同时保留了各类数据的原始动态特性,无需引入插值或重采样带来的信息失真。

其次,语义映射依托炼化领域知识库与装置级元数据,构建跨系统变量关联规则,使得 DCS、LIMS 和 MES 数据相互关联,具体包括以下三个层次:

(1)位号归一化:以装置 DCS 点位描述表为基础,完成 DCS 位号、MES 位号与工艺描述的统一映射与语义对齐;对无明确点位描述的变量,依托装置操作规程、工艺卡片等文档材料,并结合领域专家经验完成语义补充与规范命名。在此基础上,通过关键词规则匹配(如识别“温度”“压力”“流量”及

单元名称“R1”“C2”等)或轻量级文本嵌入相似度计算,可将语义相近的变量归入同一工艺变量等价类(例如所有描述含“R1 入口 温度”的测点归为一类),形成如表7所示的位号归一化描述表。该方法在保留原始位号的同时实现跨厂商、跨装置的语义对齐,为后续多源融合提供可解释的变量关联基础。

(2)跨模态关联:基于工艺机理与操作经验,建立 LIMS 分析项(如“重整生成油芳烃含量”)与

表7 位号归一化描述表

Table 7 Item number normalization description table

序号	DCS 位号	MES 位号	位号描述	工艺变量等价类
1(示例)	FIC1103	FIC-1103	C101 进料量	C101-流量

DCS、MES 关键操作变量(如“反应温度”“循环氢流量”)之间的逻辑关联,设计标准化关联模板如表8所示,形成“操作条件-产品质量”的映射关系。

表8 标准化跨模态关联模板

Table 8 Standardized cross-modal association template

序号	LIMS 分析项	DCS/MES 关键操作变量	相关性程度(高/中/低)
1(填写说明)	取样产品-化验内容	位号 A	高/中/低

(3)事件-状态融合:将操作日志中的离散事件(如“切换原料”“压缩机停机”“工艺卡片变更”)转化为时间戳标记和特征,设计标准化事件-状态模板如表9所示,由业务侧人工梳理与标注,将离散事

件转化为带时间戳的特征项,用于后续嵌入连续时序数据作为上下文信号,增强模型对非稳态工况的感知能力。

表9 标准化事件-状态模板

Table 9 Standardized event-state template

序号	离散事件	开始-结束时间	事件描述	相关位号
1(填写说明)	异常波动/原料切换等	2025-04-18T12:31:12- 2025-04-18T12:33:40	描述事件的发现、处置流程、原因分析等	位号 A、位号 B...

上述映射规则按“厂-装置-单元”三级结构存储和加载,确保该方法在多厂多装置场景下的可配置性与可扩展性。

2 结果与讨论

2.1 炼化生产运行时序数据集构建结果

基于本文提出的高质量数据集构建方法,已成功构建一套面向炼化行业的高质量生产运行时序数据集。为量化数据集数据量大小,本文采用数据点数进行统计,即以时间戳为行、变量为列构成的表格中,所有有效数值型数据点的总数量。该数据集包含多家炼化企业的18类典型装置数据(包括常减压、催化裂化、催化重整、加氢裂化、延迟焦化、汽油加氢、柴油加氢、航煤加氢、气分、硫黄回收、酸性水汽提、制氢、乙烯、聚乙烯、聚丙烯、苯乙烯、芳烃、烷基化),累计原始数据存储量达20 TB。其中,DCS 数据累计采集15类装置,因受不同系统影响,采集时间跨度不一,多数装置已连续采集一至两年的数据量,约370亿数据点,存储占比约52.91%;MES 数

据累计采集18类装置,包含22年1月起的全量数据,约330亿数据点,存储占比约47.08%;LIMS 数据累计采集12类装置,包含22年1月起的全量数据,约0.034亿数据点,存储占比约0.01%。在统一的分层分类存储架构下,所有数据均经过标准化格式转换,形成700亿总数据点规模的高质量时序数据集。

2.2 数据质量评估体系结果

根据本文制定的可量化质量评估体系,针对标准化后数据进行各维度质量评估,并进行异常数据处理,三类数据各随机选取某一装置的数据进行质量评估指标计算,如表10所示。

表10 数据质量评估指标结果

Table 10 Data quality assessment indicator results

指标	DCS		MES		LIMS	
	处理前	处理后	处理前	处理后	处理前	处理后
完整性	100.0%	100.0%	93.2%	99.9%	1.0	1.0
规范性	90.2%	100.0%	82.2%	100.0%	100.0%	100.0%
合理性	1	1	1	1	1	1
稳定性/dB	33.80	33.94	16.08	20.77	-	-

整体而言,经过插值填充、异常段剔除及逻辑校正等手段,数据集在各维度上的平均分数满足设定要求。在完整性方面,DCS与MES数据存在的数据缺失情况,主要源于装置开停工、网络丢包等情况,经处理后完整性可达约100%。规范性问题集中历史导出文件的格式异构,如时间戳混用、数据值含特殊字符等,处理后数据规范性指标均达到100%,为后续批处理提供了结构一致性保障。合理性分析结果聚焦于两类异常:一是传感器漂移导致的持续性越界(如压力负值),采用滑动窗口邻域插值予以校正或直接剔除;二是工艺瞬态引起的短暂超限,此类记录被保留但标注“合理异常”,避免误删有效工况。在稳定性方面,剔除明显噪声分量,同时最大限度保留关键动态特征,使DCS数据处理后平均稳定性达33.94dB, MES数据处理后稳定性达20.77dB。

为验证本文构建的高质量数据集的有效性,选

取生产运行参数预测作为试点任务开展数据集构建方法论效果的对比验证。分别采用原始未处理数据集与本文构建的高质量数据集训练相同结构的预测模型,以均方误差(Mean Squared Error, MSE)和平均绝对误差(Mean Absolute Error, MAE)作为评价指标。实验所用数据为某常减压装置MES时序数据,时间跨度为2023年6月–2025年10月,预测目标为装置关键工艺参数PI01。数据按8:2划分为训练集与测试集,输入时间步长为60,构建长短期记忆网络(Long Short-Term Memory, LSTM)模型并采用三步预测模式。结果表明,相较于原始数据,经本文方法处理后的数据集可使运行参数预测模型的MSE降低67.6%,MAE降低42.5%,模型性能显著提升,如图3所示。该实验验证表明高质量数据集具备良好的数据一致性与时序表征能力,可有效支撑智能化建模任务,并验证本文所提数据质量评估体系的可行性,和对下游任务支撑的有效性。

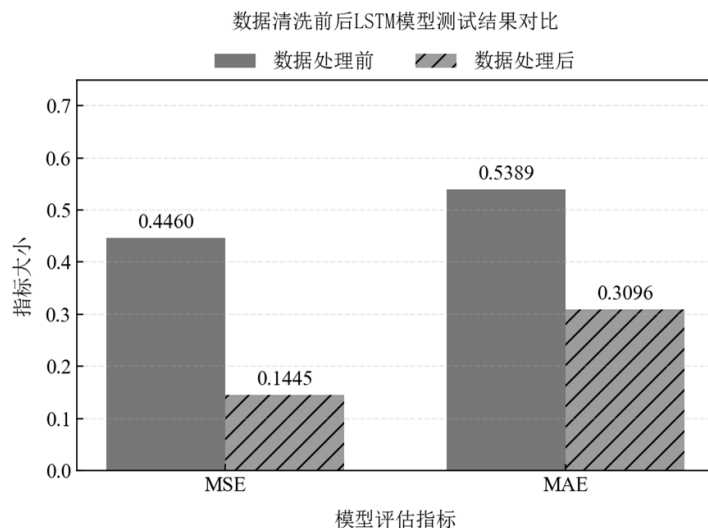


图3 数据处理前后对试点任务的验证结果对比图

Fig.3 Comparison chart of verification results for pilot tasks before and after data processing

2.3 数据处理与分析结果

数据处理与分析阶段,基于所提出的时间对齐策略,98.6%的MES记录与DCS时间基准准确匹配,LIMS标签平均覆盖每个班次内约2,880个DCS采样点,有效实现了多源数据在统一时间基准下的精准协同。针对语义映射,可针对各类装置构建“位号归一化描述”、“操作条件-产品质量关联表”、“事件-状态关联表”,将DCS、MES、LIMS数据和描述、离散事件等进行了相关关联。例如,“事件-状态关联表”中,除根据操作记录整合离散事件描述外,针对相关位号列,首先将相关系数绝对值>

0.65的操作位号进行提取映射,并结合专家审核和修改形成。

综上,通过本文提出的一套高质量多源炼化生产运行时序数据集构建方法提供了可复现、可推广的方法论框架,并形成一套高质量、标准化、语义可解释的炼化生产运行时序数据集。

3 炼化时序数据集的应用场景

基于构建的高质量炼化生产运行时序数据集,本文进一步分析讨论其在人工智能驱动下的两类

典型应用:一是面向通用炼化相关能力学习的基础时序大模型构建,二是面向具体炼化业务目标的智能体(Agent)开发[22-24]。二者分别从“底座模型”与“垂直应用”两个维度,分析数据集的技术支撑价值与工程落地潜力。

3.1 炼化生产运行基础时序大模型构建

炼化过程具有强非线性、高耦合性与长时依赖等复杂动态特性,传统机理模型或浅层数据驱动方法难以全面刻画其内在规律。为此,依托标准化后的DCS、LIMS与MES融合时序数据,可构建面向炼化行业的基础时序大模型(Foundation Time-series Model, FTM)。当前已有时序大模型大多采用基于Transformer、混合专家等架构[25-28],在此基础上,进一步基于炼化时序数据进行训练,学习跨装置、跨工况的通用表征能力,得到炼化行业的基础时序大模型。通过在包含数十套装置、数年跨度的高质量数据集上训练,可有效捕获温度-压力-流量等关键变量间的动态耦合关系,以及生产质量指标与操作参数之间的隐式映射。预训练完成后,模型可通过少量标注数据进行微调,快速适配至质量预测、能效评估、故障预警等下游任务。基于高质量数据集构建基础时序大模型,不仅为炼化领域提供了可复用的时序大模型底座,也为后续构建行业场景大模型奠定数据与模型基础。

3.2 业务场景智能体构建

在基础时序大模型能力之上,本文进一步讨论面向具体业务目标的炼化生产智能体构建。智能体以“感知—推理—决策—执行”闭环为核心架构,深度融合多源时序数据、工艺知识图谱与操作规程文档,实现对特定场景的自主认知与优化干预。为提升智能体在多变工况与极端场景下的适应性与可靠性,构建“机理模型+数据驱动”混合架构:将炼化工艺机理(如物料平衡、能量平衡、反应动力学等)嵌入到损失函数等,作为硬约束规则;以时序数据集学习获取工况模式与操作规律,两者协同推理,实现高可靠智能决策。典型应用场景包括:

(1)产品质量预测智能体:融合DCS操作变量与LIMS化验分析数据,通过历史操作参数预测产品质量,并支撑在线学习动态修正操作参数(如反应温度、进料量等)[29],使产品质量参数稳定控制在目标区间;

(2)能效优化智能体:基于MES能耗统计与装置负荷数据,结合电价时段与蒸汽平衡约束等,生

成关于水、电、气等最优调度策略,降低单位加工能耗[30];

(3)异常工况诊断与处置智能体:利用历史事件日志与DCS异常模式,结合大模型语义理解能力,自动识别故障根因并推荐标准处置流程,辅助操作员快速响应[31]。

所有智能体通过模块化方式部署,并通过人机协同接口实现“人在环路”(Human-in-the-loop)的安全可控运行,可满足炼化生产的业务需求。

综上,本文所展示的应用路径可支撑高质量多源时序数据集在推动炼化企业从“数据可用”向“智能可用”跃迁中的核心作用。

4 结 论

本文围绕炼化生产运行中DCS、LIMS与MES多源异构数据的高质量构建难题,提出了一套时序数据集构建方法。主要工作与结论如下:

(1)设计“采集—存储—标准化—评估—处理与分析”五阶段闭环流程,通过分层分类的存储架构与统一表格格式,解决了多源数据结构混乱、难以协同的问题。

(2)提出一套面向工业多源时序数据的可量化质量评估体系,将传统模糊的质量要求转化为四个维度上可自动计算、可设定阈值、可批量判定的数值化指标,显著提升了数据筛选的客观性与可操作性,为高质量时序数据集的规模化构建提供了方法论支撑。

(3)提出基于时间对齐与语义映射的多源融合策略,有效弥合了DCS、MES与LIMS在时序粒度与语义表达上的异构性,实现了多源数据在操作、状态与质量层面的协同表征。

(4)基于本文所提高质量数据集构建方法,构建700亿总数据点规模的高质量炼化生产运行时序数据集,可有效支撑炼化领域智能化应用场景的落地。

本研究不仅为炼化行业提供了可复用的高质量时序数据集构建范式,也为工业数据资产化与大模型驱动的智能化转型奠定了坚实基础。

数据可用性声明:

本文研究所用原始生产数据来自各炼化企业,属于企业专有生产数据,数据资产归属相关炼化企业所有,并受企业数据安全与保密管理规定约束,

因此数据集当前暂时无法对外提供共享。

参考文献

- [1] 国家数据局, 中共中央网络安全和信息化委员会办公室, 工业和信息化部, 等. "数据要素×"三年行动计划(2024—2026年)[Z]. 2024.
National Data Administration, Office of the Central Leading Group for Cyberspace Affairs, Ministry of Industry and Information Technology, etc. "Data Element × " Three-Year Action Plan (2024-2026) [Z]. 2024
- [2] Liang W X, Tadesse G A, Ho D, et al. Advances, challenges and opportunities in creating data for trustworthy AI[J]. *Nature Machine Intelligence*, 2022, **4**(8): 669-677.
- [3] Liu Y, Cao J H, Liu C Y, et al. Datasets for large language models: a comprehensive survey[J]. *Artificial Intelligence Review*, 2025, **58**(12): 403
- [4] Gong Y D, Liu G Z, Xue Y Z, et al. A survey on dataset quality in machine learning[J]. *Information and Software Technology*, 2023, **162**: 107268.
- [5] Pani A K, Soofastaei A. Designing intelligence: harnessing soft sensors and advanced analytics in petroleum refining for Industry 4.0[M]//Advanced Analytics for Industry 4.0. Boca Raton: CRC Press, 2025: 83-116.
- [6] 罗梦杰, 赵云鹏, 张梦轩, 等. 大模型时代的石油化工行业高质量数据集: 挑战与机遇[J]. *化工学报*, 2026, **77**(1): 435-447, F0004.
Luo M J, Zhao Y P, Zhang M X, et al. High-quality datasets for petrochemical industry in era of large language models: challenges and opportunities[J]. *CIESC Journal*, 2026, **77**(1): 435-447, F0004.
- [7] Wang C J, Zhu W B, Gao B B, et al. Real-IAD: a real-world multi-view dataset for benchmarking versatile industrial anomaly detection[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 16-22, 2024, Seattle, WA, USA. IEEE, 2024: 22883-22892.
- [8] Onyeke F O, Odujobi O, Adikwu F E, et al. Advancements in the integration and optimization of control systems: Overcoming challenges in DCS, SIS, and PLC deployments for refinery automation[J]. *Open Access Research Journal of Multidisciplinary Studies*, 2022, **4**(2): 94-101.
- [9] Andrade J M, Prada D, Muniategui S, et al. A laboratory information management system in a petrochemical company: an effective information solution[J]. *Chemometrics and Intelligent Laboratory Systems*, 1993, **21**(1): 35-49.[LinkOut]
- [10] Yuan Z H, Qin W Z, Zhao J S. Smart manufacturing for the oil refining and petrochemical industry[J]. *Engineering*, 2017, **3**(2): 179-182.
- [11] 郭佳佳. 基于元数据的石油化工行业数据治理[J]. *石油知识*, 2024(6): 46-47.
Guo J J. Data governance of petrochemical industry based on metadata[J]. *Petroleum Knowledge*, 2024(6): 46-47.
- [12] Wu Q Y, Yao D, Liu Y, et al. Research on construction methods of high-quality multimodal datasets in industrial manufacturing[C]//2025 IEEE International Conference on Blockchain (Blockchain). October 30 - November 2, 2025, Zhengzhou, China. IEEE, 2025: 241-246.
- [13] Duan R, Deng H, Tian M, et al. SODA: a large-scale open site object detection dataset for deep learning in construction[J]. *Automation in Construction*, 2022, **142**: 104499.
- [14] Li J Y, Tang J Z, Zhao X Z, et al. A large-scale, high-quality dataset for lithology identification: Construction and applications[J]. *Petroleum Science*, 2025, **22**(8): 3207-3228.
- [15] 周军, 石玉江, 张娟, 等. 统一测井数据库建设与应用[J]. *测井技术*, 2022, **46**(6): 757-761.
Zhou J, Shi Y J, Zhang J, et al. Construction and application of unified logging database[J]. *Well Logging Technology*, 2022, **46**(6): 757-761.
- [16] 李梁, 姜辉, 宋一纯, 等. AI场景下数据治理体系建设和实施策略[J]. *管道保护*, 2025, **2**(5): 78-82, 90.
Li L, Jiang H, Song Y C, et al. Construction and implementation strategy of data governance system in AI scenario [J]. *Pipeline Protection*, 2025, **2**(5): 78-82, 90.
- [17] 刘珊珊. 大语言模型在勘探开发数据治理中应用研究[J]. *北京石油化工学院学报*, 2025, **33**(2): 64-72.
Liu S S. Exploration of the application of large language models in exploration and development data governance[J]. *Journal of Beijing Institute of Petrochemical Technology*, 2025, **33**(2): 64-72.
- [18] Shu L, Mukherjee M, Pecht M, et al. Challenges and research issues of data management in IoT for large-scale petrochemical plants[J]. *IEEE Systems Journal*, 2018, **12**(3): 2509-2523.
- [19] Mohammed S, Budach L, Feuerpfeil M, et al. The effects of data quality on machine learning performance on tabular data[EB/OL]. 2022: arXiv: 2207.14529.
- [20] Johnson D H. Signal-to-noise ratio[J]. *Scholarpedia*, 2006, **1**(12): 2088.
- [21] Kini K R, Madakyaru M. Improved process monitoring scheme using multi-scale independent component analysis[J]. *Arabian Journal for Science and Engineering*, 2022, **47**(5): 5985-6000.
- [22] Kashpruk N, Piskor-Ignatowicz C, Baranowski J. Time series prediction in industry 4.0: a comprehensive review and prospects for future advancements[J]. *Applied Sciences*, 2023, **13**(22): 12374.
- [23] Bousetouane F. Agentic systems: a guide to transforming industries with vertical AI agents[EB/OL]. 2025: arXiv: 2501.00881.
- [24] Liang G N, Tong Q Q. LLM-powered AI agent systems and their applications in industry[C]//2025 IEEE World AI IoT Congress (AIoT). May 28-30, 2025, Seattle, WA, USA. IEEE, 2025: 463-471.
- [25] Ahmed S, Nielsen I E, Tripathi A, et al. Transformers in time-series analysis: A tutorial[J]. *Circuits, systems, and signal processing*, 2023, **42**(12): 7433-7466.
- [26] Shi X M, Wang S Y, Nie Y Q, et al. Time-MoE: billion-scale time series foundation models with mixture of experts[EB/OL]. 2024: arXiv: 2409.16040.
- [27] Gao S H, Hartvigsen T, Koker T, et al. UniTS: a unified multi-task time series model[C]//Advances in Neural Information Processing Systems 37. December 10-15, 2024, Vancouver, BC, Canada. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024: 140589-140631.
- [28] Liu Y, Zhang H R, Li C Y, et al. Timer: transformers for time series analysis at scale[J]. *arXiv preprint*, 2024, arXiv: 2402.02368.

- [29] Saghir H, Ahmad I, Kano M, et al. Prediction and optimisation of gasoline quality in petroleum refining: The use of machine learning model as a surrogate in optimisation framework[J]. CAAI Transactions on Intelligence Technology, 2024, **9**(5): 1185–1198.
- [30] Gholamzadeh E, Shokri A, Ghaemi A, et al. Optimization of energy consumption in the gas refinery sweetening unit using RSM, ANN, and Aspen HYSYS[J]. Scientific Reports, 2025, **15**: 38763.
- [31] Tian W D, Wang S C, Sun S L, et al. Intelligent prediction and early warning of abnormal conditions for fluid catalytic cracking process[J]. Chemical Engineering Research and Design, 2022, **181**: 304–320.